بنام خدا

هفتمین کنفرانس آمار ایران

مجموعه مقالات

جلد دوّم (فارسی)

۲ تا ۴ شهریور ماه ۱۳۸۳
دانشگاه علامه طباطبایی
تهران - ایران

مجموعه مقالات هفتمین کنفرانس آمار ایران

آماده‌سازی و صفحه‌آرایی : LaTeX-پارسی، یاشار اکبری
طراحی جلد : نیما دانش‌پرور
تیراژ : ۸۰۰ نسخه
چاپ و صحافی : مرکز آمار ایران
تاریخ انتشار : بهمن ماه ۱۳۸۳

بنام خدا

پیشگفتار

خداوند بزرگ را سپاسگزاریم که ما را یاری نمود تا هفتمین کنفرانس آمار ایران را با همکاری انجمن آمار ایران، کوشش همکاران، دانشجویان، کارکنان و حمایتهای مسؤلین محترم دانشگاه علامه طباطبایی و سازمانها و دستگاههای مختلف کشور در روزهای ۲ تا ۴ شهریور ۱۳۸۳ در دانشگاه علامه طباطبایی برگزار نمائیم.

تلاش کمیته علمی کنفرانس برای برنامه‌ریزی در جهت ارتقای سطح علمی کنفرانس از ابتدای سال ۱۳۸۲ آغاز گردید و استقبال قابل توجه اساتید، دانشجویان و محققین کشور در ارائه‌ی مقاله در زمینه‌های نظری و کاربردی آمار چشمگیر بود، به طوری که ناچار شدیم به دلیل محدودیت زمانی تعدادی از مقالات را که از کیفیت برتری برخوردار بودند انتخاب نماییم.

متقاضیان شرکت درکنفرانس تعداد ۳۹۲ مقاله به دبیرخانه ارسال نمودند وکمیته‌های تخصصی یازده گانه کمیته علمی کنفرانس تعداد ۱۹۷ مقاله را به عنوان سخنرانی، تعداد ۱۰۱ مقاله را برای ارائه به صورت پوستر و ۶ مورد را به صورت کارگاه آموزشی پذیرش نمودند. به علاوه از دانشمندان صاحب‌نام علم آمار داخل و خارج کشور برای ارائه سخنرانیهای عمومی و تخصصی دعوت به عمل آمد که تعدادی از آنها به دلایلی از شرکت درکنفرانس عذرخواهی نمودند و نهایتاً تعداد ۵۷ مقاله به صورت سخنرانی و ۴ کارگاه آموزشی توسط مدعوین داخلی و خارجی درکنفرانس ارائه گردید. مجموعهٔ حاضر شامل ۳۲ مقاله پذیرفته شده توسط داوران کمیته‌های تخصصی یازده گانه کمیته علمی کنفرانس می‌باشد.

برگزاری هفتمین کنفرانس آمار ایران بدون مساعدت‌های بی‌دریغ همکاران گرامی در دانشکده اقتصاد، تعدادی از اساتید دانشگاه‌های سراسر کشور به عنوان هیأت داوران، دانشجویان دوره کارشناسی وکارشناسی ارشد آمار و حمایتهای هیات رئیسه محترم و تلاشهای کارمندان و کارکنان دانشگاه علامه طباطبایی امکان‌پذیر نبود و جا دارد از تمامی عزیزانی که در برگزاری کنفرانس زحماتی را متحمل شدند قدردانی و تشکر شود.

هماهنگی جلسات کمیته علمی با تلاش آقای مجید رجبی و انجام کلیه‌ی امور کامپیوتری و تهیه و تنظیم مجموعه مقالات با همکاری و پشتکار بی‌دریغ و خستگی‌ناپذیر خانم تی ری تی ماندگاریان و آقای یاشار اکبری صورت پذیرفت که از این عزیزان نهایت تشکر و امتنان را دارد.

بهمن ماه ۱۳۸۳

کمیته‌های علمی و برگزاری هفتمین کنفرانس آمار ایران

حمایت کنندگان

هفتمین کنفرانس آمار ایران توسط دانشگاه علامه طباطبایی و با همکاری انجمن آمار ایران و حمایت‌های علمی و مالی مؤسسات زیر برگزار گردید که بدینوسیله مراتب قدردانی خود را از آنان ابراز می‌داریم.

پژوهشکده آمار

مرکز آمار ایران

سازمان مدیریت و برنامه‌ریزی کشور

بانک مرکزی جمهوری اسلامی ایران

بانک کشاورزی

پژوهشکده هواشناسی و علوم جو

اسامی اعضای کمیته‌های علمی و برگزاری هفتمین کنفرانس آمار ایران

اعضای کمیته برگزاری

۱. دکتر فرزاد اسکندری (مسؤل دبیرخانه کمیته علمی)
۲. آقای محمدرضا اصغری اسکویی (مسؤل امور رایانه)
۳. آقای عبدالرحیم بادامچی‌زاده (دبیر پنجمین مسابقه دانشجویی آمار و مسؤل نمایشگاه کتاب)
۴. دکتر محمد بامنی مقدم (مسؤل امور میهمانان خارجی)
۵. آقای محمدجواد زالی (دبیر کمیته اجرایی)
۶. دکتر شهرام سلیلی (مسؤل پذیرش)
۷. دکتر سید علی میریان (نماینده دانشگاه در کمیته برگزاری)
۸. دکتر **نادر نعمت‌الهی** (مسؤل کمیته و دبیر کنفرانس)
۹. دکتر حمیدرضا نواب‌پور (دبیر کمیته علمی)
۱۰. دکتر عبدالساده نیسی (مدیر امور اجرایی)

اعضای کمیته علمی

۱. دکتر فرزاد اسکندری	دانشگاه علامه طباطبایی
۲. دکتر احمد پارسیان	دانشگاه صنعتی اصفهان
۳. دکتر محمد جلوداری ممقانی	دانشگاه علامه طباطبایی
۴. دکتر صدیقه شمس	دانشگاه الزهرا
۵. دکتر عباس گرامی	پژوهشکده آمار
۶. دکتر یدا... محرابی	دانشگاه علوم پزشکی شهید بهشتی
۷. دکتر محسن محمدزاده	دانشگاه تربیت مدرس
۸. دکتر محمدرضا مشکانی	دانشگاه شهید بهشتی
۹. دکتر **نادر نعمت‌الهی** (دبیر کنفرانس)	دانشگاه علامه طباطبایی
۱۰. دکتر **حمیدرضا نواب‌پور** (دبیر و مسؤل کمیته)	دانشگاه علامه طباطبایی
۱۱. دکتر محمدقاسم وحیدی اصل	دانشگاه شهید بهشتی

فهرست مطالب

عنوان . . . . . صفحه

مقالات جلد اوّل

عنوان . . . . . صفحه

عنوان . . . . . صفحه

مقالات جلد دوّم

عنوان ..... صفحه

عنوان ..... صفحه

برآورد ضریب جینی و بررسی چگونگی توزیع درآمد نقاط شهری و روستایی استان‌های کشور

جعفر رحمانی شمسی

سازمان مدیریت و برنامه‌ریزی استان یزد

چکیده: در این مقاله روشی برای برآورد مقدار ضریب جینی[1] به عنوان مهمترین شاخص بررسی توزیع درآمد[2] ارائه شده است. این روش بر خلاف روش‌های معمول محاسبه ضریب جینی متکی بر داده‌های خام نبوده و در آن می‌توان از داده‌های گروه بندی شده مرکز آمار ایران استفاده کرد. کاربرد اصلی این روش در محاسبه ضرایب جینی استان‌ها می باشد زیرا دسترسی به داده‌های خام استان‌ها بسادگی امکان پذیر نمی‌باشد. همچنین بر اساس این روش برای اولین بار ضرایب جینی کلیه استان‌های کشور به تفکیک نقاط شهری و روستایی محاسبه گردیده تا چگونگی توزیع درآمد در استان‌ها مورد بررسی قرارگیرد همچنین با استفاده از روش آنالیز خوشه‌ای استان‌های کشور از نظر چگونگی توزیع درآمد مقایسه و طبقه‌بندی شده‌اند.

واژه‌های کلیدی: توزیع درآمد، ضریب جینی، آنالیز خوشه‌ای

۱ مقدمه

در سالهای اخیر مسئله توزیع درآمد و توزیع امکانات زندگی بین اقشار مختلف جامعه مورد توجه بیشتری قرارگرفته است. مطالعات مربوط به توزیع درآمد سابقه‌ای کمتر از یک قرن دارد و در این سالها شاخصهای گوناگونی برای محاسبه میزان نابرابری توزیع درآمد مطرح شده است که یکی از مهمترین آنها ضریب جینی است.

۲ ضریب جینی

رایج‌ترین شاخص نابرابری توزیع درآمد که در اکثر مطالعات و بررسی‌های مربوط به توزیع درآمد از آن استفاده می‌شود ضریب جینی است. این معیار در سال ۱۹۱۲ به وسیله آماردان ایتالیایی کورودن سی جینی[3] ارائه شده است و به روش زیر محاسبه می‌شود:

$$G = \frac{\triangle}{۲\mu}$$

1) Gini Coeficient 2) Income Distribution 3) Gini

و

$$\triangle = \frac{۱}{n(n-۱)} \sum_{i=۱}^{n} \sum_{j=۱}^{n} \mid x_i - x_j \mid \qquad (۱)$$

که در آن:
x_i: بیان کننده درآمدi امین فرد جامعه
n : تعداد کل افراد جامعه(نمونه)
μ: میانگین درآمد افراد جامعه(نمونه)

بر اساس رابطه فوق اگر $\triangle$ در حداقل مقدار ممکن (یعنی صفر یا حالتی که توزیع درآمد کاملا عادلانه است) باشد، اندازه ضریب جینی برابر صفر و چنانچه $\triangle$ در حداکثر مقدار ممکن یعنی ۲μ باشد (در حالت توزیع کاملا نا عادلانه درآمد) اندازه ضریب جینی برابر یک خواهد بود.

به طور کلی اندازه ضریب جینی بین صفر و یک تغییر می‌کند که مقدار صفر نشان دهنده برابری کامل توزیع درآمد و مقدار یک نشان دهنده نابرابری کامل توزیع درآمد است، مقادیر بین صفر و یک معمولا بدین صورت تفسیر می‌گردد :

۱ــ $G \leq ۰/۳$ توزیع نسبتا عادلانه درآمد
۲ــ $۰/۳ < G < ۰/۵$ توزیع نیمه عادلانه درآمد
۳ــ $G \geq ۰/۵$ توزیع نا عادلانه درآمد

۳ فرمول برآوردی ضریب جینی

با توجه به فرمول محاسباتی ضریب جینی در خواهیم یافت که محاسبه دقیق ضریب جینی نیاز به درآمد(هزینه) تک تک افراد جامعه(نمونه) دارد. این مقدار برای نقاط شهری و روستایی کشور در اکثر سالها توسط مرکز آمار ایران و یا برخی از محققان محاسبه شده است، اطلاعات مورد نیاز برای محاسبه این معیار از داده‌های خام طرح هزینه و درآمد خانوارهای شهری و روستایی مرکز آمار ایران بدست آمده است. همچنین این معیار در برخی از استان‌های کشور در بعضی از سالها بر اساس داده‌های خام این طرح محاسبه شده است ولی تاکنون تحلیل جامعی از چگونگی توزیع درآمد کلیه استان‌های کشور انجام نگرفته است و دلیل اصلی آن عدم دسترسی ساده به این داده‌ها به تفکیک استان‌های کشور و مشکلاتی است که محاسبه این ضریب از روی داده‌های خام دارد. با این وجود اطلاعات گروه بندی شده هزینه های خانوارهای شهری و روستایی به تفکیک استان‌های کشور هر ساله توسط مرکز آمار ایران منتشر می‌شود. در نشریات این مرکز خانوارهای شهری و روستایی در ۱۰ گروه هزینه‌ای گروه‌بندی شده‌اند و متوسط هزینه هر گروه و نیز فراوانی خانوارهای درون این گروهها در اختیار قرار گرفته است. در واقع بجای هزینه هر

خانوار درون گروههای هزینه‌ای تنها متوسط هزینه آن خانوارها برای هر استان در اختیار است. بر اساس ایده داده‌های گمشده(missing values) می‌توان هزینه هر خانوار درون گروهها را به عنوان داده گمشده در نظر گرفت و متوسط هزینه آن گروه را جایگزین آنها کرد. در واقع جداولی که در اختیار است را می‌توان به عنوان جداول فراوانی داده‌ها در نظر گرفت که داده‌های خام توسط آنها در ۱۰ گروه طبقه‌بندی شده‌اند و متوسط هر طبقه(نماینده طبقه) و نیز فراوانی هر طبقه در اختیار است. در این حالت فرمول برآوردی ضریب جینی را می‌توان به صورت زیر نوشت :

$$\triangle = \frac{۱}{n(n-۱)} \sum_{i=۱}^{۱۰} \sum_{j=۱}^{۱۰} f_i f_j \mid \bar{x}_i - \bar{x}_j \mid$$

که در آن:
$\bar{x}_i$: متوسط هزینه خانوارهای گروه هزینه‌ای iام
n: تعداد کل افراد جامعه (نمونه)
f_i: تعداد خانوارهای درون گروه هزینه‌ای iام

۴ برآورد ضریب جینی نقاط شهری و روستایی استان‌های کشور بین سالهای ۱۳۷۴ تا ۱۳۸۰

در این بخش بر اساس نتایج طرح هزینه و درآمد خانوارهای شهری و روستایی مرکز آمار ایران و با استفاده از فرمول برآوردی ارائه شده، برآورد ضریب جینی نقاط شهری و روستایی تمام استان‌های کشور در سالهای ۱۳۷۴ تا ۱۳۸۰ بدست آمده است، همچنین با استفاده از میانگین هندسی متوسط ضریب جینی هر استان بین سالهای ۱۳۷۴ تا ۱۳۸۰ نیز محاسبه شده است. برای سادگی محاسبات شبه نرم‌افزاری در محیط Excel طراحی شده و تنها با وارد کردن اطلاعات مورد نیاز یعنی متوسط هزینه هر گروه‌های هزینه‌ای و تعداد خانوارهای درون گروههای هزینه‌ای می‌توان ضریب جینی را مشاهده کرد. نتایج بدست آمده به تفکیک نقاط شهری و روستایی در جداول ۱ و ۲ آمده است.

لازم به ذکر است استان قم در سال ۱۳۷۶ و استان‌های قزوین و گلستان در سال ۱۳۷۷ به جمع استان‌های کشور اضافه شده‌اند و به همین دلیل نتایج سالهای قبل در مورد آنها وجود ندارد. جدول ۲ ضرایب جینی نقاط روستایی استان‌های کشور را نمایش می‌دهد.

جدول ۱: ضریب جینی نقاط شهری استان‌های کشور

استان	سال ۷۴	سال ۷۵	سال ۷۶	سال ۷۷	سال ۷۸	سال ۷۹	سال ۸۰	متوسط سالهای ۷۴ تا ۸۰
کل کشور	۰/۴۰	۰/۴۰	۰/۳۹	۰/۳۹	۰/۳۹	۰/۴۰	۰/۴۰	۰/۳۹۶
آذربایجان شرقی	۰/۳۸	۰/۳۷	۰/۳۷	۰/۳۸	۰/۳۹	۰/۴۲	۰/۴۲	۰/۳۸۷
آذربایجان غربی	۰/۴۰	۰/۳۶	۰/۳۷	۰/۳۴	۰/۳۷	۰/۳۵	۰/۳۸	۰/۳۶۶
اردبیل	۰/۳۸	۰/۳۷	۰/۳۷	۰/۳۷	۰/۳۸	۰/۳۹	۰/۳۸	۰/۳۷۶
اصفهان	۰/۴۰	۰/۴۲	۰/۴۰	۰/۴۲	۰/۳۹	۰/۴۰	۰/۴۲	۰/۴۰۶
ایلام	۰/۳۶	۰/۳۳	۰/۳۶	۰/۳۶	۰/۳۹	۰/۳۳	۰/۳۴	۰/۳۵۲
بوشهر	۰/۳۴	۰/۳۳	۰/۲۵	۰/۲۶	۰/۳۸	۰/۳۵	۰/۳۴	۰/۳۱۷
تهران	۰/۳۷	۰/۳۷	۰/۳۸	۰/۳۵	۰/۳۷	۰/۳۷	۰/۳۸	۰/۳۶۹
چهارمحال و بختیاری	۰/۴۰	۰/۳۷	۰/۳۹	۰/۳۶	۰/۳۴	۰/۳۳	۰/۳۴	۰/۳۶۰
خراسان	۰/۴۳	۰/۴۰	۰/۳۹	۰/۳۸	۰/۳۷	۰/۴۳	۰/۴۰	۰/۴۰۰
خوزستان	۰/۳۳	۰/۳۲	۰/۳۲	۰/۳۵	۰/۳۶	۰/۳۳	۰/۳۰	۰/۳۳۰
زنجان	۰/۳۹	۰/۳۹	۰/۳۹	۰/۳۴	۰/۳۸	۰/۴۰	۰/۳۴	۰/۳۷۳
سمنان	۰/۳۹	۰/۳۶	۰/۴۰	۰/۳۳	۰/۳۳	۰/۳۲	۰/۳۳	۰/۳۵۰
سیستان و بلوچستان	۰/۳۹	۰/۳۵	۰/۳۴	۰/۳۷	۰/۴۰	۰/۴۲	۰/۴۳	۰/۳۸۶
فارس	۰/۳۹	۰/۳۳	۰/۳۸	۰/۳۶	۰/۳۶	۰/۳۶	۰/۳۴	۰/۳۵۸
قزوین	--	--	--	۰/۳۵	۰/۳۸	۰/۳۸	۰/۳۷	۰/۳۶۸
قم	--	--	۰/۳۱	۰/۳۲	۰/۳۳	۰/۲۸	۰/۴۵	۰/۳۳۵
کردستان	۰/۳۳	۰/۳۲	۰/۳۱	۰/۳۴	۰/۳۶	۰/۳۷	۰/۳۱	۰/۳۲۳
کرمان	۰/۳۷	۰/۳۷	۰/۳۷	۰/۳۴	۰/۳۹	۰/۳۹	۰/۳۶	۰/۳۷۱
کرمانشاه	۰/۳۷	۰/۳۵	۰/۳۱	۰/۳۰	۰/۳۴	۰/۳۶	۰/۴۱	۰/۳۴۷
کهکیلویه و بویراحمد	۰/۳۷	۰/۳۶	۰/۳۲	۰/۳۸	۰/۳۷	۰/۳۷	۰/۳۶	۰/۳۵۹
گلستان	--	--	--	۰/۳۶	۰/۳۸	۰/۴۵	۰/۴۴	۰/۴۰۵
گیلان	۰/۳۷	۰/۳۹	۰/۳۴	۰/۴۰	۰/۴۰	۰/۳۹	۰/۳۸	۰/۳۷۹
لرستان	۰/۳۲	۰/۳۵	۰/۳۲	۰/۳۴	۰/۳۳	۰/۳۴	۰/۳۵	۰/۳۳۵
مازندران	۰/۳۷	۰/۴۰	۰/۴۰	۰/۳۸	۰/۳۶	۰/۳۸	۰/۳۸	۰/۳۷۹
مرکزی	۰/۳۷	۰/۳۵	۰/۳۰	۰/۳۶	۰/۳۷	۰/۳۷	۰/۳۷	۰/۳۵۵
هرمزگان	۰/۳۳	۰/۳۶	۰/۳۰	۰/۳۵	۰/۴۶	۰/۴۳	۰/۳۶	۰/۳۲۵
همدان	۰/۴۳	۰/۳۸	۰/۴۱	۰/۴۴	۰/۴۶	۰/۴۳	۰/۴۰	۰/۴۲۰
یزد	۰/۴۱	۰/۴۵	۰/۴۸	۰/۴۴	۰/۴۱	۰/۳۷	۰/۳۸	۰/۴۲۰

جدول ۲: ضریب جینی نقاط روستایی استان‌های کشور

استان	سال ۷۴	سال ۷۵	سال ۷۶	سال ۷۷	سال ۷۸	سال ۷۹	سال ۸۰	متوسط سالهای ۷۴ تا ۸۰
کل کشور	۰/۴۳	۰/۳۹	۰/۴۱	۰/۴۲	۰/۴۲	۰/۴۲	۰/۴۱	۰/۴۱۴
آذربایجان شرقی	۰/۳۸	۰/۳۶	۰/۴۱	۰/۴۰	۰/۳۶	۰/۴۳	۰/۴۶	۰/۳۹۹
آذربایجان غربی	۰/۳۷	۰/۳۷	۰/۳۸	۰/۴۱	۰/۳۶	۰/۳۹	۰/۳۹	۰/۳۸۰
اردبیل	۰/۳۵	۰/۳۴	۰/۴۱	۰/۳۵	۰/۴۲	۰/۴۲	۰/۴۲	۰/۳۸۵
اصفهان	۰/۴۳	۰/۴۳	۰/۴۳	۰/۴۳	۰/۴۴	۰/۴۲	۰/۴۲	۰/۴۲۶
ایلام	۰/۳۲	۰/۳۱	۰/۳۰	۰/۳۶	۰/۳۲	۰/۳۴	۰/۳۴	۰/۳۲۶
بوشهر	۰/۴۰	۰/۳۶	۰/۳۴	۰/۴۱	۰/۴۰	۰/۳۵	۰/۳۷	۰/۳۷۲
تهران	۰/۳۶	۰/۳۲	۰/۳۵	۰/۳۰	۰/۳۳	۰/۳۸	۰/۳۴	۰/۳۳۹
چهار محال و بختیاری	۰/۴۱	۰/۳۸	۰/۴۱	۰/۴۲	۰/۳۸	۰/۳۹	۰/۳۴	۰/۳۹۰
خراسان	۰/۴۶	۰/۴۲	۰/۴۲	۰/۴۵	۰/۴۲	۰/۴۳	۰/۴۱	۰/۴۲۹
خوزستان	۰/۳۲	۰/۲۹	۰/۳۷	۰/۳۵	۰/۳۶	۰/۳۱	۰/۲۲	۰/۳۱۸
زنجان	۰/۴۲	۰/۳۸	۰/۴۳	۰/۴۴	۰/۴۶	۰/۴۴	۰/۴۲	۰/۴۲۶
سمنان	۰/۴۶	۰/۴۲	۰/۴۳	۰/۳	۰/۳۹	۰/۴	۰/۴۱	۰/۳۹۷
سیستان و بلوچستان	۰/۴۷	۰/۴۳	۰/۳۸	۰/۴۲	۰/۴۲	۰/۳۷	۰/۴۲	۰/۴۱۱
فارس	۰/۳۹	۰/۳۶	۰/۳۹	۰/۴۲	۰/۴۲	۰/۳۸	۰/۳۴	۰/۳۸۵
قزوین	--	--	--	۰/۴۳	۰/۴۲	۰/۴۲	۰/۴۲	۰/۴۲۸
قم	--	--	۰/۴۶	۰/۴۶	۰/۳۶	۰/۳۴	۰/۳۸	۰/۳۷۷
کردستان	۰/۳۳	۰/۳۲	۰/۳۲	۰/۴۲	۰/۳۷	۰/۳۸	۰/۳۷	۰/۳۶۶
کرمان	۰/۴۴	۰/۴۲	۰/۳۸	۰/۴۴	۰/۴۵	۰/۴۳	۰/۴۱	۰/۴۲۲
کرمانشاه	۰/۳۲	۰/۲۹	۰/۲۹	۰/۳۲	۰/۳۷	۰/۳۳	۰/۳۷	۰/۳۲۸
کهکیلویه و بویراحمد	۰/۴۵	۰/۴۵	۰/۴۷	۰/۳۴	۰/۴۳	۰/۴۱	۰/۳۹	۰/۴۱۸
گلستان	--	--	--	۰/۴۴	۰/۴۸	۰/۴۹	۰/۴۶	۰/۴۶۵
گیلان	۰/۴۲	۰/۴۱	۰/۴۴	۰/۴۷	۰/۴۴	۰/۴۳	۰/۴۱	۰/۴۳۲
لرستان	۰/۳۷	۰/۳۱	۰/۳۵	۰/۳۶	۰/۳۴	۰/۳۶	۰/۳۴	۰/۳۴۷
مازندران	۰/۴۵	۰/۴۱	۰/۴۳	۰/۴۳	۰/۴۲	۰/۴۱	۰/۳۹	۰/۴۱۸
مرکزی	۰/۴۱	۰/۳۹	۰/۳۹	۰/۴۱	۰/۳۹	۰/۴۴	۰/۴۴	۰/۴۰۶
هرمزگان	۰/۳۹	۰/۳۵	۰/۴۲	۰/۳۸	۰/۴۰	۰/۳۴	۰/۳۶	۰/۳۷۵
همدان	۰/۴۰	۰/۴۸	۰/۳۸	۰/۳۹	۰/۴۳	۰/۴۹	۰/۴۱	۰/۴۲۳
یزد	۰/۴۴	۰/۴۰	۰/۵۰	۰/۵۵	۰/۴۵	۰/۴۵	۰/۴۶	۰/۴۶۲

جدول ۳: خوشه‌بندی استان‌های کشور با استفاده از ضریب جینی نقاط شهری

خوشه اول	بوشهر، خوزستان، قم، کردستان، کرمانشاه، لرستان، هرمزگان
خوشه دوم	آذربایجان شرقی، آذربایجان غربی، اردبیل، ایلام، تهران، چهارمحال و بختیاری، زنجان، سمنان، سیستان و بلوچستان، فارس، قزوین، کرمان، کهکیلویه و بویراحمد، گیلان، مازندران، مرکزی
خوشه سوم	اصفهان، خراسان، گلستان، همدان، یزد

۵ خوشه‌بندی استان‌های کشور بر اساس برآورد ضرایب جینی و تحلیل نتایج بدست آمده

در این بخش با استفاده از روش آنالیز خوشه‌ای[4] استان‌های کشور با استفاده از متوسط ضریب جینی نقاط شهری و روستایی طی سالهای ۷۴ تا ۸۰ در ۳ گروه قرارگرفته‌اند تا استان‌های مشابه از نظر چگونگی توزیع درآمد خانوارهای شهری و روستایی مشخص گردند. نتایج به تفکیک نقاط شهری و روستایی در جداول ۳ و ۴ آمده است.

بر اساس ضرایب جینی بدست آمده برای نقاط شهری استان‌های کشور می‌توان گفت توزیع درآمد کل کشور(نقاط شهری) در حالت نیمه متعادل است و این مورد برای استان‌ها نیز به همین گونه می‌باشد، با این وجود بین برخی از استان‌های کشور از نظر چگونگی توزیع درآمد خانوارهای شهری تفاوتهایی وجود دارد که این تفاوت را می‌توان با خوشه‌بندی آنها بیشتر مشخص کرد. براساس این خوشه بندی استان‌های کشور در ۳ خوشه واقع شده‌اند، متوسط ضریب جینی استان‌های خوشه اول برابر ۰/۳۳، خوشه دوم ۰/۳۷ و خوشه سوم ۰/۴۱ می‌باشد. در این بین استان بوشهر عادلانه‌ترین توزیع درآمد نقاط شهری را داشته و استان‌های همدان و یزد نا عادلانه‌ترین وضعیت را دارا بوده‌اند.

بر اساس ضرایب جینی بدست آمده برای نقاط روستایی استان‌های کشور می‌توان گفت توزیع درآمد کل کشور(نقاط روستایی) در حالت نیمه متعادل است و این مورد برای استان‌ها نیز به همین گونه می‌باشد، با این وجود بین برخی از استان‌های کشور از نظر چگونگی توزیع درآمد خانوارهای روستایی تفاوتهایی وجود دارد که این تفاوت را می‌توان با خوشه‌بندی آنها بیشتر مشخص کرد. براساس این خوشه بندی استان‌های کشور در ۳ خوشه واقع شده‌اند متوسط ضریب جینی استان‌های خوشه اول برابر ۰/۳۴، خوشه دوم ۰/۴۰ و خوشه سوم ۰/۴۶ می‌باشد.

4) Cluster Analysis

جدول ۴: خوشه‌بندی استان‌های کشور با استفاده از ضریب جینی نقاط روستایی

خوشه اول	ایلام، تهران، خوزستان، کردستان، کرمانشاه، لرستان
خوشه دوم	آذربایجان شرقی، آذربایجان غربی، اردبیل، اصفهان، چهارمحال و بختیاری بوشهر، خراسان، زنجان، سمنان، سیستان و بلوچستان، فارس، قزوین، قم، کرمان کهکیلویه و بویر احمد، گیلان، مازندران، مرکزی، هرمزگان، همدان
خوشه سوم	گلستان، یزد

در این بین استان خوزستان عادلانه‌ترین توزیع درآمد نقاط روستایی را داشته و استان‌های گلستان ناعادلانه‌ترین وضعیت را دارا بوده‌اند.

از نتایج دیگری که می‌توان از برآورد ضریب جینی گرفت این است که به‌طورکلی توزیع درآمد در نقاط روستایی کشور ناعادلانه‌تر از توزیع درآمد در نقاط شهری است و این موضوع در اکثر استان‌های کشور نیز دیده می‌شود و تنها در استان‌های ایلام، تهران، خوزستان و کرمانشاه توزیع درآمد خانوارهای روستایی عادلانه‌تر از نقاط شهری است.

مراجع

[۱] نتایج آمارگیری از هزینه و درآمد خانوارهای شهری، (۱۳۷۴ تا ۱۳۸۰)، مرکز آمار ایران

[۲] نتایج آمارگیری از هزینه و درآمد خانوارهای روستایی، (۱۳۷۴ تا ۱۳۸۰)، مرکز آمار ایران

[3] Kokwani, N.C (1980), Incomme Inequality and poverty, A World Bank Research Publication.

[4] Khan, Azizur Rahman (1993), Structural Adjustment and Income distribution, ILO.

شرایط آغازین برای مدلهای تصادفی ـ انتقالی در داده‌های طولی با پاسخ دودویی کامل و ناقص

زهرا رضایی قهرودی[۱]، مجتبی گنجعلی[۲]

[۱] کارشناس ارشد آمار دانشگاه شهید بهشتی
[۲] عضو هیأت علمی دانشگاه شهید بهشتی

چکیده: فرض کنید برای داده‌های طولی با پاسخ دودویی مدلی را استفاده کنیم که در آن پاسخ در زمان t به پاسخهای زمانهای گذشته و نیز به عوامل طبیعی غیرقابل مشاهده شده به عنوان اثرهای تصادفی وابسته است. چنین مدلی مدل انتقالی ـ تصادفی نامیده می‌شود. همچنین فرض کنید فرایند در زمان صفر شروع ولی در زمان $J \geq ۱$ مشاهده شده باشد. بنابراین برای فرد iام فرض کنید مشاهدات در زمانهای $J, J+۱, ..., T$ مشاهده شده است. از آنجا که برای مثال در مدل مارکوف مرتبهٔ اول پاسخ در زمان J توسط مدل انتقالی ـ تصادفی به پاسخ در زمان $J-۱$ وابسته است ولی پاسخ $J-۱$ مشاهده نمی‌شود، مسئلهٔ شرایط آغازین ایجاد می‌شود. در این مقاله با استفاده از مفهوم متغیر پنهان برای پاسخ دودویی یک فرایند مارکوف مرتبهٔ اول در نظرگرفته و فرض می‌کنیم که اختلالی که فرایند را ایجاد می‌کند به دو قسمت خطای اندازه‌گیری و تاثیر تصادفی تجزیه می‌شود. روشهای مختلف مواجهه با شرایط آغازین را در دو مورد (i) پاسخهای دودویی کامل و (ii) پاسخهای دودویی ناقص مورد بحث و بررسی قرار می‌دهیم.

واژه‌های کلیدی: داده های طولی، مدل تصادفی ـ انتقالی، فرایند مارکوف مرتبهٔ اول، شرایط آغازین

مقدمه

اخیرا مطالعات طولی از اهمیت به سزائی برخوردار شده است. از مشخصه‌های بارز مطالعات طولی اندازه‌های مکرر برای افراد مختلف در طول زمان است. در مطالعات طولی همبستگی بین اندازه‌ها در طول زمان مورد بررسی قرار می‌گیرد. در داده‌های طولی امکان گم شدن داده‌ها نیز وجود دارد. در این مقاله هدف بررسی مسئلهٔ شرایط آغازین در مدل های اثرات تصادفی ـ انتقالی با پاسخ دودویی کامل و ناقص است. فرض می‌کنیم فرایند در زمان صفر شروع ولی در زمان $J \geq ۱$ مشاهده شده باشد. بنابراین برای فرد iام فرض کنید مشاهدات در زمانهای $J, J+۱, ..., T$ مشاهده شده است. از آنجا که برای مثال در مدل مارکوف مرتبهٔ اول پاسخ در زمان J توسط مدل انتقالی ـ تصادفی به پاسخ در زمان $J-۱$ وابسته است ولی پاسخ

$J-۱$ مشاهده نمی‌شود، مسئلهٔ شرایط آغازین ایجاد می‌شود. با توجه به اینکه بحث ما جنبه‌های ضروری مسئلهٔ شرایط آغازین و راه‌حل‌های آن می‌باشد، در این مقاله روش‌های مختلف مواجهه با مقادیر مشاهده نشده (در داده‌های کامل و ناقص) که ممکن است دنباله‌ای از y_{it} برای $t = ۰, \ldots, J-۱$ باشد و یا در حالت خاص تنها $y_{i۰}$ $(J = ۱)$ باشد را مورد بحث و بررسی قرار می‌دهیم.

۱ گم شدن داده‌ها در مطالعات طولی

در داده‌های طولی گم شدن داده‌ها به دو صورت الگوی عمومی و الگوی یکنوا که همان انصراف است طبقه‌بندی می‌شود. در الگوی عمومی هیچ روند خاصی مشاهده نمی‌شود و مقادیر گمشده در هر زمانی می‌تواند رخ دهد. بالاخص آزمودنی خاصی می‌تواند از مطالعه در زمان خاصی خارج و در زمان دیگری دوباره به مطالعه بازگشت کند. ولی انصراف در داده‌های طولی به این معناست که وقتی آزمودنی در زمان t پاسخ نمی‌دهد هیچوقت به مطالعه بازگشت ندارد و ما پاسخهای زمانهای $t+۱, \ldots$ را نیز از دست می‌دهیم. به منظور مدل‌بندی همزمان انصراف و پاسخ از دیدگاه تابع درستنمایی تعاریف اساسی را که اولین بار در رابین (۱۹۷۶)[1] مطرح شده است، می‌آوریم. فرض کنید در صورت عدم داشتن دادهٔ گمشده برای پاسخهای آزمودنی iام پاسخهای دودویی y_{ij} برای $j = ۱, ۲, \ldots, T$ مشاهده شود.

برای نشان دادن آن که متغیر y گمشده است از تابع نشانگر R استفاده می‌کنیم که به‌صورت زیر تعریف می‌شود.

$$y_{it} = \begin{cases} ۱ & \text{اگر پاسخ فرد } i\text{ام در زمان } t \text{ مشاهده شده باشد} \\ ۰ & \text{در غیر این صورت} \end{cases}$$

همچنین پاسخ دودویی y_{it} را برآمده از یک متغیر پنهان y^*_{it} فرض می‌کنیم، که به صورت زیر تعریف می‌شود.

$$y_{it} = \begin{cases} ۱ & y^*_{it} > c \\ ۰ & \text{در غیر اینصورت} \end{cases}$$

برای درک مفهوم متغیر پنهان فرض کنید که تنها ثبت شده‌است که آزمودنی چاق یا لاغر است. اما اگر فردی چاق $(y_{it} = ۱)$ است این به خاطر وزن زیاد $(y^*_{it} > c)$ او بوده برای بردار تصادفی $Y_i = (Y_{i۱}, \ldots, Y_{in})^T$ بردار نشانگر پاسخ $R_i = (R_{i۱}, \ldots, R_{in})^T$ رااستفاده می‌کنیم.

1) Rubin

۲ مدل اثرهای تصادفی و انتقالی

این مدل ترکیبی از مدل اثرهای تصادفی و مدل انتقالی است به این صورت که پاسخ در زمان t به پاسخ در زمانهای گذشته و نیز به عوامل طبیعی غیر قابل مشاهده شده به عنوان اثرهای تصادفی وابسته است. به عنوان مثال برای بررسی علت تصادف برای آزمودنی‌ها باید این مسئله را بررسی کنیم که آیا علت تصادف به دلیل شتابزدگی فرد (اثر تصادفی) و یا به دلیل عادت فرد به تصادف (پاسخهای گذشته) بوده است. برای بیان مدلی مناسب به صورت زیر عمل می‌کنیم. فرض کنید که $y_i = (y_{i1}, \ldots, y_{iT_i})^T$ پاسخهای دودوئی برای n آزمودنی $(i = 1, \ldots, n)$ در T_i موقعیت متفاوت $(t = 1, \ldots, T_i)$ با بردار p بعدی متغیرهای تبیینی $x_{it} = (x_{it1}, \ldots, x_{itp})^T$ که به زمان وابسته است فرض شود. مدل زیر مدلی مناسب برای بررسی اثرهای تصادفی و پاسخهای گذشته است.

$$Y_{it}^* = x_{it}'\beta + z_{it}'\tau_i + \sum_{j=1}^{q} \gamma_j y_{i,t-j} + \varepsilon_{it}$$

که در آن Y_{it}^* متغیر پنهانی متناظر با Y_{it} است که قبلاً معرفی شده است و $y_{i,t-j}$ پاسخ آزمودنی iام در زمان $t-j$ است که تاثیر آن بر y_{it} توسط پارامتر γ_j سنجیده می‌شود. β بردار p بعدی ضرایب متغیرهای تبیینی و $\tau_i = (\tau_{i1}, \ldots, \tau_{is})^T$ اثرهای تصادفی برای $i = 1, \ldots, n$، با تابع چگالی چندمتغیره معلوم $g(.;\phi)$ فرض می‌شوند. همچنین z_{it} زیر مجموعه‌ای از x_{it} است.

در برخورد با رویهٔ انصراف، برای مدل‌بندی مکانیسم گم شدن، به روشهای مختلف می‌توان عمل کرد. مدل مکانیسم گم شدن که الگوی گم‌شدن انصراف و تناوبی را در نظر می‌گیرد به صورت زیر بیان می‌شود.

$$R_{it}^* = v_{it}'\alpha_l + \theta_2 \tau_i + \varepsilon_{it}$$

که R_{it}^* متغیر پنهان برای مدل مربوط به مکانیسم گم شدن است. α_l بردار پارامتری حاصل از تاثیر متغییرهای تبیینی بر مکانیسم گمشدن و θ_2 پارامتر منعکس کنندهٔ تاثیر تصادفی بر مکانیسم گمشدن است. این مدل به مدل تاثیر تصادفی مشترک (۱۹۹۵)[2] موسوم است.

۳ مسئلهٔ شرایط آغازین و بعضی راه‌حل های معمول

با توجه به اینکه بحث ما جنبه‌های ضروری مسئلهٔ شرایط آغازین و راه‌حل‌های آن می‌باشد، بنابراین یک فرآیند مارکوف مرتبهٔ اول را درنظر می‌گیریم و متغیرهای برون‌زا (متغیرهای تبیینی) را فعلاً در نظر نمی‌گیریم. به علاوه فرض شده است که اختلالی که فرآیند را ایجاد می‌کند به دو قسمت

2) Follmann D. and Wu M.

خطای اندازه‌گیری و تاثیر تصادفی تجزیه می‌شود. فرآیند به وسیلهٔ یک متغیر پنهان دودویی Y_{it}^* تعریف شده است که از مدل زیرین تبعیت می‌کند:

$$Y_{it}^* = \beta_۰ + \gamma y_{i,t-۱} + \varepsilon_{it} \quad i = ۱, ..., n, \quad t = ۱, ..., T \quad (۱۰۳)$$

که در آن

$$\varepsilon_{it} = \tau_i + U_{it}$$

و

$$E(U_{it}) = ۰ = E(\tau_i),\ E(\tau_i^۲) = \sigma_\tau^۲,\ E(U_{it}^۲) = \sigma_u^۲ = ۱$$
$$E(U_{it}U_{i't''}) = ۰,\ \ t \neq t'',\ E(\tau_i U_{i't''}) = ۰ \qquad \forall i, i', t''$$

که در آن τ_i اثر تصادفی و U_{it} خطای مدل است و احتمالهای تغییر وضعیت برای فرد iام در زمان tام به شرط τ_i به صورت زیر است.

$$f(y_{it}|y_{it-۱}, \tau_i) = \Phi\{[\beta_۰ + \gamma y_{i,t-۱} + \tau_i].[۲y_{it} - ۱]\}$$

که در آن Φ تابع توزیع تجمعی نرمال است. در اینجا از یک مدل پروبیت استفاده می‌شود، اما بدون از دست دادن هیچ کلیتی مدل لوژستیک نیز می‌تواند مورد بررسی قرار گیرد. اگر شروع مشاهدهٔ فرآیند در زمان J باشد به این معنی که شروع اولیهٔ فرآیند زمان ۰ بوده و محقق مشاهدات خود را از زمان J ثبت کرده است، مسئلهٔ شرایط آغازین ایجاد می‌شود. در حالتی که $J > ۱$ باشد، $f(y_{iJ}|\tau_i)$ نشان می‌دهد که چگونه مسئلهٔ مقادیر اولیه به وجود می‌آید. توجه کنید که $f(y_{iJ}|\tau_i)$ به تمام دنباله‌های ممکن پیشامدهای قبل از زمان Jام که مشاهده نمی‌شود و منجر به یک مقدار خاص برای $y_{i,J}$ می‌شود وابسته است. لازم به ذکراست که عموماً محققین مقادیر متغیرهای تبیینی مناسب را در دوره‌های زمانی $t = ۱, ..., J-۱$ نیز نمی‌دانند.

اکنون پنج روش مختلف مواجهه با مقادیر مشاهده نشده که ممکن است دنباله‌ای از y_{it} برای $t = ۰, ..., J-۱$ باشد و یا در حالت خاص $(J = ۱)$ $(y_{i۰}$ مشاهده نشده است) را که ما در اینجا در نظر گرفته‌ایم، بکار می‌بریم. در هر حالت سعی شده‌است که تابع درستنمایی برای بهینه‌سازی داده شود.

روش اول:

اگر $J = ۱$ باشد، $y_{i۱}$ را یک مقدار ثابت غیرتصادفی و معلوم درنظر می‌گیریم و تابع درستنمایی را به شرط $y_{i۱}$ به صورت زیر می‌نویسیم یعنی:

$$\prod_{i=۱}^{n} f(y_{i۲}, ..., y_{iT_i}|y_{i۱}) = \prod_{i=۱}^{n} \int \prod_{t=۲}^{T_i} f(y_{it}|\tau_i, y_{i,t-۱}) g(\tau_i) d\tau_i \quad (۲۰۳)$$

که در این روش عملاً اثر پاسخهای مشاهده شدهٔ زمان ۱ به عنوان متغیر تبیینی در نظر گرفته می‌شود و از یک مدل مارکوف مرتبهٔ اول استفاده شده است.

روش دوم:
حالتی است که ایتکن و آلفو[3] (۲۰۰۰) معرفی کرده‌اند که آنها نیز مانند روش اول عمل می‌کنند با این تفاوت که پاسخ مشاهده شدهٔ y_{i1} اثر تصادفی را نیز تحت تاثیر قرار می‌دهد. تابع درستنمایی ارائه شده توسط ایتکن و آلفو (۲۰۰۰) به صورت زیر است.

$$\prod_{i=1}^{n} f(y_{i2},...,y_{iT_i}|y_{i1}) = \prod_{i=1}^{n} \int \prod_{t=2}^{T_i} f(y_{it}|\tau_i, y_{i,t-1}) g(\tau_i|y_{i1}) d\tau_i \quad (3\cdot 3)$$

که در آن همانطور که مشاهده می‌کنید توزیع شرطی اثر تصادفی به شرط y_{i1} بایستی در نظر گرفته شود.

روش سوم:
روشی است که برای پاسخ در زمان ۱ تابع توزیع مجزایی ارائه دهیم، که در این روش به صورت زیر عمل می‌شود.

$$\prod_{i=1}^{n} f(y_{i1},...,y_{iT_i}) = \prod_{i=1}^{n} \int f(y_{i1}|\tau_i) \prod_{t=2}^{T_i} f(y_{it}|\tau_i, y_{i,t-1}) g(\tau_i) d\tau_i \quad (4\cdot 3)$$

توجه کنید که در رابطهٔ فوق $f(y_{i1}|\tau_i)$ نیز مدل‌بندی می‌شود.

روش چهارم:
اگر فرآیند تعادلی باشد، احتمالهای حاشیه‌ای برای حالت $y_{it}=1$ برای همهٔ tها (با فرض گذشتهٔ بی‌نهایت) به صورت زیر است.

$$\Pi_1(\tau_i) = \frac{\Phi(\beta_0 + \tau_i)}{1 - \Phi(\beta_0 + \gamma + \tau_i) + \Phi(\beta_0 + \tau_i)} \quad (5\cdot 3)$$

و احتمال حاشیه‌ای برای حالت $y_{it}=0$، $\Pi_0(\tau_i) = 1 - \Pi_1(\tau_i)$ است. که Π_1 و Π_0 به ترتیب احتمالهای حاشیه‌ای در حالت پاسخ ۱ و ۰ است. لازم به ذکر است که احتمالهای تعادلی به صورت زیر به دست می‌آیند:

$$P(Y_{it}=1|\tau_i) = \sum_{y_{i,t-1}} P(Y_{it}=1|Y_{i,t-1}=y_{i,t-1},\tau_i) P(Y_{i,t-1}=y_{i,t-1}|\tau_i)$$

که چون در شرایط تعادلی هستیم، خواهیم داشت

$$P(Y_{i,t-1}=y_{i,t-1}|\tau_i) = P(Y_{it}=y_{it}|\tau_i)$$

3) Aitkin and Alfo

با استفاده از این واقعیت که $\Pi_{\circ}(\tau_i) = ۱ - \Pi_{۱}(\tau_i)$ است، خواهیم داشت:

$$\begin{aligned} P(Y_{it} = ۱|\tau_i) =& P(Y_{it} = ۱|Y_{i,t-۱} = \circ, \tau_i)[۱ - \Pi_{۱}(\tau_i)] \\ &+ P(Y_{it} = ۱|Y_{i,t-۱} = ۱, \tau_i)\Pi_{۱}(\tau_i) \end{aligned}$$

یعنی

$$\Pi_{۱}(\tau_i) = \Phi(\beta_{\circ} + \tau_i)\Pi_{\circ}(\tau_i) + \Phi(\beta_{\circ} + \gamma + \tau_i)\Pi_{۱}(\tau_i)$$

و بنابراین اگر فرآیند در وضعیت تعادل باشد،

$$P(Y_{iJ} = y_{iJ}|\tau_i) = \Pi_{۱}(\tau_i)^{y_{iJ}}\Pi_{\circ}(\tau_i)^{۱-y_{iJ}}$$

است. روش دیگر مواجهه با شرایط آغازین استفاده از احتمالات تعادلی است.

روش پنجم:

روش آخر که هکمن[4] (۱۹۸۱) پیشنهاد کرده است حالتی است که در آن برای پاسخ‌ها در زمان $t = ۱$ اثر تصادفی‌ای متفاوت با اثر تصادفی در زمانهای $t = ۲, \ldots, T_i$ در نظر می‌گیریم. مدل ارائه شده توسط هکمن (بدون حضور متغیرهای تبیینی) به صورت زیر است:

$$Y_{it}^{*} = \beta_{\circ} + \gamma y_{i,t-۱} + \varepsilon_{it} \qquad i = ۱, \ldots, n, t = ۲, \ldots, T$$

که در آن

$$\varepsilon_{it} = \tau_{i۱} + U_{it}, \qquad t = ۲, ..., T$$

$$Y_{i۱}^{*} = \beta_{\circ} + \varepsilon_{i۱}$$

و

$$\varepsilon_{i۱} = \tau_{i۲} + U_{i۱}$$

و $\tau_{i۱}$ و $\tau_{i۲}$ اثرات تصادفی با میانگین صفر هستند که به طور نرمال توزیع شده است. همچنین $\tau_{i۱}$ با $\tau_{i۲}$ همبسته‌اند. بنابراین تابع درستنمایی ارائه شده توسط هکمن به صورت زیر است:

$$\begin{aligned} \prod_{i=۱}^{n} f(y_{i۱}, ..., y_{iT_i}) &= \prod_{i=۱}^{n} \int \int f(y_{i۱}|\tau_{i۲}) \\ &\times \prod_{t=۲}^{T_i} f(y_{it}|\tau_{i۱}, y_{i,t-۱}) g(\tau_{i۱}, \tau_{i۲}) d\tau_{i۱} d\tau_{i۲} \qquad (۶ \cdot ۳) \end{aligned}$$

که برآورد پارامترهای مدل به روش ماکسیمم درستنمایی با درنظرگرفتن همبستگی بین $\tau_{i۲}$ و $\tau_{i۱}$ به روش مربع بندی گاوسی به‌دست می‌آید. هکمن (۱۹۸۱) در یک مطالعهٔ شبیه‌سازی برتری این روش را نسبت به برخی روشهای دیگر مورد بررسی قرار داده است. در بخش بعد

4) Heckman

مسئلهٔ شرایط آغازین را با درنظرگرفتن پاسخ گمشده در پنج روش فوق مورد بحث و بررسی قرار می‌دهیم.

مسئلهٔ دیگری که در شرایط آغازین مطرح است، مسئلهٔ پارامترهای فرعی است. در این حالت اثرهای ثابت τ_i را به عنوان پارامتر درنظر می‌گیریم و تابع درستنمایی شرطی $y_{i2}, \ldots, y_{iT}$ به شرط y_{i1} که خود y_{i1} تابعی از پاسخهای قبلی مشاهده نشده است، را به صورت زیر بیان می‌کنیم:

$$f(y_{i2}, ..., y_{iT}|y_{i1}) = \prod_{i=1}^{n}\prod_{t=2}^{T} \frac{\Phi\{[\beta_0 + \gamma y_{i,t-1} + \tau_i][2y_{it} - 1]\} \times P(y_{i1}|\tau_i)}{P(y_{i1}|\tau_i)}$$
$$= \prod_{i=1}^{n}\prod_{t=2}^{T} \Phi\{[\beta_0 + \gamma y_{i,t-1} + \tau_i][2y_{it} - 1]\} \qquad (۷.۳)$$

همانگونه که ملاحظه می‌شود توزیع $f(y_{i1}|\tau_i)$ از معادلهٔ بالا حذف می‌شود و این نشان می‌دهد که استفاده از روش پارامترهای فرعی، راه‌حلی برای رفع شرایط آغازین است. این مدل یک مدل رگرسیون خطی با اثرات ثابت است و محاسبهٔ آن بسیار راحت است. با ماکسیمم کردن تابع (۷.۳) نسبت به τ_i برای $i = 1, ..., n$، $\hat{\beta}$ و $\hat{\gamma}$ زمانی که n و T بزرگند برآوردهای سازگاری هستند (هکمن، ۱۹۸۱). در حالتی که T کوچک است زیاد بودن پارامترها مشکلی است که این روش با آن مواجه است.

۴ شرایط آغازین برای مدلهای با پاسخ دودویی و ناقص

همانطور که در بخش قبل مسئلهٔ شرایط آغازین را بیان کردیم، در این بخش نیز به این مسئله می‌پردازیم با این تفاوت که در این بخش پاسخهای گمشده در داده‌های ما وجود دارد. مدل در نظرگرفته شده برای پاسخ مدل مارکوف مرتبهٔ اول همراه با اثر تصادفی است و مدلی که برای پاسخ‌های گمشده درنظرگرفته می‌شود مدل پیشنهادی دیگل وکنورد (۱۹۹۴) است که مکانیسم گم شدن در زمان t به پاسخهای زمانهای t و $t-1$ بستگی دارد با این تفاوت که در مدل دیگل وکنورد اثرهای تصادفی را نیز وارد مدل می‌کنیم. مدل فوق به صورت زیر بیان می‌شود.

$$Y_{it}^* = \beta + \gamma_1 y_{i,t-1} + \varepsilon_{it1} \qquad i = 1, ..., n, \quad t = 1, ..., T \qquad (۱۰۴)$$

که در آن

$$\varepsilon_{it1} = \theta_1 \tau_i + U_{it1}$$

برای مکانیسم گم شدن داریم:

$$R_{it}^* = \alpha + \psi_{12} y_{i,t-1} + \psi_{22} y_{it} + \varepsilon_{it2} \qquad i = 1, ..., n, \quad t = 2, ..., T$$

که در آن فرض شده برای پاسخ در زمان ۱ مشاهدهٔ گمشده نداریم و $\psi_{۱۲}$ و $\psi_{۲۲}$ اثر پاسخهای قبلی و جاری را برگمشدن پاسخ در زمان جاری منعکس می‌کنند. همچنین

$$\varepsilon_{it۲} = \theta_۲\tau_i + U_{it۲}$$

و داریم:

$$E(U_{it۱}) = E(U_{it۲}) = ۰ = E(\tau_i), \quad E(\tau_i^۲) = ۱$$

$$E(U_{it۱}^۲) = E(U_{it۲}^۲) = \sigma_u^۲ = ۱, E(U_{it۱}U_{i't''۱}) = E(U_{it۲}U_{i't''۲}) = ۰, \quad t \neq t''$$

$$E(\tau_i U_{i't''۱}) = E(\tau_i U_{i't''۲}) = ۰ \qquad \forall i, t', t''$$

که در آن τ_i اثرهای تصادفی و $U_{it۱}$ و $U_{it۲}$ به ترتیب خطاهای مدل پاسخ و مکانیسم گم شدن هستند. توجه داشته باشید که مکانیسم گمشدن در صورتی کاملاً تصادفی است که پارامترهای $\psi_{۱۲}$، $\psi_{۲۲}$ و $\theta_۲$ معنی‌دار نباشند. احتمالهای تغییر وضعیت برای فرد iام در زمان tام به شرط τ_i به صورت زیر است.

$$f(y_{it}|y_{i,t-۱}, \tau_i) = \Phi\{[\beta + \gamma_۱ y_{i,t-۱} + \tau_i][۲y_{it} - ۱]\}$$

و

$$P(R_{it}|y_{i,t-۱}, y_{it}, \tau_i) = \Phi\{[\alpha + \psi_{۱۲}y_{i,t-۱} + \psi_{۲۲}y_{it} + \theta\tau_i][۲R_{it} - ۱]\}$$

که در آن Φ تابع توزیع تجمعی نرمال است.

در مواجهه با مسئله شرایط آغازین وقتی داده‌های گمشده داریم ممکن است به پنج روش برخورد کنیم. در زیر این روشها برای حالتی که فرآیند در زمان ۱ شروع به مشاهده شدن می‌کند و مشاهدات قبل از این در دسترس نیست، مورد بررسی قرارگرفته‌اند. در هر حالت تابع درستنمایی برای بهینه‌سازی داده شده است.

روش اول:

در این روش پاسخ $y_{i۱}$ را به عنوان یک پاسخ ثابت و معلوم در نظر می‌گیریم، تابع درستنمایی به شکل زیر است.

$$\prod_{i=۱}^{n} f(y_{i۲}, ..., y_{iT}, R_{i۲}, ..., R_{iT}|y_{i۱})$$
$$= \prod_{i=۱}^{n} \int \prod_{t=۲}^{T} f(y_{it}|\tau_i, y_{i,t-۱}) P(R_{it}|y_{it}, y_{i,t-۱}, \tau_i) g(\tau_i) d\tau_i \quad (۲۰۴)$$

که با توجه به تابع چگالی‌های ارائه شده برای پاسخ و مکانیسم گم شدن، تابع درستنمایی فوق را می‌توان ماکسیمم نمود.

روش دوم:
روش دوم مانند روش اول است با این تفاوت که در آن اثر تصادفی به y_{i1} وابسته است که تابع درستنمایی در این روش به شکل زیر است.

$$\prod_{i=1}^{n} f(y_{i2},...,y_{iT},R_{i2},...,R_{iT}|y_{i1})$$
$$=\prod_{i=1}^{n}\int\prod_{t=2}^{T} f(y_{it}|\tau_i,y_{i,t-1})P(R_{it}|y_{it},y_{i,t-1},\tau_i)g(\tau_i|y_{i1})d\tau_i \quad (۳\cdot۴)$$

روش سوم:
روشی است که در آن توزیع مجزایی را برای پاسخ مشاهده شدهٔ y_{i1} درنظرگیریم. در این حالت تابع درستنمایی به شکل زیر است.

$$\prod_{i=1}^{n} f(y_{i1},...,y_{iT},R_{i1},...,R_{iT})$$
$$=\prod_{i=1}^{n}\int f(y_{i1}|\tau_i)\prod_{t=2}^{T} f(y_{it}|\tau_i,y_{i,t-1})P(R_{it}|y_{it},y_{i,t-1},\tau_i)g(\tau_i)d\tau_i \quad (۴\cdot۴)$$

که در آن $f(y_{i1}|\tau_i)$ می‌تواند جداگانه مدل‌بندی شود.

روش چهارم:
اگر فرآیند در وضعیت تعادلی باشد آنگاه خواهیم داشت:

$$f(y_{i1}|\tau_i)=\Pi_1(\tau_i)^{y_{i1}}\Pi_0(\tau_i)^{1-y_{i1}} \qquad (۵\cdot۴)$$

که در آن Π_1 و Π_0 به ترتیب احتمالهای حاشیه ای برای $y_{it}=1$ و $y_{it}=0$ است و می‌توان از این توزیع در (۴.۴) استفاده کرد.

روش پنجم:
نهایتاً روش آخر استفاده از مدل هکمن است که با کمی تغییر و با درنظرگرفتن مکانیسم گم شدن در تابع درستنمایی، طبق مدل درنظرگرفته شدهٔ زیر خواهیم داشت:

$$Y_{it}^{*}=\beta+\gamma_1 y_{i,t-1}+\varepsilon_{it} \qquad i=1,...,n,t=2,...,T$$

$$\varepsilon_{it}=\tau_{i1}+U_{it}$$

و

$$Y_{i1}^{*}=\beta+\varepsilon_{i1}$$

که در آن

$$\varepsilon_{i1} = \tau_{i2} + U_{i2}$$

و

$$R^*_{it} = \alpha + \psi_{12} y_{i,t-1} + \psi_{22} y_{it} + \varepsilon_{it2} \qquad t = 2, ..., T$$

اگر فرض کنیم ε_{it3} از ε_{it2} و ε_{it1} مستقل است، تابع درستنمایی مدل فوق به صورت زیر است.

$$\prod_{i=1}^{n} f(y_{i1}, ..., y_{iT}, R_{i2}, ..., R_{iT}) = \prod_{i=1}^{n} \int\int f(y_{i1}|\tau_{i2}) \prod_{t=2}^{T_i} f(y_{it}|\tau_{i1}, y_{i,t-1}) \times P(R_{it}|y_{it}, y_{i,t-1}) g(\tau_{i1}, \tau_{i2}) d\tau_{i1} d\tau_{i2} \qquad (۶\cdot۴)$$

به عنوان کار بعدی مقایسه‌ای عمومی بین این روشها پیشنهاد می‌شود. می‌توان همچون هکمن (۱۹۸۱) که مدل با استفاده از روشهای (۱)، (۲)، (۳) و (۵) را در داده‌های کامل با استفاده از شبیه‌سازی مونت کارلو ، مقایسه کرده و به این نتیجه رسیده است که روش (۵) نتایج بهتری ارایه می‌دهد، روش (۴) را نیز با این روشها مقایسه نمود. همچنین مقایسه‌ای بین روشهای ارایه شده برای داده‌های با پاسخ‌های گمشده، پیشنهاد می‌شود. برای ملاحظهٔ چگونگی کاربرد برخی از روشهای ذکر شده در داده‌های واقعی به رضایی (۱۳۸۲) مراجعه کنید.

مراجع

[1] Aitkin, M. and Alfo, M. (2000). Random coefficient models for binary longitudinal responses with attrition. Statistics and Computing. 10, 279-287.

[2] Diggle, P. and Kenword, M.G. (1994). Informative drop-out in longitudinal data analysis (with discussion). Appl. Statist., 43, 49-93.

[3] Follmann D. and Wu M. (1995). An approximate generalized linear model with random effects for informative missing data. Biometrics 51: 151-168.

[4] Heckman, J., (1981). Statistical models for discrete panel data, in Manski, C. & McFadden D., Structural analysis of discrete data with econometric applications. 114-195, Cambridge, Mass: MIT press.

[5] Rubin D. B. (1976). Inference and missing data. Biometrica 63, 581-592.

[۶] رضایی قهرودی، ۱۳۸۲، تحلیل داده‌های طولی دودویی ناشی از متغیرهای پنهان با پاسخهای گمشده، دانشگاه شهید بهشتی.

تحلیل ممیزی با استفاده از آمیخته‌های نرمال

قاسم رکابدار[۱]، رحیم چینی‌پرداز[۲]

[۱] دانشگاه آزاد اسلامی واحد آبادان، خرمشهر
[۲] دانشگاه شهید چمران، گروه آمار

چکیده: تحلیل ممیزی خطی فیشر[۱] (LDA) ابزار مهمی برای رده‌بندی بین چندگروه می‌باشد. هنگامیکه کلاسها دارای توزیع نرمال چند متغیره با ماتریس کواریانس مشترک باشند طبق لم نیمن-پیرسن LDA بهینه بوده و معادل با قاعده رده‌بندی ماکزیمم درستنمایی است. LDA دارای معایبی است از جمله فرض نرمال بودن کلاسها اغلب بندرت پیش می‌آید وکرانهای تصمیم خطی، کلاسها را به قدر کافی از هم جدا نمی‌کنند. برای تعمیم LDA می‌توان فرض کرد که کلاسها از زیر کلاسهایی تشکیل شده‌اند که دارای توزیع نرمال با میانگینهای متفاوت و ماتریس کواریانس مشترک هستند. ممیزی در این حالت آمیخته[۲] (MDA) می‌باشد وکرانهای تصمیم بین کلاسها غیر خطی و پیچیده بوده و برآورد پارامترها تنها از طریق الگوریتم EM امکان‌پذیر است.
کاهش ابعاد در ممیزی یکی از اهداف مهمی است که به وسیله ممیزی خطی فیشر امکان‌پذیر می‌گردد و بوسیله متغیرهای ممیزی فیشر می‌توان کلاسها را در نمودارهای دو بعدی مشاهده نمود. ممیزی آمیخته نیز این خاصیت را در بردارد وکاهش ابعاد داده‌ها در فضای برازش یافته توسط میانگین زیرکلاسها امکان‌پذیر است. در این مطالعه ممیزی بوسیله آمیخته‌های نرمال توسعه داده می‌شود و سپس داده‌های گزارش توسعه انسانی در سال ۲۰۰۲ را برای مقایسه روشهای ممیزی برای رده‌بندی کشورها به سه کلاس کشورها با توسعه انسانی بالا، کشورها با توسعه انسانی متوسط وکشورها توسعه انسانی پایین در نظر گرفته شدند و نشان داده شد که ممیزی آمیخته دارای نرخ خطای کمتری نسبت به دیگر روشها برای رده‌بندی کشورها است.

واژه‌های کلیدی: رده‌بندی، تحلیل خوشه‌ای، الگوریتم EM، تحلیل ممیزی انعطاف پذیر، شاخص توسعه انسانی

۱ مقدمه

رده‌بندی دو یا چند کلاس یا ممیزی از موضوعات مهم در آمار چند متغیره است که دارای کاربردهای مختلف در حوزه‌های گوناگون است. در حالت کلی ممیزی اختصاص یک یا چند مشاهده x با

1) Linear Discriminant Analysis 2) Mixture Discriminant Analysis

کلاس نامعلوم به جوامعی معلوم $G_1, G_2, ..., G_J$ است. برای این کار فرض می‌شود مجموعه راهنما $\{x_i, g_i\}_{i=1}^n$ است که در آن $x_i \in \Re^p$ موجود و g_i مشخص کننده کلاس نمونه iام است. با استفاده از مجموعه راهنما پارامترها در مدل برآورد و سپس با استفاده از پارامترهای برآورد شده هدف بدست آوردن یک قاعده برای ممیزی بصورت یک تابع جهت پیش‌بینی مشاهده جدید x_0 به یکی از کلاسها است.

روش سنتی وکلاسیک آماری ممیزی خطی و رگرسیون لجستیک است این روش‌ها تا حدود زیادی توسعه یافته و معمولاً در متون آماری موجود هستند. به عنوان مثال (ماردیا و دیگران ۱۹۷۹ و آندرسن ۱۹۸۴) برای ممیزی خطی و ممیزی لجستیک مراجع خوبی محسوب می‌شوند. هرگاه جوامع نرمال باشند تابع ممیزی خطی نیز نرمال و بنابراین خطای ممیزی نیز براحتی قابل محاسبه خواهد بود. روش درجه دوم[3] در ممیزی بدلیل پیچیده بودن توزیع ممیزی بوسیله آماردانان هنوز در حال توسعه است. وب (۱۹۹۹) روش‌های ممیزی غیر خطی از جمله درجه دوم، ناپارامتری و ... را مورد بررسی قرار داده است.

متاسفانه فرض نرمال بودن کلاسها بندرت پیش می‌آید. در چنین صورتی تابع ممیزی داده‌ها را بصورت موثری از هم جدا نمی‌کند و خطای ممیزی افزایش خواهد یافت. یک طریق برای حل این مساله فرض تعمیم LDA است به اینکه فرض شود کلاسها شامل زیرکلاسهائی نرمال‌اند یعنی کلاسها دارای توزیع آمیخته‌اند و بنابراین ممیزی بکار رفته نیز آمیخته خواهد بود. این موضوع در نوشتجات آماری بوسیله مک لاچلان (۱۹۹۲) و تکست و دیگران (۱۹۹۱) پیشنهاد شده است. هاستی و تیبشیرانی (۱۹۹۶) این مطلب را گسترش داده‌اند. آنها فرض کرده‌اند هرکلاس از زیر کلاسهایی با میانگین متفاوت و ماتریس کواریانس درون و بین کلاس‌های مشترک هستند. LDA فقط برای رده‌بندی استفاده نمی‌شود بلکه ابزار سودمندی برای کاهش ابعاد داده‌هاست با LDA می‌توان چندین کلاس از داده‌ها را در غالب نمودارهای دو بعدی نمایش داد بگونه‌ای که بیشترین جداسازی بین میانگین کلاسها را داشته باشند MDA نیز این خاصیت را دارد و می‌توان کاهش ابعاد را روی میانگین کلاسها (زیر کلاسها) مشاهده نمود. در LDA کاهش ابعاد در حالتی که ممیزی بین دو کلاس باشد امکان‌پذیر نیست اما در MDA این امر امکان‌پذیر است.

در این مقاله تحلیل ممیزی آمیخته بررسی و الگوریتم EM برای برآورد پارامترها و ممیزی پیشنهاد شده است. این الگوریتم بوسیله هاستی و تیبشیرانی پیشنهاد و بوسیله نویسندگان مقاله انطباق و سپس در داده‌های توسعه انسانی بکارگرفته شده است و در بخش دوم مقاله‌های ممیزی خطی مرور و بخش سوم مقاله توسعه LDA به MDA را شامل می‌شود. بخش چهارم به انطباق الگوریتم EM برای استفاده در ممیزی MDA می‌پردازد و در بخش پنجم داده‌های توسعه انسانی را برای مقایسه بین MDA با LDA و QDA در نظرگرفته شده‌اند سرانجام بخش‌های نهایی مقاله یک نتیجه‌گیری در خصوص MDA خواهد بود.

3) Quadratic Discriminant Analysis

۲ تحلیل ممیزی خطی LDA

فرض کنید در مساله ممیزی مجموعه راهنما $\{x_i, g_i\}_{i=1}^n$ در اینصورت چگالی شرطی X در کلاس jام به صورت زیر خواهد بود:

$$p_j(x) = \phi(\mu_j, \Sigma) = P(X = x \mid G = j)$$

$$= \mid 2\pi\Sigma \mid^{-\frac{1}{2}} exp\left\{-\frac{1}{2}D(x_i, \mu_j)\right\} \quad (1)$$

در اینصورت لگاریتم درستنمایی مشاهدات به صورت زیر خواهد بود:

$$2\ell(\mu_j, \Sigma) = -\sum_{j=1}^{J}\sum_{g_i=j} D(x_i, \mu_j) - nlog \mid \Sigma \mid \quad (2)$$

در اینجا $D(x, \mu) = (x-\mu)'\Sigma^{-1}(x-\mu)$ فاصله ماهالانوبیس بین x و μ و $\Sigma_{g_i=j}$ به معنی مجموع مشاهدات مجموعه راهنما در کلاس jام است. گیریم احتمال پیشین کلاس jام بصورت $P(G = j) = \Pi_j$ باشد که معمولاً از قبل معلوم است و یا از مجموعه راهنما برآورد می‌شود اگر هیچ دلیلی بر ترجیح کلاسها بر یکدیگر نباشد $\Pi_1 = \Pi_2 = ... = \Pi_J = \frac{1}{J}$ در نظر گرفته می‌شود سپس مشاهده x_0 به کلاس jام رده‌بندی می‌شود $C(x_0) = j$ اگر احتمال پسین اینکه از کلاس jام باشد در میان بقیه کلاسها ماکزیمم شود:

$$P(G = j \mid X = x_0) = max_\ell P(G = \ell \mid X = x_0) \quad (3)$$

هرگاه $P(G = j) = \frac{1}{J}$ باشد این روش با روش ممیزی ماکزیمم درستنمایی معادل می‌شود. با توجه به (۱) اگر کلاسها دارای توزیع نرمال با ماتریس کواریانس مشترک باشند چگالی پسین کلاس jام به صورت زیر خواهد بود:

$$\begin{aligned} P(G = j \mid X = x) &= \frac{\Pi_j \phi(\mu_j, \Sigma)}{\sum_{l=1}^{J} \Pi_l \phi(\mu_l, \Sigma)} \\ &\propto exp\left\{x'\Sigma^{-1}\mu_j - \frac{1}{2}\mu_j'\Sigma^{-1}\mu_j + \log \Pi_j\right\} \\ &\propto exp\left\{x'\beta_j + \alpha_j\right\} \end{aligned}$$

توجه کنید ثابت بودن مخرج کسر نسبت به j تاثیری در ماکزیمم کردن کلاسها ندارد. کران تصمیم بین دو کلاس مانند i و jام به صورت مجموعه‌ای از نقاط تعریف می‌شود که دارای احتمالات پسین مساوی هستند یعنی:

$$b_{i,j} = \{x \in \Re^p; P(G = j \mid X = x) = P(G = i \mid X = x)\}$$

و بنابراین کران تصمیم بین دو کلاس بصورت زیر است:

$$(\beta_j - \beta_i)'x + (\alpha_j - \alpha_i) = ۰$$

که تابعی خطی از x است . تابع ممیزی برای کلاس jام به صورت زیر تعریف می‌شود:

$$\delta_j(x) = -(x'\beta_j + \alpha_j) \qquad (۴)$$

و قاعده رده‌بندی انتساب $x_۰$ به کلاس jام است اگر رابطه زیر برقرار باشد:

$$\delta_j(x) = -min_\ell(x'\beta_\ell + \alpha_\ell)$$

باید توجه شود که در حالت مساوی بودن احتمال پیشین کلاسها مشاهده جدید $x_۰$ به کلاسی رده‌بندی می‌شود که کمترین فاصله ماهالانوبیس را با مرکز آن داشته باشد. α_ℓ و β_ℓ با استفاده از مجموعه راهنما برآورد می‌شوند بنابراین ماکزیمم (۲) بصورت زیر حاصل می‌شوند.

$$\hat{\mu}_j = \frac{۱}{n_j}\sum_{g_i=j} x_i \qquad \hat{\Sigma} = \frac{۱}{n}\sum_{g_i=j}^{J}\sum_{g_i=j}(x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)' \qquad (۵)$$

لازم بذکر است اگر کلاسهای دارای ماتریس کواریانس مختلف باشند یعنی:

$$p_j(x) = \phi(\mu_j, \Sigma_j)$$

کران تصمیم بین دو کلاس i و j غیر خطی به صورت تابع درجه دوم از x خواهد بود در این حالت تعداد پارامترهایی که باید برآورد شوند $JP + \frac{P(P+۱)}{۲}$ خواهد بود چون اندازه نمونه‌های مجموعه راهنما ثابت است بنابراین برآورد پارامترها در ممیزی درجه دوم نیرومند نیست و برای حالت $P > ۲$ پیشنهاد نمی‌شود.

۱.۲ کاهش ابعاد در LDA

کاهش ابعاد در ممیزی را می‌توان با تجزیه فیشر (۱۹۳۶) بدست آورد در روش پیشنهادی فیشر هدف بدست آوردن $\alpha \in \Re^p$ بگونه‌ای که کسر زیر ماکزیمم شود:

$$\frac{\alpha' B \alpha}{\alpha' W \alpha}$$

در اینجا B ماتریس کواریانس بین گروهها و W ماتریس کواریانس درون گروهی است. بهترین جواب برای α بردار ویژه مطابق با بزرگترین مقدار ویژه ماتریس $W^{-۱}B$ است در حالت کلی این ماتریس حداکثر دارای $min(J-۱, p)$ مقدار ویژه غیر صفر است بنابراین متغیرهای ممیزی

فیشر به صورت $y_k = \alpha_k' x$ و $k = ۱, ۲, ..., K$ می‌باشند که $K \leq min(J - ۱, p)$ یعنی عملاً از K ممیز اول فیشر در LDA به عنوان متغیر ممیزی استفاده می‌شود دو متغیر ممیزی اول فیشر بیشترین جداسازی را بین میانگین کلاسها دارند و برای نمایش کلاسها در غالب نمودار استفاده می‌شوند. در ممیزی خطی فیشر مشاهده جدید $x_۰$ به کلاس jام رده‌بندی می‌شود اگر فاصله اقلیدسی زیر کمترین در میان همه کلاسها باشد:

$$\delta_j^۲(x_۰) = \sum_{k=۱}^{K} (y_k(x_۰) - \bar{y}_k^j)^۲$$

که $\bar{y}_k^j$ میانگین متغیر ممیزی kام در کلاس jام است. در حالت نرمال بودن کلاسها هاستی و تیبشیرانی (۱۹۹۶) نشان داده‌اند که رده‌بندی در فضای ممیزی خطی فیشر معادل با LDA است. همچنین هاستی و دیگران (۱۹۹۴) نشان دادن که فضای ممیزی خطی فیشر (LDA) معادل با رگرسیون خطی چند متغیره است که از مقادیر بهینه برای نشان دادن کلاسها استفاده شده است در حالت کلی LDA را می‌توان بصورت دنباله‌ای از رگرسیونهای چندگانه خطی $\eta_k(X) = X'\beta_k$ و $k = ۱, ..., K$ نمایش داد این شکل از ممیزی پس از تعیین مقادیر بهینه $\theta_۱, ..., \theta_K$ و انتساب آنها به کد کلاسها بدست آورده می‌شود. روش کار بصورت زیر است فرض شود که Y ماتریسی است که اگر x_i متعلق به کلاس jام باشد آنگاه در ردیف iام و ستون jام مقدار آن یک است و بقیه ستونها در ردیف iام مقدار صفر را دارند:

۱) رگرسیون چند متغیره خطی: رگرسیون خطی چند متغیره از Y در مقابل X (ماتریس پیش بینی‌گرها) برازش می‌شود گیریم $\hat{Y}$ ماتریس مقادیر برازش شده باشد و $\eta(X)$ بردار توابع رگرسیونی باشد.

۲) مقادیر بهینه: K تا از بزرگترین بردارهای ویژه $Y'\hat{Y}$ با شرط $\Theta' D_\Pi \Theta = I_K$ که $D_\Pi = \frac{Y'Y}{n}$ ماتریس قطری از احتمال پیشین کلاسهاست فرض شود $\Theta J \times K$ ماتریس مقادیر بهینه برآورد شده باشد.

۳) با استفاده از ماتریس مقادیر بهینه در مرحله (۲) و بردار توابع رگرسیونی در مرحله (۱) $\eta(X) \longleftarrow \Theta'\eta(X)$.

در این حالت مشاهده جدید $x_۰$ به کلاس jام رده‌بندی می‌شود اگر فاصله اقلیدسی زیر در میان همه کلاسها کمترین باشد:

$$\delta_j^۲(x_۰) = \sum_{k=۱}^{K} w_k(\eta_k(x_۰) - \bar{\eta}_k^j(x)^۲$$

که $\bar{\eta}_k^j(x)$ میانگین مقادیر برازش شده تابع رگرسیونی $\eta_k(x)$ در کلاس jام است و $w_k = \frac{۱}{r_k^۲(۱-r_k^۲)}$ وزنهایی هستند که فضای برازش شده رگرسیونی را به فضای ممیزی فیشر تبدیل می‌کنند و r_k، kام مقدار ویژه‌ای است که در مرحله (۲) الگوریتم محاسبه می‌شود. اگر در مرحله (۱) از الگوریتم

از رگرسیون ناپارامتری استفاده شود در این صورت ممیزی غیر خطی خواهد بود. هاستی و دیگران (۱۹۹۴) هر روشی را که از یک پیش‌پردازش رگرسیونی برای مسائل ممیزی استفاده شود را ممیزی انعطاف‌پذیر[4] (FDA) نامیده‌اند بنابراین اگر پیش‌پردازش رگرسیون خطی باشد ممیزی نیز خطی خواهد بود.

۳ تعمیم LDA با استفاده از آمیخته‌های نرمال

فرض کنید در بخش قبلی تابع چگالی در هر کلاس آمیخته نرمال باشد که در آن هر مولفه دارای میانگین مربوط به خود و ماتریس کواریانس مشترک با بقیه مولفه‌ها باشد بنابراین در کلاس jام به صورت زیر خواهد بود:

$$p_j(x) = \sum_{r=1}^{R_j} \pi_{jr}\phi(x_i, \mu_{jr}, \Sigma)$$

$$= | \, 2\pi\Sigma \, |^{\frac{-1}{2}} \; exp\left\{-\frac{1}{2}D(\mu_{jr}, \Sigma)\right\} \qquad (۶)$$

در اینجا R_j تعداد مولفه‌های کلاس jام و π_{jr} نسبتهای آمیختگی هر یک از مولفه‌های نرمال است که $\sum_{r=1}^{R_j} \hat{\pi}_{jr} = 1$ می باشد. در مدل ممیزی آمیخته احتمال پسین کلاس jام در صورتی که x مشاهده شده باشد به صورت زیر است:

$$P(G = j \mid X = x) = \frac{\sum_{r=1}^{R_j} \pi_{jr}\phi(x_i, \mu_{jr}, \Sigma)}{\sum_{j=1}^{J}\sum_{r=1}^{R_j} \pi_{jr}\phi(x_i, \mu_{jr}, \Sigma)} \qquad (۷)$$

لگاریتم درستنمایی شرطی در اینجا بصورت زیر است:

$$\ell_m(\mu_{jr}, \Sigma, \pi jr) \propto \sum_{j=J}^{J} log(\sum_{r=1}^{R_j} \pi_{jr} exp\left\{-\frac{1}{2}D(\mu_{jr}, \Sigma)\right\}) - \frac{n}{2} log \mid \Sigma \mid \quad (۸)$$

در MDA مشاهده جدید x_0 به کلاس jام نسبت داده می‌شود، $C(x_0) = j$، اگر رابطه (۷) در میان تمام کلاسهای دیگر ماکزیمم شود:

$$P(G = j \mid X = x_0) = max_\ell P(G = \ell \mid X = x_0)$$

4) Flexible Discriminant Analysis

در اینجا پارامترهای مدل $R_j, \pi_j r, \mu_j r, \Sigma$ که $j = ۱, ..., J, \quad r = ۱, ..., R_j$ می‌باشد این پارامترها از طریق مجموعه راهنما با ماکزیمم کردن رابطه (۸) بدست می‌آیند بدیهی است که در MDA ماکزیمم کردن رابطه (۸) از طریق مشتق‌گیری امکان‌پذیر نیست بنابراین باید از طریق روشهای عددی ماکزیمم شود.

۴ برآورد پارامتر مدل ممیزی آمیخته

دمپستر و دیگران (۱۹۷۷) برای برآورد پارامترها مطرح کرده‌اند که در هنگام ناکامل بودن مشاهدات کاربرد دارد. الگوریتم EM روشی تکراری تا رسیدن به همگرایی در لگاریتم درستنمایی داده‌هاست که در هر تکرار مرحله E (امید ریاضی) و مرحله M (برآورد ماکزیمم درستنمایی) وجود دارد. در مدل ممیزی آمیخته بطور دقیق مشخص نیست که هر نمونه در مجموعه راهنما به کدام زیرکلاس تعلق دارد بنابراین برای ماکزیمم کردن (۲) باید از الگوریتم عددی EM استفاده نمود.

مرحله E:

مرحله E از الگوریتم EM شامل امید ریاضی لگاریتم درستنمایی داده‌های کامل است به شرط اینکه مقدار مشاهده شده x و مقدار صحیح پارامترها در هر تکرار در اختیار باشد در داده‌های آمیخته نرمال می‌توان مشاهدات را به صورت داده‌های ناکامل در نظرگرفت که مقادیرگمشده در آن کد زیرکلاس هر مولفه می‌باشد در اینجا با فرض اینکه پارامترهای ابتدایی مشخص هستند احتمال یا وزن اینکه مشاهده x_i متعلق به زیرکلاس rام ازکلاس jام باشد برای همه نمونه‌های مجموعه راهنما درکلاس jام محاسبه می‌شود که برای اثبات می‌توان به مک لاچلان (۱۹۹۲) مراجعه کرد:

$$P(C_{jr}(x_i) \mid x_i, j) = \frac{\pi_{jr}\phi(x_i, \mu_{jr}, \Sigma)}{\sum_{l=۱}^{R_j} \pi_{jl}\phi(x_i\mu_{jl}, \Sigma)} \qquad (۹)$$

مرحله M:

با استفاده از وزنهای (احتمالات) برآورد شده در مرحله E رابطه (۹) برآوردهای ماکزیمم درستنمایی برای پارامترهای هر مولفه نرمال در هرکلاس محاسبه می‌شوند:

$$\hat{\pi}_{jr} \propto \sum_{g_i=j} P(C_{jr}(x_i) \mid x_i, j) \qquad \sum_{r=۱}^{R_j} \hat{\pi}_{jr} = ۱ \qquad (۱۰)$$

$$\hat{\mu}_{jr} = \frac{\sum_{g_i=j} P(C_{jr}(x_i) \mid x_i, j)x_i}{\sum_{g_i=j} P(C_{jr}(x_i) \mid x_i, j)} \qquad (۱۱)$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{J}\sum_{g_i=j}\sum_{r=1}^{R_j} P(C_{jr}(x_i) \mid x_i, j)(x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)' \qquad (۱۲)$$

همانطور که از مقایسه رابطه (۵) با رابطه‌های (۱۱) و (۱۲) دیده می‌شود برآوردهای ماکزیمم درستنمایی مانند حالت نرمال کامل هستند با این تفاوت که وزنهای $P(C_{jr}(x_i) \mid x_i, j)$ جایگزین تابع نشانگر کلاسها شده‌اند. با توجه به اینکه وزنهای $P(C_{jr}(x_i) \mid x_i, j)$ خود تابعی از μ_{jr} و Σ هستند رابطه‌های (۱۰) تا (۱۲) در رابطه (۹) جایگزین شده تا وزنهای بعدی محاسبه شوند این مراحل آنقدر تکرار می‌شوند تا همگرایی در لگاریتم درستنمایی شرطی (۸) روی دهد یکی از خواص مناسب الگوریتم EM این است که در هر تکرار لگاریتم درستنمایی داده‌ها نسبت به تکرار قبلی افزایش می‌یابد بنابراین اگر در مرحله‌ای رابطه (۸) نسبت به تکرار قبل از آن تغییر کمی داشته باشد الگوریتم متوقف می‌شود و برآوردهای نهایی بدست آورده می‌شوند.

تکرارهای الگوریتم EM به اندازه خوشه‌ها R_j و برآوردهای ابتدایی برای پارامترهای μ_{jr} و Σ و وزنهای $P(C_{jr}(x_i) \mid x_i, j)$ دارند برای این منظور می‌توان از الگوریتم خوشه‌بندی میانگین Kام[5] یک مقدار ثابت برای خوشه‌ها (تعداد زیرکلاسها) تعیین کرد و سپس از این الگوریتم برای برآورد مراکز زیرکلاسها μ_{jr} در هر کلاس استفاده شود سپس برای همه مشاهدات مجموعه راهنما در کلاس jام $P(C_{jr}(x_i) \mid x_i, j)$ برابر با یک است اگر μ_{jr} نزدیکترین مرکز و در غیر اینصورت صفر است. پس از برآورد پارامترها مشاهده جدید x_0 مطابق با رابطه (۷) به خوشه مورد نظر نسبت داده می شود. در MDA با مساوی قرار دادن احتمال پسین کلاسها به تابعی پیچیده و غیر خطی خواهیم رسید بنابراین در MDA حتی با فرض مساوی بودن ماتریس کواریانس کلاسها (زیرکلاسها) کرانهای تصمیم بین کلاسها غیرخطی است که این یکی از تعمیمهای مهمی است که هنگام فرض نرمال آمیخته برای کلاسها در LDA بدست می‌آید.

فرض مساوی بودن ماتریس کواریانس زیرکلاسها این امکان را بوجود می‌آورد تا کاهش ابعاد داده‌ها در MDA بدست آورده شود به عبارت دیگر می‌توان فرض نمود که با نسخه وزنی LDA روبرو هستیم که زیرکلاسها جایگزین کلاسها شده‌اند اگر $R = \sum_j^J R_j$ تعداد کل زیرکلاسها باشد که هر کدام از مشاهدات با وزن خود به هر یک از زیرکلاسها تعلق می‌گیرد در این حالت کلاس jام شامل $n_j R_j$ مشاهده است بنابراین مجموعه راهنما در این حالت مجموعه‌ای وزنی و افزوده شده به تعداد $n' = \sum_j^J n_j R_j$ از مشاهدات است. دوباره از الگوریتم EM کمک گرفته می‌شود و پس از محاسبه وزنها در مرحله E در مرحله M مسئله LDA وزنی حل می‌شود نرم افزار مورد استفاده این امکان را می‌دهد که در مرحله M از مقادیر بهینه برای محاسبه متغیرهای ممیزی استفاده شود اگر $Z_{n\times R}$ ماتریس پاسخ ظاهری باشد که ردیفهای این ماتریس وزن مشاهدات است بگونه‌ای که اگر مشاهده iام متعلق به کلاس ام باشد jام باشد آنگاه زیرکلاسهای مربوط به کلاس jام با وزن این مشاهده پر می‌شوند و بقیه زیرکلاسها در ردیف iام مقدار صفر را می‌گیرند باید توجه داشت که مجموع وزنها برابر n است. $(n = \sum_{j=1}^{J}\sum_{g_i=j}\sum_{r=1}^{R_j} P(C_{jr}(x_i) \mid x_i, j))$

5) Kmeans

جدول ۱: نرخ خطای رده‌بندی نادرست برای داده‌های توسعه انسانی

روش	مجموعه راهنما	مجموعه آزمون
LDA	۰/۱۰۳	۰/۱۰۳
QDA	۰/۹۰۶	۰/۱۲۱
MDA	۰/۰۲۶	۰/۰۵۲

پس از محاسبه ماتریس ظاهری Z با استفاده از مراحل (۳-۱) گفته شده در بخش ۱.۲ مدل کاهش یافته MDA در هر تکرار محاسبه می‌شود. در LDA کاهش ابعاد داده‌ها توسط تعداد کلاسها محدود شده است بنابراین در حالتی که ممیزی بین دو کلاس باشد ($J = ۲$) کاهش ابعاد امکان‌پذیر نیست اما در MDA چون زیرکلاسها جایگزین کلاسها می‌شوند بنابراین همواره متغیرهای ممیزی قابل محاسبه بوده و کاهش ابعاد امکان‌پذیر است.

۵ مثال کاربردی: رده‌بندی کشورها

از سال ۱۹۹۰ سازمان ملل متحد هر ساله گزارشی تحت عنوان گزارش توسعه انسانی را منتشر می‌کند در این گزارش برای رده‌بندی کشورها از نظر توسعه انسانی آنها متغیرهای زیر در نظرگرفته می‌شود:

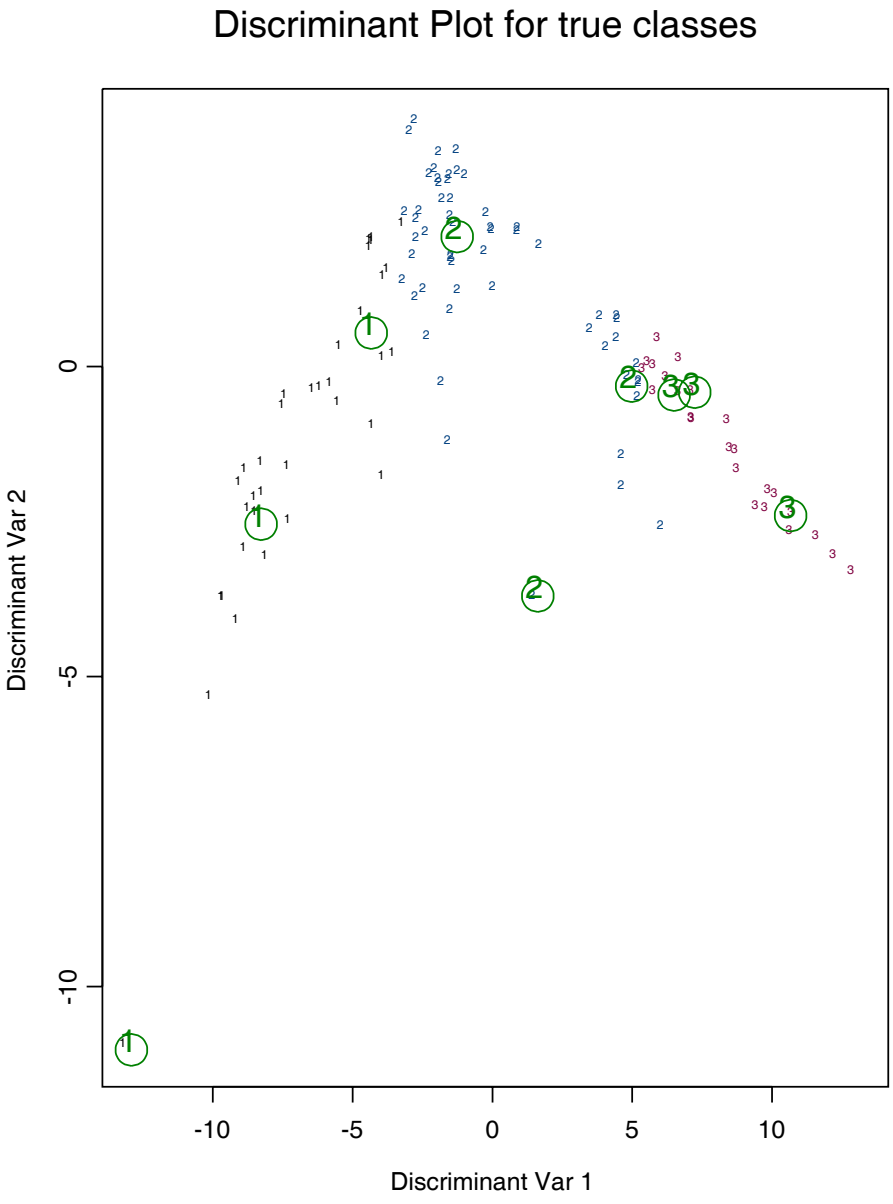
$x_۱$: امید به زندگی در بدو تولد.
$x_۲$: نرخ باسوادی بزرگسالان.
$x_۳$: نرخ رشد نام نویسی در دوره‌های تحصیلی.
$x_۴$: سرانه محصول ناخالص داخلی.

جمع‌آوری اطلاعات معمولا با استفاده از اطلاعات رسمی منتشره کشورها و در صورت غیر قابل اعتماد بودن اطلاعات بوسیله برنامه عمران سازمان ملل برآورد می‌شوند. در این گزارش از داده‌های بدست آمده شاخص توسعه انسانی[6] (HDI) محاسبه می‌شود و با استفاده از آن کشورها به سه رده کشورها با توسعه انسانی بالا، کشورها با توسعه انسانی متوسط وکشورها با توسعه انسانی پایین رده‌بندی می‌شوند.

۱.۵ رده‌بندی کشورها با تحلیل ممیزی

در گزارش توسعه انسانی سال ۲۰۰۲ آمارها برای ۱۷۳ کشور گزارش شده است برای مقایسه روشهای ممیزی خطی LDA، درجه دوم QDA و ممیزی آمیخته MDA، ۱۱۵ کشور بطور تصادفی برای مجموعه راهنما انتخاب شدند و نرخ خطای رده‌بندی نادرست برای مجموعه راهنما

6) Human Development Index



شکل ۱: متغیرهای ممیزی: متغیرهای ممیزی اول و دوم بهترکشورها را از هم جدا کرده‌اند

محاسبه شد سپس برای بررسی اعتبار مدلهای ارائه شده ۵۸ کشور دیگر به عنوان مجموعه آزمون در نظر گرفته شده‌اند و نرخ خطای رده‌بندی نادرست برای مشاهدات مجموعه آزمون با استفاده از توابع ممیزی برآورد شده توسط مجموعه راهنما محاسبه شده است همانطورکه از جدول (۱) مشاهده می‌شود MDA با سه زیرکلاس در هرکلاس دارای نرخ خطای کمتری نسبت به روشهای ممیزی خطی و درجه دوم است بنابراین از ممیزی آمیخته برای رده‌بندی کشورها استفاده شده است.

شکل (۱) نمودار مربوط متغیرهای ممیزی برای ممیزی خطی و ممیزی آمیخته را برای مجموعه راهنما نشان می‌دهد همانطورکه مشاهده می‌شود در MDA چون زیرکلاسها جایگزین کلاسها شده‌اند بنابراین تعداد متغیرهای ممیزی بیشتر از حالت خطی است اعداد درون دایره‌ها مشخص کننده مرکز زیرکلاسها می‌باشند و متغیرهای ممیزی اول و دوم بهتر از متغیرهای ممیزی دوم و سوم کلاسها (زیرکلاسها) را از هم جدا کرده‌اند.

کشورها با توسعه انسانی بالا:

آرژانتین، آنتیگوآ و باربودا، آلمان، اتریش، اروگوئه، اسپانیا، استرالیا، استونی، اسلواکی، اسلوونی، ایتالیا، امارات متحده عربی، انگلستان ، ایالات متحده آمریکا، ایرلند، ایسلند، باربادوس، باهاماس، بحرین، برونئی ، بلژیک، پرتغال، ترنیداد و توباگو، جمهوری چک، دانمارک، ژاپن، سنگاپور، سوئد، سوئیس، سیشل، فرانسه، رژیم اشغالگر قدس، فنلاند، قبرس، قطر، کاستاریکا، کانادا، کرواسی، کره‌جنوبی، کویت، لوکزامبورگ، لهستان، ماریتیوس، مالت، مالزی، مجارستان،مکزیک ، نروژ، نیوزلند، هلند، هنگ کنگ، یونان.

کشورها با توسعه انسانی متوسط:

آذربایجان، آفریقای جنوبی، آلبانی، اردن، ارمنستان، ازبکستان، اکراین، اکوادور، اکوتوریال، الجزایر، السالوادور، اندونزی، ایران، برزیل، بلغارستان، بلیز، بولیوی، بوتسوانا، پاپوآ، پاراگوئه، پاناما، پرو، تاجیکستان، تایلند، ترکمنستان، ترکیه، تونس، جامائیکا، جزایر سلیمان، دمنیکا، جمهوری دمنیکن، چین، روسیه، روسیه سفید، رومانی، زیمباوه، ساموآ، ساوتومو، سری لانکا، سنت اوسیا، سنت وینت، سوازیلند، سورینام، سوریه، عربستان، عمان، غنا، فیلیپین، فیجی، قرقیزستان، قزاقستان، کاپورد، کامبوج، کلمبیا، کنگو، کنیا، کوبا، گرجستان، گرانادا، گواتمالا، گوانا جدید، گویان، مالدیو، مراکش، مصر، مغولستان، مقدونیه، ملداوی، میانمار، لاتویا، لیتوانی، لبنان، لسوتو، لیبی، لیتونی، نامبیا، نیکاراگوئه، ونزوئلا، ویتنام، هند، هندوراس.

کشورها با توسعه انسانی پایین:

آفریقای مرکزی، آنگولا، اتیوپی، اریتره، اوگاندا، بنگلادش، بنین، بوتان، بورکینافاسو، بروندی، پاکستان، تانزانیا، توگو، چاد، جمهوری مردمی لئو، جیبوتی، روآندا، زامبیا، سنگال، سودان، سیرالئون، کامرون، کنگو، کوته‌دویری، کوموروس، گابن، گامبیا، گینه، گینه‌بیسائو، ماداگاسکار، مالاوی، مالی، موریتانی، موزامبیک، نپال، نیجر، نیجریه، وانوتوآ، هائیتی، یمن.

۶ نتیجه‌گیری

همانطورکه گفته شد بدلیل غیر نرمال بودن کلاسها ممیزی متداول کلاسیک توان کافی در جداسازی کلاسها را ندارند. یک روش پارامتری را در ممیزی میتوان با فرض داشتن توزیع نرمال آمیخته در کلاسها بدست آورد با این فرض دو تعمیم مهم در ممیزی خطی بدست می‌آید:
۱) کرانهای تصمیم بین کلاسها غیر خطی و پیچیده است.
۲) هنگامیکه ممیزی بین دو کلاس باشد کاهش ابعاد بر خلاف LDA امکان‌پذیر است.
روش ممیزی آمیخته اگر چه منجر به محاسبات پیچیده و استفاده از روشهای عددی بجای تحلیلی می‌گردد ولی با توجه به دسترسی به رایانه‌ها و نرم‌افزارهای دقیق می‌توان از روش MDA استفاده کرد.

مراجع

[1] Anderson, T. W.(1984), An Introduction to Multivariate Statistical Analysis, John Wiely, New York.

[2] Dempster, A. P.Laird, N. M. and Rubin, D. B.(1977), Maximum likelihood from incomplete data via EM algorithm(with discussion), R. Statis. Sac. B, 39:1-38.

[3] Hastie, T. Tibshirani, R. and Buja, A.(1994), Flexible Discriminant Analysis by Optimal Scoring, J. Amer. Statis. Assoc. 89, 1225-1270.

[4] Hastie, T. Tibshirani, R. (1996), Discriminant Analysis by Gaussian Mixtures, J. R. Statis. Soc. B, 58:155-176.

[5] Mardia, K. Kent, J. and Bibby, J.(1979), Multivariate Analysis, Academic Press, London.

[6] McLachlan, G. J. (1992), Discriminant Analysis and Statistica Pattern Recognition, John Wiley, New York.

[7] Taxt, T. Hjot, N. and Eikvil, L. (1991), Statistical classification using a linear mixture of multinormal probability densities, Pattern Recognition Letters, 12:731-737.

[8] UNDP. (2002), Human Development Report 2002, New York.

[9] Webb, A. (1999), Statistical Pattern Recognition, Arnold, London.

خطای مطلق یا نسبی در تعیین اندازه نمونه

علیرضا زاهدیان

مرکز آمار ایران

چکیده: یکی از مهمترین مسایلی که در شروع طراحی یک طرح نمونه‌گیری با آن روبرو هستیم این است که اندازه نمونه مورد نیاز باید چقدر باشد تا با اطمینان کافی انتظار داشته باشیم خطای نمونه‌گیری در برآورد پارامترهای مورد نظر در حد قابل قبولی باشد. در تعیین اندازه نمونه، عواملی مانند پراکندگی واحدهای جامعه آماری از نظر صفت مورد بررسی، اندازه جامعه، توزیع صفت مورد بررسی و حداکثر خطای نمونه‌گیری قابل قبول، نقش تعیین کننده‌ای دارند. در بین عوامل فوق، خطای نمونه‌گیری تنها عاملی است که کنترل آن در اختیار طراح آمارگیری قرار دارد و با تغییر آن می‌تواند حجم کار و هزینه طرح را تا جایی که به اهداف آن صدمه وارد نشود، کاهش دهد. در این مقاله روشهای مختلف تعیین حداکثر خطای نمونه‌گیری قابل پذیرش برای تعیین اندازه نمونه در طرحهای آمارگیری و اثرات آنها در فرمولهای تعیین اندازه نمونه مورد بحث قرار می‌گیرد.

واژه‌های کلیدی: اندازه نمونه، خطای نسبی نمونه‌گیری، خطای مطلق نمونه‌گیری

۱ مقدمه

یکی از عواملی که در دقت نتایج یک آمارگیری نمونه‌ای تاثیر مستقیم دارد، اندازه نمونه است. بگونه‌ای که با افزایش اندازه نمونه خطای نمونه‌گیری کاهش پیدا کرده و به تعبیری نتایج بدست آمده از اعتبار آماری بیشتری برخوردار خواهند بود. ولی در عمل با افزایش اندازه نمونه، هزینه آمارگیری و میزان خطای غیر نمونه‌گیری خصوصا در آمارگیریهای بزرگ افزایش پیدا می‌کنند. علاوه بر این در برخی از زمینه‌ها مانند تحقیقات پزشکی که واحدهای آماری انسانها هستند، برای افزایش اندازه نمونه محدودیتهای بسیاری وجود دارد. بنابراین آمارشناس باید به کمک محقق طرح اندازه نمونه را بگونه‌ای تعیین کند که بین دقت آمارگیری و هزینه آن تعادل لازم برقرار شود. مسئله تعیین اندازه نمونه از جنبه‌های مختلف در کتب و مقالات آماری مورد بحث و بررسی قرار گرفته است. برای مثال می‌توان به مراجعی مانند دمینگ[1] (۱۹۶۶)، کوکران[2] (۱۹۷۷)، کرامر و تیمن[3] (۱۹۸۷)، کوهن[4] (۱۹۸۸و۱۹۹۷)، شیفر[5] و همکاران (۱۹۹۰)، کیش[6] (۱۹۹۵)، لنت[7] (۲۰۰۱) و تامپسون[8] (۲۰۰۲) اشاره کرد. در این مراجع عواملی مانند تعیین یا برآوردهای تخمین

1) Deming 2) Cochran 3) Kramer and Thiemann 4) Cohen 5) Schaeffer 6) Kish 7) Lenth 8) Thompson

میزان پراکندگی صفت مورد بررسی، توزیع آماری صفت مورد بررسی، توان آزمونهای آماری مربوط به میانگین، مجموع و نسبت، در ارتباط با اندازه نمونه، مورد تجزیه و تحلیل قرار گرفته‌اند و در مورد مسئله تعیین حداکثر خطای نمونه‌گیری قابل پذیرش و رابطه آن با اندازه نمونه نیز اشاراتی شده است ولی این مسئله به طور خاص مورد بررسی قرار نگرفته است. این در حالی است که حداکثر خطای نمونه‌گیری قابل پذیرش (یا به اختصار «خطا») در بین عوامل فوق تنها عاملی است که تعیین آن در اختیار طراح طرح بوده و با تغییر آن در حدی که به کیفیت و اهداف طرح صدمه‌ای وارد نشود، می‌توان از اندازه نمونه و به تبع آن از هزینه طرح کاست. در این مقاله سعی شده است فرمولهای تعیین اندازه نمونه مورد نیاز برای پارامترهایی مانند میانگین، مجموع و نسبت، با تمرکز بر نقش خطای مذکور در تعیین اندازه نمونه مورد بررسی قرار گیرد.

تعبیر ریاضی حداکثر خطای نمونه‌گیری قابل پذیرش را می‌توان در قالب رابطه زیر بیان کرد:

$$Pr(|\hat{\Theta} - \Theta| \leq d) = ۱ - \alpha \qquad (۱)$$

که بر اساس آن با احتمال $۱ - \alpha$ می‌توانیم انتظار داشته باشیم که خطای نمونه‌گیری در برآورد پارامتر مورد نظر (Θ) حداکثر برابر با d یا همان حداکثر خطای نمونه‌گیری قابل پذیرش باشد. در رابطه بالا خطای نمونه‌گیری به صورت مطلق و برابر با تفاوت بین مقدار واقعی پارامتر و برآورد آن بیان شده است. حال اگر بخواهیم خطای نمونه‌گیری را به صورت نسبی در نظر بگیریم، رابطه (۱) به صورت زیر تبدیل می‌شود:

$$Pr\left(\left|\frac{\hat{\Theta} - \Theta}{\Theta}\right| \leq r\right) = ۱ - \alpha$$

و یا به عبارت دیگر

$$Pr\left(\left|\hat{\Theta} - \Theta\right| \leq r\Theta\right) = ۱ - \alpha \qquad (۲)$$

که در آن r حداکثر خطای نسبی قابل پذیرش است. با مقایسه این رابطه با رابطه (۱) به این نتیجه می‌رسیم که اندازه نمونه با حداکثر خطای مطلق d معادل با اندازه نمونه با حداکثر خطای مطلق $r\Theta$ است. بنابراین در تعیین اندازه نمونه، خطای نمونه‌گیری را به دو صورت می‌توان بیان کرد:

۱ــ خطای مطلق ۲ــ خطای نسبی.

استفاده از خطای نسبی در بین آمارشناسان از عمومیت بیشتری برخوردار است ولی باید توجه داشت حداکثر خطای قابل پذیرش، توسط محقق طرح تعیین می‌شود و باید به گونه‌ای بیان شود که برای وی به راحتی قابل درک باشد. به عبارت دیگر در تعیین حداکثر خطای قابل پذیرش، شیوه بیان خطا نیز از اهمیت خاصی برخوردار است.

همانطورکه اشاره شد، حداکثر خطای قابل پذیرش باید توسط محقق طرح تعیین شود ولی معمولاً

وی نمی‌تواند تعابیر آماری نظرات خود را بطور دقیق بیان نماید. از اینرو باید آمارشناس با توجه به اهداف طرح سوالاتی را از محقق طرح بپرسد تا بتواند حداکثر خطای قابل پذیرش را با تعابیر آماری بیان کند. برخی از این سوالات را می‌توان به صورت زیر بیان نمود:

- چقدر خطا در برآورد پارامتر مورد بررسی قابل چشم‌پوشی است؟
- چه میزان تغییر در پارامتر مورد بررسی را می‌توان نادیده گرفت؟
- پارامتر مورد بررسی را در چه فاصله‌ای تخمین می‌زنید؟
- آیا از نظر شما تفاوت بین $\Theta \pm d$ و Θ معنی‌دار است؟
- آیا از نظر شما $\Theta \pm r\Theta$ با Θ تفاوت معنی‌داری دارد؟ (منظور مقادیر عددی r و Θ است که توسط آمارشناس پیشنهاد می‌شود)
- مقدار پارامتر مورد بررسی را تا چند رقم می‌توان به بالا یا پایین گرد کرد؟

برای مثالهای بیشتری در این زمینه می‌توانید مقاله لنت (۲۰۰۱) را ببینید. علاوه بر این در تعیین حداکثر خطای قابل پذیرش باید به نکات زیر توجه داشت:

- حداکثر خطای قابل پذیرش نباید از خطای گرد کردن قابل پذیرش بیشتر باشد.
- در بعضی موارد بیان خطا به صورت نسبی منطقی نیست. برای مثال، در برآورد متوسط دمای بدن بیماران مورد آزمایش نمی‌توان خطا را به صورت درصدی از دمای بدن بیان کرد یا در برآورد متوسط قطر قطعات تراش داده شده توسط یک دستگاه نمی‌توان خطا را به صورت درصدی از متوسط قطر قطعات تعیین نمود.
- در مواردی که تعاریف یا استانداردهای خاصی برای اندازه‌گیری صفت مورد بررسی وجود دارد باید حداکثر خطا را بر اساس آنها بیان کرد. برای مثال، حداکثر خطا در مورد برآورد دمای بدن انسان باید ۰٫۵ درجه باشد.
- استفاده از خطای نسبی مستلزم تخمینی از پارامتر مورد بررسی است و در مواردی که این پارامتر دارای نوسانات زیادی باشد استفاده از خطای نسبی توصیه نمی‌شود. بنابراین معمولا از خطای نسبی در مواردی که آمارگیری دارای سابقه تکرار (در سال گذشته) است، استفاده می‌شود.
- در مواردی که پارامتر مورد بررسی طی زمان دارای روند صعودی است (برای مثال، مجموع درآمد سالیانه خانوارهای کشور هر سال افزایش پیدا می‌کند)، استفاده از خطای نسبی محتاطانه است و در مواردی که پارامتر مورد بررسی روند نزولی دارد (برای مثال، نسبت جمعیت روستایی به کل جمعیت کشور) غیر محتاطانه است.
- در مواردی که جامعه آماری به تعدادی زیرجامعه (مانند استان، مرد و زن، شهری و روستایی و . . .) شکسته شده و اندازه نمونه برای هر زیر جامعه به صورت جداگانه تعیین می‌شود، استفاده از خطای نسبی موجب می‌شود علاوه بر عوامل دیگر، عامل بزرگی زیر جامعه از

نظر صفت مورد بررسی نیز در توزیع اندازه نمونه بین زیر جامعه‌ها دخالت داده شود. به عبارت دیگر در این حالت از زیرجامعه بزرگتر نمونه بیشتری انتخاب می‌شود.

حال رابطه (۱) را در نظر بگیرید. اگر $\hat{\Theta}$ برآورد نا اریبی از Θ بوده و دارای توزیع نرمال باشد، $\frac{\hat{\Theta}-\Theta}{\sqrt{V(\hat{\Theta})}}$ دارای توزیع نرمال استاندارد است. بنابراین داریم:

$$Pr\left(\frac{\left|\hat{\Theta}-\Theta\right|}{\sqrt{V(\hat{\Theta})}} \leq Z_{(۱-\frac{\alpha}{۲})}\right) = Pr\left(\left|\hat{\Theta}-\Theta\right| \leq Z_{(۱-\frac{\alpha}{۲})}\sqrt{V(\hat{\Theta})}\right) = ۱-\alpha$$

که در آن $Z_{(۱-\frac{\alpha}{۲})}$ صدک $۱۰۰(۱-\frac{\alpha}{۲})$ ام توزیع نرمال استاندارد است. با توجه به اینکه واریانس $\hat{\Theta}$ با افزایش n کاهش پیدا می‌کند، اندازه نمونه (n) را بگونه‌ای باید انتخاب کرد که رابطه $Z_{(۱-\frac{\alpha}{۲})}\sqrt{V(\hat{\Theta})} \leq d$ برقرار باشد. در قسمتهای بعد چگونگی تعیین n با توجه به رابطه فوق، برای پارامترهایی مانند میانگین، مجموع و نسبت در روشهای نمونه‌گیری که از عمومیت بیشتری برخوردار هستند، مورد بحث قرار خواهد گرفت.

۲ برآورد میانگین با نمونه‌گیری تصادفی ساده

برآورد هر پارامتری معمولا با دو هدف انجام می‌گیرد:
۱ــ ساخت فاصله اطمینان ۲ــ آزمون فرض. در قسمتهای بعد روش تعیین اندازه نمونه برای هر یک از این دو هدف، مورد بررسی قرار می‌گیرد.

۱.۲ برآورد میانگین با هدف ساخت فاصله اطمینان (برآورد فاصله‌ای)

در این حالت n را طوری انتخاب می‌کنیم که رابطه زیر برقرار باشد:

$$Pr\left(|\bar{x}-\mu| \leq Z_{(۱-\frac{\alpha}{۲})}\sqrt{V(\bar{x})}\right) = ۱-\alpha$$

با توجه به اینکه $V(\bar{x}) = \frac{N-n}{N}\frac{\sigma^۲}{n}$ می‌توانیم بنویسیم:

$$d = Z_{(۱-\frac{\alpha}{۲})}\sqrt{\frac{N-n}{N}\frac{\sigma^۲}{n}}$$

و بر این اساس اندازه نمونه مورد نیاز از رابطه زیر بدست می‌آید:

$$n = \frac{n_۰}{۱+\frac{n_۰}{N}}$$

که

$$n_{\circ} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})}\sigma^{۲}}{d^{۲}} \qquad (۳)$$

در روابط بالا اگر از ضریب تصحیح جامعه (fpc) که برابر $\frac{N}{N-n}$ است، صرف نظر کنیم، $n_{\circ}$ همان اندازه نمونه مورد نیاز خواهد بود. $Z_{(۱-\frac{\alpha}{۲})}$ تعیین کننده اطمینان مورد نظر است و معمولا به ازای $\alpha = ۰/۰۵$ برابر با ۱/۹۶ یا تقریباً ۲ در نظر گرفته می‌شود ولی اگر سطح اطمینان بالایی مدنظر باشد، (مثلاً در آزمایشات پزشکی، پذیرش محموله‌های دارو یا کنترل تولید قطعات خاص در صنعت) α برابر ۰/۰۱ و در نتیجه $Z_{(۱-\frac{\alpha}{۲})} = ۲/۵۸$ یا تقریباً ۳ در نظر گرفته می‌شود. روشهای تعیین $\sigma^{۲}$ عبارتند از:

۱ــ آمارگیریهای قبلی ۲ــ آمارگیری اولیه ۳ــ تقسیم آمارگیری به دو مرحله، یکی برای تخمین $\sigma^{۲}$ و دیگری برای برآورد میانگین که اطلاعات مرحله اول را نیز شامل می‌شود. در حالتی که هیچ اطلاعی در مورد $\sigma^{۲}$ در دست نباشد، با فرض نرمال بودن توزیع صفت مورد بررسی، می‌توان $\sigma^{۲}$ را بر اساس دامنه تغییرات تخمین زد:

$$\sigma^{۲} = \frac{range}{۶}$$

که تخمین محافظه‌کارتر $\sigma^{۲} = \frac{range}{۴}$ است. زیرا ۹۵ درصد از مقادیر یک جامعه نرمال در فاصله ۲ انحراف معیار و ۹۹/۷ در صد از مقادیر آن در فاصله ۳ انحراف معیار از میانگین قرار دارند. شیفر و همکارانش (۱۹۹۰) و لر[9] (۱۹۹۹) مثالهایی را در مورد استفاده از این روش ارائه کرده‌اند.

حال به سراغ d می‌رویم همانطورکه قبلتر اشاره شد، می‌توان d را به صورت مطلق یا نسبی تعیین کرد. اگر d را به صورت نسبی تعیین کنیم با صرف نظر از ضریب جامعه محدود (fpc) داریم:

$$n = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})}\sigma^{۲}}{r^{۲}\mu^{۲}} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})}}{r^{۲}} \quad CV$$

که در این حالت به جای واریانس $(\sigma^{۲})$ از ضریب تغییرات (CV) برای تعیین اندازه نمونه استفاده می‌شود. با توجه به اینکه ضریب تغییرات یک متغیر در طول زمان نسبت به واریانس آن از پایایی بیشتری برخوردار است، استفاده از خطای نسبی در آمارگیریهایی که در دوره‌های زمانی تکرار می‌شوند، مناسب‌تر است. شیفر و همکارانش (۱۹۹۰) از رابطه زیر برای تعیین اندازه نمونه استفاده کرده‌اند:

$$n = \frac{CV}{CV_{\circ}(\bar{x})}$$

که در آن $CV_{\circ}(\bar{x})$ ضریب تغییرات مطلوب در برآورد میانگین است.

9) Lohr

۲.۲ برآورد میانگین با هدف آزمون فرض

فرض کنید برآورد میانگین با هدف انجام آزمون فرض $H_۰ : \mu = \mu_۰$ در مقابل $H_۱ : \mu \neq \mu_۰$ انجام شود. در این حالت ناحیه رد فرض $H_۰$ عبارتست از:

$$R = \{\bar{x} > \mu_۰ + Z_{(۱-\frac{\alpha}{۲})}\sigma_{\bar{x}}\} \cup \{\bar{x} < \mu_۰ - Z_{(۱-\frac{\alpha}{۲})}\sigma_{\bar{x}}\}$$

و ناحیه قبول عبارتست از:

$$A = \{|\bar{x} - \mu_۰| < Z_{(۱-\frac{\alpha}{۲})}\sigma_{\bar{x}}\}$$

در این صورت توان آزمون (احتمال رد یک فرض صفر نادرست) از رابطه زیر بدست می‌آید:

$$\begin{aligned} Power &= ۱ - \beta = Pr(\bar{x} > \mu_۰ + Z_{(۱-\frac{\alpha}{۲})}\sigma_{\bar{x}}) + Pr(\bar{x} < \mu_۰ - Z_{(۱-\frac{\alpha}{۲})}\sigma_{\bar{x}}) \\ &= Pr(\frac{\bar{x}-\mu}{\sigma_{\bar{x}}} > \frac{\mu_۰ - \mu}{\sigma_{\bar{x}}} + Z_{(۱-\frac{\alpha}{۲})}) + Pr(\frac{\bar{x}-\mu}{\sigma_{\bar{x}}} < \frac{\mu_۰ - \mu}{\sigma_{\bar{x}}} - Z_{(۱-\frac{\alpha}{۲})}) \\ &= \Phi(-\sqrt{n}\frac{\mu-\mu_۰}{\sigma} - Z_{(۱-\frac{\alpha}{۲})}) + \Phi(\sqrt{n}\frac{\mu-\mu_۰}{\sigma} - Z_{(۱-\frac{\alpha}{۲})}) \end{aligned}$$

که در آن، α مقدار خطای نوع اول قابل پذیرش (معمولاً برابر با ۰٫۰۵) و $\Phi(.)$ تابع احتمال توزیع نرمال استاندارد است. اگر $\sqrt{n}\frac{\mu-\mu_۰}{\sigma} \geq ۱$ باشد، می‌توانیم $\Phi(-\sqrt{n}\frac{\mu-\mu_۰}{\sigma} - Z_{(۱-\frac{\alpha}{۲})})$ را تقریباً مساوی صفر در نظر بگیریم. در این صورت داریم:

$$Power = \Phi(\sqrt{n}\frac{\mu-\mu_۰}{\sigma} - Z_{(۱-\frac{\alpha}{۲})})$$

و اندازه نمونه مورد نیاز از رابطه زیر بدست می‌آید:

$$n = (Z_{(۱-\frac{\alpha}{۲})} + Z_{(۱-\beta)})^۲ \frac{\sigma^۲}{(\mu-\mu_۰)^۲}$$

معمولاً α=۰٫۰۵ و β=۰٫۲ در نظرگرفته می‌شود. فرمول توان آزمون نشان می‌دهد با افزایش n توان آزمون در سطح معنی‌داری α افزایش پیدا می‌کند. معمولاً فرضهای $H_۰$ و $H_۱$ به گونه‌ای تعیین می‌شوند که اهمیت α بیشتر از β باشد. به عبارت دیگر در بسیاری از موارد محقق فقط α را تحت کنترل قرار می‌دهد و نیاز چندانی به کنترل β احساس نمی‌کند و در نتیجه فرضهای $H_۰$ و $H_۱$ را نیز بر همین اساس تعریف می‌کند. در این موارد خصوصاً در نمونه‌گیری از جوامع بزرگ می‌توان از $Z_{۱-\beta}$ در فرمول اندازه نمونه چشم‌پوشی کرد که در این صورت داریم:

$$n_۰ = \frac{Z^۲_{(۱-\frac{\alpha}{۲})}\sigma^۲}{(\mu-\mu_۰)^۲} \qquad (۴)$$

از مقایسه روابط (۳) و (۴) می‌توانیم نتیجه بگیریم در صورتی که کنترل β مد نظر نباشد، در فرمول تعیین اندازه نمونه، رابطه $d \leq \mu - \mu_{\circ}$ باید برقرار باشد. در این صورت اگر $\mu_{\circ}$ پس از انجام آمارگیری خارج از فاصله اطمینان قرار گیرد، فرض $H_{\circ}$ در سطح معنی‌داری α رد می‌شود. برای مثال فرض کنید بخواهیم آزمون کنیم افزایش متوسط درآمد قشر خاصی از جامعه نسبت به سال گذشته بیشتر از نرخ تورم (برای مثال ۲۰ درصد) بوده است یا خیر. در این حالت باید $d \leq ۰/۲\mu'$ باشد که μ' متوسط درآمد در سال گذشته است.

۳ برآورد مجموع با نمونه‌گیری تصادفی ساده

در این حالت $\hat{X} = N\bar{x}$ که در آن N تعداد واحدهای جامعه است، برآوردی ناریب از مجموع صفت مورد بررسی می‌باشد. بنابراین مشابه قسمت قبل، اندازه نمونه مورد نیاز برای برآورد مجموع بر اساس خطای مطلق از رابطه زیر بدست می‌آید:

$$n_{\circ} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} N^{۲}\sigma^{۲}}{d^{۲}} \qquad (۵)$$

حال اگر $d = r \times X$ در نظر گرفته شود، اندازه نمونه بر اساس خطای نسبی عبارتست از:

$$n_{\circ} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} N^{۲}\sigma^{۲}}{r^{۲}X^{۲}} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} N^{۲}\sigma^{۲}}{r^{۲}N^{۲}\mu^{۲}} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} \sigma^{۲}}{r^{۲}\mu^{۲}}$$

بنابراین فرمول تعیین اندازه نمونه برای برآورد مجموع براساس خطای نسبی با فرمول تعیین اندازه نمونه برای برآورد میانگین یکسان است.

حال فرض کنید آزمون فرض $H_{\circ} : X = X_{\circ}$ در مقابل $H_{۱} : X \neq X_{\circ}$ مد نظر باشد. در این حالت، اندازه نمونه مورد نیاز از رابطه زیر بدست می‌آید:

$$n = (Z_{(۱-\frac{\alpha}{۲})} + Z_{(۱-\beta)})^{۲} \frac{N^{۲}\sigma^{۲}}{(X - X_{\circ})^{۲}}$$

۴ برآورد نسبت با نمونه‌گیری تصادفی ساده

می‌دانیم $\hat{P}$ که برابر با تعداد واحدهای دارای صفت مورد نظر در نمونه تقسیم بر تعداد واحدهای نمونه است، برآوردگری ناریب برای نسبت در جامعه (P) است و واریانس آن از رابطه

$\frac{N-n}{N}\frac{P(۱-P)}{n}$ $V(\hat{P}) =$ بدست می‌آید. بنابراین مشابه میانگین، اندازه نمونه مورد نیاز برای برآورد P بر اساس خطای مطلق را می‌توان از رابطه زیر بدست آورد:

$$n_۰ = \frac{Z^۲_{(۱-\frac{\alpha}{۲})}P(۱-P)}{d^۲} \qquad (۶)$$

در رابطه بالا P با یکی از روشهای زیر تعیین می‌شود:

- استفاده از اطلاعات آمارگیریهای قبلی
- استفاده از آمارگیری اولیه
- تقسیم آمارگیری به دو مرحله، یکی برای برآورد اولیه P و دیگری برای برآورد نهایی P
- مساوی گرفتن P با $\frac{۱}{۲}$ که موجب ماکزیمم شدن n به ازای Z و d ثابت می‌شود.

اگر به جای d از خطای نسبی استفاده کنیم فرمول تعیین اندازه نمونه به صورت زیر تبدیل می‌شود:

$$n_۰ = \frac{Z^۲_{(۱-\frac{\alpha}{۲})}P(۱-P)}{r^۲P^۲} = \frac{Z^۲_{(۱-\frac{\alpha}{۲})}(۱-P)}{r^۲P} \qquad (۷)$$

موضوع مهمی که باید به آن توجه داشت این است که اندازه نمونه در رابطه (۶) به ازای $P = \frac{۱}{۲}$ ماکزیمم می‌شود ولی اندازه نمونه در رابطه (۷) با کاهش P بزرگتر می‌شود. برای مثال نرخ بیکاری را به عنوان یک نسبت در نظر بگیرید. اگر اطلاعات قبلی حاکی از این باشد که نرخ بیکاری در یک کشور بین ۱۰ تا ۱۵ درصد در نوسان است، بر اساس رابطه (۶)، n به ازای $P =$ ۰٫۱ و بر اساس رابطه (۷) به ازای $P =$ ۰٫۱۵ ماکزیمم خواهد بود. بنابراین در تعیین اندازه نمونه برای برآورد P استفاده از خطای نسبی در مقایسه با خطای مطلق، موجب بروز تغییرات مهمی می‌شود و به این نتیجه می‌رسیم که استفاده از خطای نسبی در مواردی قابل توصیه است که واحدهای دارای صفت مورد بررسی، نادر باشند. کوکران (۱۹۷۷) واحدهایی را که نسبت آنها کمتر از ۱۰ درصد باشد، نادر معرفی کرده است. کوکران (۱۹۷۷) و زاکولا[10] (۱۹۹۹) استفاده از خطای نسبی را برای تعیین اندازه نمونه مورد نیاز برای برآورد تعداد ($N \times P$) مناسب دانسته‌اند ولی دلیل خاصی برای آن ذکر نکرده‌اند. شاید دلیل آن این باشد که اندازه نمونه مورد نیاز برای برآورد $N \times P$ بر اساس خطای مطلق و خطای نسبی به ترتیب از دو رابطه $n_۰ = \frac{Z^۲_{(۱-\frac{\alpha}{۲})}N^۲P(۱-P)}{d^۲}$ و $n_۰ = \frac{Z^۲_{(۱-\frac{\alpha}{۲})}(۱-P)}{r^۲P}$ بدست می‌آیند و در فرمول دوم نیازی به N نیست و اندازه نمونه برای برآورد P و $N \times P$ یکسان است.

در تعیین حداکثر خطای قابل پذیرش در برآورد P باید توجه داشت که معمولاً نسبتها به صورت درصد و تا دو رقم اعشار گزارش می‌شوند (در اینگونه موارد حداکثر خطای مطلق نباید بیشتر

10) Zakkula

از ۰/۰۰۵ یا همان خطای گرد کردن باشد). بنابراین در تعیین اندازه نمونه برای برآورد نسبت راحتر می‌توان حداکثر خطای مطلق قابل پذیرش را تعیین نمود و استفاده از خطای نسبی فقط در مواردی که واحدهای دارای صفت مورد نظر، نادر هستند، قابل توجیه است. بدیهی است در اینگونه موارد نیز برای آنکه حداکثر خطای مطلق کمتر از خطای گرد کردن (معمولاً ۰/۰۰۵) باشد، باید حداکثر خطای نسبی را ۰/۰۵ در نظر بگیریم.

۵ نمونه‌گیری با طبقه‌بندی و با انتساب متناسب

در این حالت کوکران (۱۹۷۷) و شیفر (۱۹۹۰) از رابطه زیر برای تعیین اندازه نمونه مورد نیاز برای برآورد میانگین استفاده کرده‌اند:

$$n_0 = \frac{1}{V_0} \sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2$$

که در آن N تعداد واحدهای جامعه، N_h تعداد واحدهای جامعه در طبقه hام، σ_h^2 واریانس صفت مورد بررسی در طبقه hام، H تعداد طبقه و V_0 واریانس مطلوب در برآورد میانگین است که می‌توان آن را از رابطه $V_0 = \frac{d^2}{Z^2_{(1-\frac{\alpha}{2})}}$ بدست آورد. V_0 را می‌توان بر اساس ضریب تغییرات مطلوب در برآورد میانگین (CV_0) نیز بیان کرد که در این صورت $V_0 = CV_0\mu^2$ خواهد بود. μ و حداکثر خطای نسبی برابر $0/1Z_{(1-\frac{\alpha}{2})}$ خواهد بود. با این توضیحات می‌توان $0/1Z_{(1-\frac{\alpha}{2})}\mu$ فرمول اندازه نمونه بر اساس خطای مطلق را به صورت زیر بیان کرد:

$$n_0 = \frac{Z^2_{(1-\frac{\alpha}{2})} \sum N_h \sigma_h^2}{d^2 N}$$

در این فرمول d از همان تعابیر مربوط به نمونه‌گیری تصادفی ساده برخوردار است. همچنین اندازه نمونه بر اساس خطای نسبی از رابطه زیر به دست می‌آید:

$$n_0 = \frac{Z^2_{(1-\frac{\alpha}{2})} \sum N_h \sigma_h^2}{r^2 \mu^2 N}$$

برای برآورد مجموع نیز اندازه نمونه مورد نیاز بر اساس خطای مطلق و خطای نسبی از روابط زیر بدست می‌آید:

$$n_0 = \frac{Z^2_{(1-\frac{\alpha}{2})} N \sum N_h \sigma_h^2}{d^2}$$

$$n_{\circ} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} N \sum N_h \sigma_h^{۲}}{r^{۲} X^{۲}} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} N \sum N_h \sigma_h^{۲}}{r^{۲} N \mu^{۲}}$$

و بطور مشابه برای برآورد P اندازه نمونه بر اساس خطای مطلق و خطای نسبی از روابط زیر بدست می‌آید:

$$n_{\circ} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} \sum N_h P_h (۱ - P_h)}{d^{۲} N}$$

$$n_{\circ} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} \sum N_h P_h (۱ - P_h)}{r^{۲} P^{۲} N}$$

بنابراین تغییری که به دلیل استفاده از خطای نسبی در فرمول تعیین اندازه نمونه در نمونه‌گیری تصادفی ساده بوجود می‌آید، در اینجا مشاهده نمی‌شود.

۶ نمونه‌گیری با طبقه‌بندی و با انتساب متناسب

در این حالت فرمولهای زیر توسط کوکران (۱۹۷۷) و شیفر (۱۹۹۰) پیشنهاد شده است:

$$n = \frac{(\sum N_h \sigma_h)^{۲}}{V_{\circ} N^{۲} + \sum N_h \sigma_h^{۲}} \qquad n_{\circ} = \frac{(\sum N_h \sigma_h)^{۲}}{V_{\circ} N^{۲}}$$

بر این اساس می‌توان اندازه نمونه مبتنی بر خطای مطلق را از روابط زیر بدست آورد:

$$n = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} (\sum N_h \sigma_h)^{۲}}{d^{۲} N^{۲} + Z^{۲}_{(۱-\frac{\alpha}{۲})} \sum N_h \sigma_h^{۲}} \qquad n_{\circ} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} (\sum N_h \sigma_h)^{۲}}{d^{۲} N^{۲}}$$

و برای اندازه نمونه مبتنی بر خطای نسبی داریم:

$$n = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} (\sum N_h \sigma_h)^{۲}}{r^{۲} N^{۲} \mu^{۲} + Z^{۲}_{(۱-\frac{\alpha}{۲})} \sum N_h \sigma_h^{۲}} \qquad n_{\circ} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} (\sum N_h \sigma_h)^{۲}}{r^{۲} N^{۲} \mu^{۲}}$$

برای برآورد مجموع، اندازه نمونه مورد نیاز بر اساس خطای مطلق از روابط زیر بدست می‌آید:

$$n = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} (\sum N_h \sigma_h)^{۲}}{d^{۲} + Z^{۲}_{(۱-\frac{\alpha}{۲})} \sum N_h \sigma_h^{۲}} \qquad n_{\circ} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})} (\sum N_h \sigma_h)^{۲}}{d^{۲}}$$

این اندازه نمونه بر اساس خطای نسبی عبارتست از:

$$n = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})}(\sum N_h\sigma_h)^{۲}}{r^{۲}X^{۲} + Z^{۲}_{(۱-\frac{\alpha}{۲})}\sum N_h\sigma_h^{۲}} \qquad n_{\circ} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})}(\sum N_h\sigma_h)^{۲}}{r^{۲}X^{۲}}$$

که اگر در روابط بالا X را برابر $N\mu$ بگیریم، اندازه نمونه بر اساس خطای نسبی برای برآورد مجموع برابر با اندازه نمونه بر اساس خطای نسبی برای برآورد میانگین خواهد بود. به همین ترتیب اندازه نمونه مورد نیاز بر اساس خطای مطلق برای برآورد P از روابط زیر بدست می‌آید:

$$n = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})}\left(\sum N_h\sqrt{P_h(۱-P_h)}\right)^{۲}}{d^{۲}N^{۲} + Z^{۲}_{(۱-\frac{\alpha}{۲})}\sum N_hP_h(۱-P_h)}$$

$$n_{\circ} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})}\left(\sum N_h\sqrt{P_h(۱-P_h)}\right)^{۲}}{d^{۲}N^{۲}}$$

و اندازه نمونه مورد نیاز بر اساس خطای نسبی برای برآورد P نیز عبارتست از:

$$n = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})}\left(\sum N_h\sqrt{P_h(۱-P_h)}\right)^{۲}}{r^{۲}N^{۲}P^{۲} + Z^{۲}_{(۱-\frac{\alpha}{۲})}\sum N_hP_h(۱-P_h)}$$

$$n_{\circ} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})}\left(\sum N_h\sqrt{P_h(۱-P_h)}\right)^{۲}}{r^{۲}N^{۲}P^{۲}}$$

اگر برای برآورد پارامتر مورد نظر Θ از نمونه‌گیری خوشه‌ای استفاده کنیم، تعبیر حداکثر خطای قابل پذیرش همانند نمونه‌گیری تصادفی ساده است زیرا در نمونه‌گیری خوشه‌ای ابتدا اندازه نمونه از طریق فرمولهای مربوط به نمونه‌گیری تصادفی ساده محاسبه می‌شود و سپس بر اساس متوسط اندازه خوشه و ضریب همبستگی درون خوشه‌ای تعدیل می‌گردد. در برآورد نسبتی نیز اگر برآورد نسبت $R = \frac{\bar{Y}}{\bar{X}}$ مد نظر باشد، اندازه نمونه مورد نیاز بر اساس خطای مطلق از رابطه زیر [شیفر و همکاران (۱۹۹۰)] بدست می‌آید:

$$n_{\circ} = \frac{Z^{۲}_{(۱-\frac{\alpha}{۲})}\sigma^{۲}\mu_x}{d^{۲}}$$

که d همان تعبیرهای مربوط به نمونه‌گیری تصادفی ساده را دارا است.
همچنین اگر صفت مورد بررسی دارای توزیع غیر نرمال باشد، می‌توان بر اساس فرمول میانگین و واریانس این توزیعها فرمول تعیین اندازه نمونه را برای توزیعهای خاص به صورت تقریبی بدست آورد. برای مثال، در توزیع پواسون رابطه $\sigma^۲ = \mu$ برقرار است. بنابراین اندازه نمونه بر اساس خطای نسبی از رابطه $n = \frac{Z(CV)}{r}$ بدست می‌آید. بنابراین می‌توان اظهار داشت با تغییر توزیع، فرمول تعیین اندازه نمونه تغییر می‌کند ولی تعبیر حداکثر خطای قابل پذیرش به قوت خود باقی است. در مورد آزمون فرضهای پیچیده‌تر مانند $H_۰ : \mu_۱ = \mu_۲$ یا $H_۰ : P_۱ = P_۲$ نیز این امر صادق است. برای جزئیات بیشتر در این زمینه می‌توانید به کتاب رن بیل[11] (۲۰۰۲) مراجعه کنید.

۷ نتیجه‌گیری

- برای برآورد میانگین و مجموع اندازه مورد نیاز بر اساس خطای مطلق تفاوتی با اندازه مورد نیاز بر اساس خطای نسبی ندارد ولی برای برآورد P استفاده از خطای نسبی موجب تغییر اساسی در فرمول اندازه نمونه می‌شود.
- استفاده از خطای نسبی در تعیین اندازه نمونه برای برآورد P فقط برای صفات نادر قابل توجیه است.
- در آمارگیریهایی که به صورت دوره‌ای (سالیانه) اجرا می‌شوند، استفاده از خطای نسبی مناسبتر است.
- در مواردی که جامعه آماری به چند زیرجامعه شکسته می‌شود و برای هر زیر جامعه برآورد جداگانه‌ای مورد نیاز است، استفاده از خطای نسبی مناسبتر است.
- استفاده از خطای نسبی موجب می‌شود در تعیین اندازه نمونه، به جای واریانس از ضریب تغییرات استفاده شود که نسبت به واریانس از پایایی بیشتری برخوردار است.
- در استفاده از خطای نسبی باید توجه داشت حداکثر خطای نسبی باید در حدی باشد که خطای مطلق متناظر با آن قابل پذیرش باشد.

مراجع

[1] Cochran, William G. (1977), Sampling Techniques. New York: John Wiley.

[2] Cohen, J. (1988), Statistical Power Analysis For The Behaviural Sciences. New York: Academic Press.

11) Ran Belle

[3] Deming, W. E. (1966), Some Theory of Sampling. New York: John Wiley.

[4] Hansen, M. H. Horwitz, W. N. and Madow, W. G. (1953), Sample Survey Methods and Theory. New York: John Wiley.

[5] Kish, L. (1995), Survey Sampling. New York: John Wiley.

[6] Lenth, R. V. (2001), Some Practical Guidelines for Effective Sample Size Determination. The American Statistician, Vol. 55, No. 3, 187-193.

[7] Lohr, S. L. (1999), Sampling Design and Analysis. Duxbury Press.

[8] Ran Belle, G. (2002), Statistical Rules of Thump. New York: John Wiley.

[9] Scheaffer, R. L. Mendenhall, W. and Ott, L. (1990), Elementary Survey Sampling. Boston: PWS-KENT.

[10] Thompson, S. K. (2002), Sampling. New York: John Wiley.

[11] Zakkula, G. (1999), Elements of Sampling Theory and Methods. Prentice Hall.

مطالعه آنتروپی باقیمانده توزیع‌های طول عمر

یونس زهره‌وند[۱]، مجید اسدی[۲]

[۱] دانشجوی کارشناسی ارشد آمار کاربردی دانشگاه اصفهان
[۲] عضو هیئت علمی دانشگاه اصفهان

چکیده: در مبحث قابلیت اعتماد برای مطالعه طول عمر یک سیستم، تاکنون معیارهای مختلفی از قبیل؛ تابع نرخ شکست، میانگین باقیمانده عمر و ... توسط محققین معرفی شده‌اند و بر اساس هر یک از این معیارها کلاسهای متفاوتی از توزیعها ارائه شده و مطالعات فراوانی روی هرکدام از این کلاسها صورت گرفته است. در این مقاله با استفاده از معیار میزان عدم حتمیت یک متغیر تصادفی طول عمر و مباحث موجود در نظریه اطلاع دوکلاس جدید از توزیع‌های طول عمر را معرفی می‌کنیم و خواص آنها را مورد بررسی قرار می‌دهیم.

واژه‌های کلیدی: نظریه اطلاع، آنتروپی شانون، توزیع‌های طول عمر، میانگین باقیمانده عمر، تابع نرخ شکست، تابع نرخ شکست معکوس

۱ مقدمه

فرض کنید T یک متغیر تصادفی مطلقاً پیوسته و نامنفی با تابع توزیع $F(t)$ و تابع چگالی احتمال $f(t) = F'(t)$ باشد. در مباحث نظریه قابلیت اعتماد اندازه‌های متفاوتی برای مدل‌سازی و تجزیه و تحلیل داده‌های مورد مطالعه معرفی شده‌اند. مهمترین آنها عبارت است از تابع قابلیت اعتماد که آن را با $\bar{F}(t)$ نمایش می‌دهیم و عبارت است از: $\bar{F}(t) = P(T > t)$
یکی دیگر از اندازه‌های مطرح در قابلیت اعتماد تابع نرخ شکست می‌باشد که آن را با $\lambda_F(t)$ نمایش می‌دهیم و به صورت زیر تعریف می‌شود:

$$\lambda_F(t) = \lim_{\varepsilon \longrightarrow 0} \frac{P(t < T \leq t + \varepsilon | T > t)}{\varepsilon}$$

$$= \frac{f(t)}{\bar{F}(t)}$$

احتمال موجود در فرمول فوق در واقع احتمال از کار افتادن بلافاصله بعد از زمان t است با این فرض که تا زمان t کارکرده است. همچنین میانگین باقیمانده عمر T نیز از اندازه های مهم در

مطالعات طول عمر است که آن را با $\mu_{F(t)}$ نمایش می‌دهیم و بصورت زیر تعریف می‌شود:

$$\mu_{F(t)} = E(T-t|T>t) = \frac{\int_t^\infty \bar{F}(x)dx}{\bar{F}(t)}$$

تابع بقا توسط $\lambda_F(t)$ و $\mu_F(t)$ به ترتیب بصورت زیر نمایش داده‌می‌شود:

$$\bar{F}(t) = \exp(-\int_\circ^t \lambda_F(x)dx)$$

$$\bar{F}(t) = \frac{\mu_F(\circ)}{\mu_F(t)} \exp(-\int_\circ^t \frac{1}{\mu_F(x)}dx)$$

همچنین روابط بین λ_F و μ_F بصورت زیر است:

$$\lambda_F(t) = \frac{\mu_F'(t)+1}{\mu_F(t)}$$

میزان عدم حتمیت T بر مبنای تعریف شانون (۱۹۴۸) بصورت زیر تعریف می‌شود:

$$H(f) = -\int_\circ^\infty f(t)\log f(t)dt$$

$H(f)$ معیاری برای اندازه‌گیری میزان عدم حتمیت و یا قابلیت پیش‌بینی متغیر تصادفی T می‌باشد. هر چه مقدار $H(f)$ برای یک متغیر تصادفی بیشتر باشد، توزیع آن متغیر یکنواخت‌تر می‌باشد و پیش‌بینی برآمدهای متغیر تصادفی T سخت‌تر است و بالعکس هر چه میزان $H(f)$ کمتر باشد توزیع T متمرکز وکشیده‌تر است و پیش‌بینی مقادیر برآمدهای آتی T ساده‌تر خواهد بود. آنتروپی شانون در حالت گسسته همواره کمیتی مثبت می‌باشد ولی در حالت پیوسته می‌تواند هر مقدار حقیقی را اختیار کند.

اغلب در مطالعات طول عمر و تحلیل بقا با حالاتی روبرو هستیم که در آنها واحد مورد مطالعه طول عمری بیشتر از مقدار معین t دارد. به عبارت دیگر فرض کنید T طول عمر یک قطعه باشد. حال اگر این قطعه تا زمان t عمرکرده باشد، یعنی $T>t$ آنگاه باقیمانده عمر آن $T-t$ است. لذا در چنین حالتی $H(f)$ معیار مناسبی برای اندازه‌گیری عدم حتمیت باقیمانده عمر سیستم نمی‌باشد. در این مقاله سعی می‌کنیم مفهوم آنتروپی باقیمانده و آنتروپی گذشته متغیر تصادفی T را که به ترتیب توسط ابراهیمی (۱۹۹۵) کرسنزو و لانگوباردی (۲۰۰۲) ارایه شده است را معرفی کرده و بعضی از خواص آنها را مورد بررسی قرار دهیم. به عبارت دیگر آنتروپی متغیرهای $T-t|T>t$ و $t-T|T<t$ را مورد مطالعه قرار می‌دهیم و بعضی از خواص آنها را استخراج می‌کنیم. به عنوان مثال نشان می‌دهیم آنتروپی باقیمانده و آنتروپی گذشته تحت بعضی شرایط تابع توزیع متغیر تصادفی T را مشخص می‌کنند. همچنین رفتار آنتروپی باقیمانده و آنتروپی گذشته را در سیستم‌های موازی و متوالی مورد مطالعه قرار می‌دهیم.

۲ آنتروپی باقیمانده

قبل از معرفی آنتروپی باقیمانده مثال زیر را در نظر بگیرید.
مثال: فرض کنید متغیر تصادفی T دارای توزیع احتمالی بصورت زیر باشد:

$$p(T=t)=\begin{cases} ۱/۵ & t=۱,۲,۳,۴,۵ \\ ۰ & o.w \end{cases}$$

با توجه به فرم توزیع T پیش‌بینی برآمد متغیر تصادفی T دشوارترین حالت ممکن در بین تمام توزیع‌های گسسته ۵ مقداری است. حال فرض کنید به طریقی مطلع می‌شویم که $T \geq ۳$. در این صورت:

$$p(T=t|T\geq ۳)=\begin{cases} ۱/۳ & t=۳,۴,۵ \\ ۰ & o.w \end{cases}$$

ملاحظه می‌شود که در توزیع جدید پیش‌بینی مقدار T از بین سه مقدار ممکن $۵,۴,۳$ خواهد بود که بوضوح ساده‌تر از قبل است. یعنی آنتروپی متغیر تصادفی $(T|T\geq ۳)$ کمتر از آنتروپی T می‌باشد.

فرض کنید $f_t(x)$ تابع چگالی احتمال متغیر تصادفی $T|T>t$ باشد آنگاه

$$f_t(x)=\frac{f(x)}{\bar{F}(t)} \qquad ,x>t$$

ابراهیمی (۱۹۹۵) آنتروپی $f_t(x)$ را بصورت زیر تعریف کرد:

$$\begin{aligned} H(f;t) &= -\int_t^\infty (\tfrac{f(x)}{\bar{F}(t)})\log(\tfrac{f(x)}{\bar{F}(t)})dx \\ &= ۱-\tfrac{۱}{\bar{F}(t)}\int_t^\infty (\log\lambda_F(x))f(x)dx \qquad (۲-۱) \end{aligned}$$

$H(f;t)$ به عنوان یک معیار پویا برای اندازه گیری عدم حتمیت T، تمام خواص $H(f)$ را داراست. واضح است که

$$H(f;۰)=H(f)$$

لازم به ذکر است که: $-\infty<H(f;t)<+\infty$.
به راحتی می‌توان نشان داد که: $H(f;t)=c$ (c یک عدد ثابت است) اگر و تنها اگر T دارای توزیع نمایی باشد. ابراهیمی (۱۹۹۵) بر مبنای معیار $H(f;t)$ دو کلاس از توزیع‌های احتمال به صورت زیر تعریف کرد.

تعریف ۲.۱: متغیر تصادفی مطلقاً پیوسته و نامنفی T دارای آنتروپی باقیمانده عمر نزولی (صعودی) $DURL$ $(IURL)$ است هرگاه $H(f;t)$ تابعی نزولی (صعودی) از t باشد.

با توجه به تعریف توزیع نمایی با تابع نرخ شکست $\lambda_F(t) = c$ مرز بین توزیع‌های $IURL, DURL$ می‌باشد و این دو کلاس از توزیعها را از هم تفکیک می‌کند.

تبصره ۲.۱: فرض کنید توزیع متغیر تصادفی T در کلاس توزیع‌های $DURL$ قرارگیرد بنابراین با گذشت زمان $(t > ۰)$، توزیع متغیر تصادفی جدید $(T|T > t)$ اطلاع بخش‌تر از توزیع T خواهد بود زیرا با گذشت زمان عدم حتمیت این متغیر کمتر می‌شود.

ابراهیمی (۱۹۹۵) نتایج زیر را اثبات کرد:

نتیجه ۲.۱: اگر در توزیع F, $\mu_F(t) < \infty$ آنگاه $H(f;t) \leq ۱ + \mu_F(t)$

نتیجه ۲.۲: اگر توزیع F، $(IURL)\ DURL$ باشد آنگاه $H(f;t) \leq (\geq) ۱ - \log \lambda_F(۰)$

نتیجه ۲.۳: اگر توزیع F، $DURL$ باشد آنگاه μ $H(f;t) \leq ۱ + \log$.

یکی از خواص مهم توابعی مانند $\mu_F(t)$ و $\lambda_F(t)$ آن است که توزیع احتمال را به صورت منحصربفرد مشخص می‌کنند. اکنون این سوال مطرح است که آیا رابطه بین $H(f;t)$ و F یک به یک است. در ادامه نشان می‌دهیم که در کلاس وسیعی از توزیعها، $H(f;t)$ به طور منحصر به فرد f را مشخص می‌کند.

قضیه ۲.۱: فرض کنید X و Y دو متغیر تصادفی نامنفی و مطلقاً پیوسته به ترتیب با توابع توزیع F و G باشند و F و G از خانواده توزیع‌های $IURL$ باشند و به ازای هر t داشته باشیم

$$H(f;t) = H(g;t) \qquad (۲-۲)$$

آنگاه: $G(t) = F(t)$

اثبات : با مشتق‌گیری از طرفین تساوی $(۲-۲)$ داریم:

$$H'(f;t) = H'(g;t)$$

که این نتیجه می‌دهد:

$$\lambda_F(t)[H(f;t) - ۱ + log\lambda_F(t)] = \lambda_G(t)[H(g;t) - ۱ + log\lambda_G(t)] \qquad (۲-۳)$$

اگر به ازای هر t، $\lambda_F(t) = \lambda_G(t)$ آنگاه $G(t) = F(t)$ و قضیه اثبات می‌شود فرض کنید:

$$A = \{t; t > ۰, \lambda_F(t) \neq \lambda_G(t)\}$$

و A تهی نباشد، یعنی

$$\exists \quad t_۰ \in A : \quad \lambda_F(t_۰) \neq \lambda_G(t_۰)$$

بدون خلل به کلیت مسئله فرض کنید $\lambda_F(t_۰) > \lambda_G(t_۰)$ با توجه به صعودی بودن $H(f;t)$ و $H(g;t)$ به ازای هر t و به خصوص برای $t = t_۰$ داریم: $۰ \leq H(f;t_۰) - ۱ + log\lambda_F(t_۰)$ و $۰ \leq H(g;t_۰) - ۱ + log\lambda_G(t_۰)$. و لذا از $(۲-۳)$ نتیجه می‌گیریم:

$$H(f;t_۰) - ۱ + log\lambda_F(t_۰) \leq H(g;t_۰) - ۱ + log\lambda_G(t_۰)$$

در نتیجه $\lambda_F(t_\circ) \leq \lambda_G(t_\circ)$ و این با فرض $\lambda_F(t_\circ) > \lambda_G(t_\circ)$ تناقض دارد. بنابراین A شامل هیچ عضوی نیست و لذا به ازای هر t

$$\lambda_F(t) = \lambda_G(t)$$

که معادل است با

$$F(t) = G(t)$$

بلاک و همکارانش (۱۹۸۵) نشان دادند که اگر F و G توابع توزیع مربوط به دو متغیر تصادفی مطلقاً پیوسته X و Y با توابع نرخ خطر λ_F و λ_G باشند و به ازای هر t داشته باشیم $\lambda_G(t) = \theta(t)\lambda_F(t)$ که در آن $\circ \leq \theta(x) \leq ۱$ یک تابع صعودی (نزولی) از x است، در اینصورت اگر F، IFR، (DFR)، $(IMRL)DMRL$، $(DFRA)IFRA$، $(NWU)NBU$ باشد آنگاه G نیز آن خاصیت را خواهد داشت. اسدی و ابراهیمی (۲۰۰۰) نتیجه زیر رابرای $H(f;t)$ نشان دادند.

لم ۲.۱: فرض کنید X و Y دو متغیر تصادفی مطلقاً پیوسته با توابع توزیع F و G و توابع نرخ شکست λ_F و λ_G و آنتروپی باقیمانده عمر $H(f;t)$ و $H(g;t)$ باشند و به ازای هر $t > \circ$

$$\lambda_G(t) = \theta(t)\lambda_F(t) \qquad (۲-۴)$$

که در آن $\circ \leq \theta(t) \leq ۱$ تابعی صعودی از t است. در اینصورت اگر F، $DURL$ باشد آنگاه G نیز $DURL$ است مشروط بر آنکه $\lim_{t\longrightarrow\infty} \frac{\bar{G}(t)}{\bar{F}(t)} < \infty$

اثبات: داریم

$$H(f;t) = ۱ - \frac{۱}{\bar{F}(t)} \int_t^\infty log\lambda_F(x) f(x) dx = ۱ - E_X(\lambda_F(X) \mid X > t)$$

$$H(g;t) = ۱ - \frac{۱}{\bar{G}(t)} \int_t^\infty log\lambda_G(x) g(x) dx = ۱ - E_Y(\lambda_G(Y) \mid Y > t)$$

بنابراین برای اثبات نزولی بودن $H(g;t)$ کافی است نشان دهیم که $E_Y(\lambda_G(Y) \mid Y > t)$ تابعی صعودی از t است از طرفی از $(۴-۲)$ داریم:

$$log\lambda_G(y) = log\theta(y) + log\lambda_F(y)$$

در نتیجه

$$E_Y(\lambda_G(Y) \mid Y > t) = E_Y(\theta(Y) \mid Y > t) + E_Y(\lambda_F(Y) \mid Y > t)$$

از طرفی چون $\theta(t)$ تابعی صعودی از t است بنابراین $E_Y(\theta(Y) \mid Y > t)$ نیز تابعی صعودی از t خواهد بود. لذا برای صعودی بودن $E_Y(\lambda_G(Y) \mid Y > t)$ کافی است نشان دهیم

$E_Y(\lambda_F(Y) \mid Y > t)$ صعودی است. بدین منظور m_1 و m_2 را بصورت زیر تعریف می‌کنیم:

$$m_1(t) = E_X(\lambda_F(X) \mid X > t)$$

$$m_2(t) = E_Y(\lambda_F(Y) \mid Y > t)$$

و قرار می‌دهیم

$$\beta(t) = \bar{G}(t)[m_1(t) - m_2(t)]$$

داریم:

$$\beta'(t) = \bar{G}(t)\{[log\lambda_F(t) - m_1(t)]\lambda_G(t) + m_1'(t)\}$$

از طرفی $m_1'(t) = \lambda_F(t)[m_1(t) - log\lambda_F(t)]$ که نتیجه می‌دهد:

$$\beta'(t) = \bar{G}(t)m_1'(t)[1 - \frac{\lambda_G(t)}{\lambda_F(t)}] > 0$$

در نتیجه $\beta(t)$ تابعی صعودی است . حال با فرض آنکه $\lim_{t\longrightarrow\infty} \frac{\bar{G}(t)}{\bar{F}(t)} < \infty$ داریم:

$$\lim_{t\longrightarrow\infty} \beta(t) = \lim_{t\longrightarrow\infty} [\frac{\bar{G}(t)}{\bar{F}(t)} \int_t^\infty log\lambda_F(x)f(x)dx - \int_t^\infty log\lambda_F(x)g(x)dx] = 0$$

و بنابراین به ازای هر t، $\beta(t) \leq 0$ و $m_1(t) \leq m_2(t)$ و چون $\lambda_F(t) \leq m_1(t)$ در نتیجه به ازای هر t، $\lambda_F(t) \leq m_2(t)$ و بنابراین $m_2(t)$ تابعی صعودی است و بنابراین $H(g;t)$ نیز نزولی است.

در مباحث قابلیت اعتماد یک روش برای افزایش قابلیت سیستم، اضافه کردن مؤلفه بیشتر به سیستم است. یکی از مهمترین ساختارهای سیستم‌هایی که دارای مؤلفه اضافی هستند، سیستم‌های k ازn می‌باشد.

سیستم k از n: سیستمی که عملکرد آن مستلزم عملکرد حداقل k تا از کل n مولفه‌اش می‌باشد. در حالت خاص اگر $k = n$ سیستم را سری گوییم. و در حالت $k = n$ سیستم را موازی گوییم.

فرض کنید $T_1, T_2, ..., T_n \overset{iid}{\sim} F(.)$ طول عمر n مولفه یک سیستم را نشان دهند و $T_{(1)}, ..., T_{(n)}$ آماره‌های مرتب نظیر این نمونه باشند. در این صورت $T_{(k)}$ توزیع طول عمر یک سیستم $(n - k + 1)$ از n را نشان می‌دهد و توابع چگالی و توزیع و نرخ شکست طول

عمر این سیستم به ترتیب بصورت زیر می‌باشد.

$$f_{(k)}(t) = \frac{n!}{(k-1)!(n-k)!}F(t)^{k-1}f(t)\bar{F}(t)^{n-k}$$

$$F_{(k)}(t) = \sum_{i=k}^{n}\binom{n}{i}F(t)^i\bar{F}(t)^{n-i}$$

$$\lambda_{F_{(k)}}(t) = \frac{f_{(k)}(t)}{\bar{F}_{(k)}(t)}$$

$$= \frac{n!}{(k-1)!(n-k)!}.\lambda_F(t).\frac{(F(t)/\bar{F}(t))^{k-1}}{\sum_{i=0}^{k-1}\binom{n}{i}(\frac{F(t)}{\bar{F}(t)})^i}$$

از لم (۱ − ۲) قضیه زیر نتیجه می‌شود.

قضیه ۲.۲: اگر $T_1,...,T_n \overset{iid}{\sim} F$ با تابع چگالی f و تابع نرخ شکست $\lambda_F(t)$ و آنتروپی باقیمانده عمر نزولی $H(f;t)$ باشند، آنگاه $H(f_n;t)$ (آنتروپی باقیمانده عمر آمارهٔ مرتب nام) نیز نزولی است.

اثبات: داریم

$$\lambda_{F_{(n)}}(t) = \frac{n(\frac{F(t)}{\bar{F}(t)})^{n-1}}{\sum_{i=0}^{n-1}\binom{n}{i}(\frac{F(t)}{\bar{F}(t)})^i}\lambda_F(t)$$

$$\theta(t) = n\frac{(F(t)/\bar{F}(t))^{n-1}}{\sum_{i=0}^{n-1}\binom{n}{i}(F(t)/\bar{F}(t))^i}$$

اگر قرار دهیم

$$\theta(t) = n\frac{(F(t)/\bar{F}(t))^{n-1}}{\sum_{i=0}^{n-1}\binom{n}{i}(F(t)/\bar{F}(t))^i}$$

آنگاه به راحتی ملاحظه می‌شود که $0 \le \theta(t) \le 1$ و $\theta(t)$ تابعی صعودی از t است. بنابراین با استفاده از لم قبل $H(f_n;t)$ هم تابعی نزولی از t است یعنی $F_{(n)}(t)$، $DURL$ است.

در واقع قضیهٔ فوق بیان می‌کند که اگر یک سیستم موازی با n مولفه با توزیع طول عمر مشترک F و آنتروپی باقیمانده نزولی داشته باشیم آنگاه آنتروپی باقیمانده عمر کل سیستم نیز نزولی خواهد بود.

۳ آنتروپی گذشته

در بسیاری از حالات سیستم‌های مورد استفاده بصورت پیوسته مورد بازرسی قرار نمی‌گیرند. اکنون فرض کنید قطعه‌ای با طول عمر T را در زمان $t = ۰$ بکارگرفته‌ایم و در زمان $t > ۰$ مورد بازرسی قرار داده‌ایم و مشاهده نموده‌ایم که سیستم ازکار افتاده است. به عبارت دیگر طول عمر سیستم کمتر از t است یعنی $T < t$. اکنون این سوال مطرح است که آنتروپی $t - T$ (یعنی زمان ازکار افتادگی قطعه تازمان t) $(t - T|T < t)$ را بصورت زیر تعریف کردند:

$$\bar{H}(f;t) = -\int_{۰}^{t} (\frac{f(x)}{F(t)}) \log(\frac{f(x)}{F(t)}) dx$$

$$= ۱ - \frac{۱}{F(t)} \int_{۰}^{t} (\log r_F(x)) f(x) dx \qquad (۳-۱)$$

که در آن $y < t$, $\frac{f(y)}{F(t)}$ تابع چگالی $(T|T < t)$ است و $r_F(t) = \frac{f(t)}{F(t)}$ تابع نرخ شکست معکوس می‌باشد. در زیر نشان می‌دهیم که $\bar{H}(f;t)$ تحت شرایطی F را بصورت منحصربفرد مشخص می‌کند. به عبارت دیگر یک رابطه یک به یک بین F و $\bar{H}(f;.)$ وجود دارد.

قضیه ۳.۱: فرض کنید X و Y دو متغیر تصادفی نامنفی و مطلقاً پیوسته به ترتیب با توابع توزیع F و G باشند و به ازای هر t، $\bar{H}(f;t) = \bar{H}(g;t)$ آنگاه درکلاس همه توزیعهایی که $\bar{H}$ آنها تابعی نزولی از t است، داریم:

$$F(t) = G(t)$$

اثبات: با مشتق‌گیری از طرفین تساوی فوق داریم:

$$\bar{H}'(f;t) = \bar{H}'(g;t)$$

$$r_F(t)[۱ - \bar{H}(f;t) - log r_F(t)] = r_G(t)[۱ - \bar{H}(g;t) - log r_G(t)] \qquad (۳-۱)$$

ازآنجایی که r_F از اندازه‌های مشخصه‌ساز توزیع احتمال است اگر به ازای هر t، $r_F(t) = r_G(t)$ آنگاه $F(t) = G(t)$ و قضیه اثبات می‌شود فرض کنید

$$\exists \quad t_۰ > ۰ \quad r_F(t_۰) \neq r_G(t_۰)$$

بدون خلل به کلیت مسئله فرض کنید $r_F(t_\circ) > r_G(t_\circ)$ بنابراین با توجه به رابطه (۲ − ۳) و نزولی بودن $\bar{H}(f;t)$ به ازای هر t و به خصوص $t = t_\circ$ داریم:

$$1 - \bar{H}(f;t_\circ) - log r_F(t_\circ) > 1 - \bar{H}(g;t_\circ) - log r_G(t_\circ)$$

و یا

$$1 - \bar{H}(f;t_\circ) - log r_F(t_\circ) = 1 - \bar{H}(g;t_\circ) - log r_G(t_\circ)$$

که از دو رابطهٔ فوق به راحتی نتیجه می‌شود $r_F(t_\circ) \leq r_G(t_\circ)$ و این با فرض $r_F(t_\circ) > r_G(t_\circ)$ تناقض دارد. بنابراین به ازای هر t

$$r_F(t) = r_G(t)$$

که معادل است با

$$F(t) = G(t)$$

نتیجه مهم زیر را برای آنتروپی گذشته نشان می‌دهیم.

لم ۳.۱: فرض کنید X و Y دو متغیر تصادفی مطلقاً پیوسته با توابع توزیع F و G و توابع نرخ شکست معکوس r_F و r_G باشند و داشته باشیم:

$$\forall \ t \geq \circ; \quad r_G(t) = \theta(t) r_F(t)$$

که در آن $\circ \leq \theta(t) \leq 1$ تابعی نزولی از t باشد. در اینصورت اگر $\bar{H}(f;t)$ تابعی صعودی از t باشد آنگاه $\bar{H}(g;t)$ نیز چنین است مشروط بر آنکه $\lim_{t \longrightarrow \circ} \frac{G(t)}{F(t)} < \infty$.

اثبات: مشابه اثبات لم (۳.۱) می‌باشد و از آوردن آن صرفنظر می‌شود.

در قضیه زیر رفتار $\bar{H}(f;t)$ یک سیستم موازی را بر اساس آنتروپی گذشته مولفه‌هایش تحت شرایطی خاص بررسی می‌کنیم.

قضیه ۳.۲: اگر $T_1, ..., T_n \overset{iid}{\sim} F$ با تابع چگالی مشترک f و تابع نرخ شکست معکوس $r_F(t)$ و آنتروپی گذشته صعودی $\bar{H}(f;t)$ باشند. آنگاه $\bar{H}(f_1;t)$ (آنتروپی گذشتهٔ آمارهٔ مرتب اول) نیز صعودی است.

اثبات: اگر $T_{(1)} = min\{T_1, ..., T_n\}$ آنگاه داریم

$$r_{T_{(1)}}(t) = \frac{f_{T_{(1)}}(t)}{F_{T_{(1)}}(t)}$$

$$= \frac{f(t)}{F(t)} . \frac{n\bar{F}(t)^{n-1}}{\sum_{i=\circ}^{n-1} \bar{F}(t)^i}$$

با قرار دادن

$$\theta(t) = \frac{n\bar{F}(t)^{n-1}}{\sum_{i=0}^{n-1} \bar{F}(t)^i}$$

بوضوح $0 \leq \theta(t) \leq 1$ و $\theta(t)$ تابعی نزولی از t است بنابراین با استفاده از لم قبل $\bar{H}(f_{(1)};t)$ تابعی صعودی از t است و آنتروپی گذشته $F_{(1)}(t)$ هم صعودی است.

در واقع قضیهٔ فوق بیان می‌کند که اگر یک سیستم سری با n مولفه با توزیع طول عمر مشترک F و آنتروپی گذشته صعودی داشته باشیم آنگاه آنتروپی گذشته کل سیستم نیز صعودی خواهد بود.

مراجع

[1] Ebrahimi, N.(1996). How to measure uncerttainty in the residual life time distribution. *Sankhya*, vol 58 p;48-56.

[2] Asadi, M. Ebrahimi, N.(2000).Residual entropy and its characterizaations in terms of hazard function and mean residual life function. *Statistics and Probability letters*, vol 49 P;263-269.

[3] DI Cresenzo,A. Longobardi,M.(2002) Entropy-based measure of uncertainty in past lifetime distributions.*J. Appl. Prob.* vol 39 P;434-440.

[4] Ebrahimi, N. Pellerey, F.(1995) New partial ordering of survival functions based on the notion of uncerttainty.*Appl. Prob.* vol 32 p;202-211.

یک الگوریتم مقدماتی برای شبیه‌سازی عددی معادلات دیفرانسیل تصادفی

پرویز سرگلزایی[۱]، محمد امینی[۱]، محمود دادخواه[۲]

[۱] دانشگاه سیستان و بلوچستان، گروه ریاضی
[۲] دانشجوی کارشناسی ارشد ریاضی دانشگاه سیستان و بلوچستان

چکیده: در این مقاله چند روش عددی سودمند برای حل معادلات دیفرانسیل تصادفی ارایه می‌شود. انتگرالهای تصادفی، روشهای اولر ـ مارویاما و ملشتاین برای حل عددی معادلات دیفرانسیل تصادفی و همگرایی قوی و ضعیف مورد توجه قرارگرفته است. در ادامه چند مثال عددی هم ارایه می‌گردد.

واژه‌های کلیدی: روش اولر ـ مارویاما، روش ملشتاین، روش مونت کارلو، همگرایی ضعیف و قوی، بسط تصادفی تیلور

۱ مقدمه

امروزه مدلهای معادلات دیفرانسیل تصادفی نقشی بسیار مهم و برجسته در حوزه‌های مختلف علوم ایفا می‌کنند. از جمله‌ی این علوم می‌توان به زیست شناسی، شیمی، مکانیک، اقتصاد، پزشکی، الکترونیک، ریاضی و ... اشاره کرد. تفهیم و آشنایی کامل با نظریه‌ی معادلات دیفرانسیل تصادفی منوط به آشنایی پیشرفته با احتمال و فرایندهای تصادفی است اما فهم شبیه‌سازی عددی مورد بحث در این مقاله برای معادلات دیفرانسیل تصادفی، نیازمند آشنایی و داشتن زمینه‌ای در مورد روش اولر برای حل عددی معادلات دیفرانسیل معمولی و فهم شهودی متغیرهای تصادفی است. بعلاوه آشنایی با روشهای عددی حل معادلات دیفرانسیل معمولی، گام مفید و موثر اولیه برای کار با معادلات دیفرانسیل تصادفی را در اختیار ما قرار می‌دهد. بهر حال ما در این مقاله فرض کرده‌ایم که خواننده حداقل به طور سطحی با مفاهیمی از قبیل متغیرهای تصادفی، استقلال، مقادیر مورد انتظار (امیدها)، واریانس و بخصوص مفهوم متغیرهای تصادفی با توزیع نرمال آشناست. آزمایشها و نتایج عددی ما از روش مونت کارلو اقتباس شده است (که بیان می‌دارد: متغیرهای تصادفی به کمک تولید اعداد تصادفی و مقادیر مورد انتظار با میانگینهای محاسبه شده، قابل حصول‌اند). برای اجرای هر چه سریع‌تر و بهتر برنامه‌ها در $Matlab$، یک تکنیک جدید ارایه کرده‌ایم و سعی کرده‌ایم کدهای برنامه را به صورت برداری در آورده و اجرا نماییم.

۲ حرکت براونی (فرایند وینر)

حرکت براونی استاندارد یا فرایند وینر استاندارد روی $[۰, T]$، یک متغیر تصادفی $W(t)$ است که بطور پیوسته به $t \in [۰, T]$ وابسته بوده و در سه شرط ذیل هم صدق می نماید:

۱) $W(۰) = ۰ \quad (w.p_{/}۱)$

۲) برای هر $۰ \leq s < t \leq T$، متغیرهای تصادفی داده شده با نمو $W(t) - W(s)$، توزیع نرمال با میانگین صفر و واریانس $t - s$ داشته باشند به عبارت دیگر داشته باشیم:
$$W(t) - W(s) \sim N(۰, t - s)$$

۳) برای هر $۰ \leq s < t < u < v \leq T$، نموهای $W(t) - W(s)$ و $W(v) - W(u)$ مستقل از هم باشند.

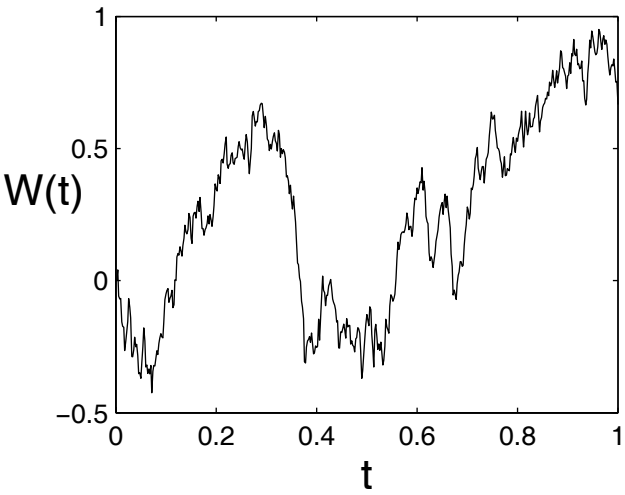
برای کارهای محاسباتی، در نظرگرفتن حرکت براونی گسسته‌سازی شده که در آن $W(t)$ در مقادیر گسسته‌سازی شده‌ی t معین شده است، مناسب است. بنابراین فرض می کنیم $\delta t = \frac{T}{N}$ (N عدد صحیح) و نیز $t_j = j\delta t$ ، $W_j = W(t_j)$. بنابر شرط (۲) و (۳) داریم:

$$W_j = W_{j-۱} + dW_j \quad , \quad j = ۱, ۲, ..., N \qquad (۱)$$

که هر dW_jیی یک متغیر تصادفی مستقل به فرم $N(۰, \delta t)$ می‌باشد. برنامه‌ی $Matlab$ ذیل شبیه‌سازیهای ساده و متفاوتی را با توجه به مقادیر اخذ شده از کاربر برای فرایند وینر روی $[۰, T]$ می‌دهد. از تابع مولد اعداد تصادفی $randn$ استفاده کرده‌ایم تا در هر بار بازخوانی آن عددی تصادفی با توزیع $N(۰, ۱)$ داشته باشیم. از وضعیت $state$ هم برای دریافت نتایج عددی مشابه در هر تکرار استفاده شده است. برای تکرار پذیر بودن نتایج عددی، مقدار اولیه را هم از کاربر دریافت کرده‌ایم اما معمول آنست که آنرا صفر در نظر بگیریم. ما نمودار را برای $T = ۱$ و $N = ۵۰۰$ رسم کرده‌ایم.

```
randn('state',100)
T=input(' final time:');
N=input(' number of divisions:');
dt=T/N;   dW=zeros(1,N);
W=zeros(1,N);
w(1)=dW(1);
for j=2:N
dW(j)=sqrt(dt)*randn;
W(j)=W(j-1)+dW(j);
end
subplot(2,2,1);
plot([0:dt:T],[0,W],'r-')
xlabel('t','FontSize',16) , ylabel('W(t)','FontSize',16,'Rotation',0)
```

*** win.m : Brownian Path Simulation ***



حالا می‌خواهیم با یک تکنیک خوب، معادل برنامه‌ی قبلی را به صورتی کاراتر نوشته و از مزاحمتهای حلقه‌ی For رهایی پیدا کنیم. این کار را با جانشین سازی حلقه‌ی For با دستور سطح بالاتری بنام $Cumsum$ انجام می‌دهیم. این کار یکی از تکنیکهای بالا بردن کارایی برنامه است. jامین مولفه‌ی این دستور بصورت زیر محاسبه می‌شود:

$$dW(۱) + dW(۲) + ... + dW(j)$$

در برنامه‌ی $win۲.m$ این کار را انجام داده‌ایم. اگر شکل مربوط به این برنامه را هم رسم کنیم، دقیقاً همانند شکلی است که در حالت قبلی برای حلقه‌ی For رسم کرده بودیم، با این تفاوت که از مزاحمت حلقه‌ی For در اینجا خبری نیست.

```
randn('state',100)
T=input('Input final time:');
N=input('Input number of divisions:');
dt=T/N; dW=sqrt(dt)*randn(1,N);
W=cumsum(dW); subplot(2,2,1);
plot([0:dt:T],[0,W],'r-')
xlabel('t','FontSize',16)
ylabel('W(t)','FontSize',16,'Rotation',0)
```

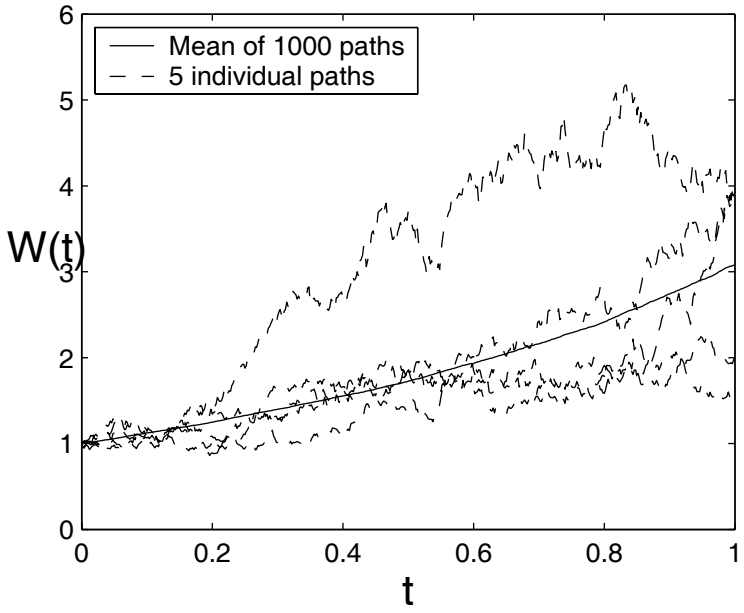
$***\ win2.m\ ***$

حالا می‌خواهیم یک تابع ساده‌ی تصادفی را شبیه‌سازی کنیم. در برنامه‌ی $bpath.m$ ذیل، تابع $U(W(t)) = e^{t+\frac{۱}{۲}W(t)}$ را حول ۱۰۰۰ مسیر براونی گسسته‌سازی شده رسم و ارزیابی کرده‌ایم.

```
randn('state',100)
T=1; N=500;
```

```
dt=T/N; t=[dt:dt:T]; M=1000;
dW=sqrt(dt)*randn(M,N);
W=cumsum(dW,2);
U=exp(repmat(t,[M 1])+0.5*W);
Umean=mean(U);
subplot(1.5,1.5,1);
plot([0,t],[T,Umean],'b-'),
hold on
plot([0,t],[ones(5,T),U(T:5,:)],...
'r–'), hold off
xlabel('t','FontSize',16)
ylabel('W(t)','FontSize',16,...
'Rotation',0)
legend('Mean of 1000 paths','5 individual paths',2)
avver=norm((Umean-exp(9*t/8)),'inf')
```

**** bpath.m : Function along a Brownian Path ****



این تابع به عنوان حالت خاصی از جوابهای معادلات دیفرانسیل خطی مطرح است. با فرض $T = ۱$ و $N = ۵۰۰$، ۵ مسیر مجزا و میانگین این مسیرها را هم رسم کرده‌ایم. برای افزایش کارایی برنامه سعی کرده‌ایم برنامه را برداری نماییم، بطوریکه dW یک آرایه $M \times N$ است که $dW(i,j)$ نمو dW_j برای iامین مسیر را می‌دهد. دستور $repmat$ برای معرفی یک آرایه‌ی $M \times N$ بکار می‌رود بطوریکه سطرهای آن تماماً نسخه‌ای از t می‌باشد. با توجه به

شکل ترسیمی در می‌یابیم که علی‌رغم ناهموار بودن $U(W(t))$ در طی مسیرهای مجرد، میانگین آن هموار به نظر می‌رسد. امید یا مقدار مورد انتظار برای $U(W(t))$ برابر $e^{\frac{۹t}{۸}}$ است. مقدار $avver = ۰/۰۵۰۴$ بدست آمده است.
هر گاه جواب یک معادله‌ی دیفرانسیل برای مسیر معینی خواسته شده باشد، گوییم جواب قوی مورد نیاز است و این منجر به مفهوم همگرایی قوی می‌شود. در مقابل هر گاه فقط اطلاعات نوع مقدار مورد انتظار (امید) در مورد جواب مورد مطالعه قرار گیرد، منجر به مفهوم همگرایی ضعیف خواهد شد.

۳ انتگرالهای تصادفی

اگر تابع مناسبی همانند f داده شده باشد، انتگرال $\int_۰^T f(t)dt$ را ممکن است با جمع سری ریمن ذیل تقریب بزنیم:

$$\sum_{j=۰}^{N-۱} f(t_j)(t_{j+۱} - t_j) \qquad (۲)$$

که نقاط گسسته‌سازی شده $t_j = j\delta t$ در بخش (۲) معرفی شدند.
هر گاه از سری (۲) وقتی $\delta t \to ۰$ حد گرفته شود، این عبارت به شکل انتگرال تعریف خواهد شد. به طریقی مشابه می‌توان مجموعی به شکل ذیل را هم در نظر گرفت:

$$\sum_{j=۰}^{N-۱} f(t_j)(W(t_{j+۱}) - W(t_j)) \qquad (۳)$$

که در مقایسه با سری (۲) در بالا، می‌توان آنرا به عنوان تقریبی برای انتگرال تصادفی $\int_۰^T f(t)dW(t)$ در نظر گرفت (انتگرال ایتو). از طرفی می‌توان یک تقریب دیگر نیز برای سری (۲) به فرم زیر، برای بدست آوردن انتگرال، در نظر گرفت:

$$\sum_{j=۰}^{N-۱} f(\frac{t_j + t_{j+۱}}{۲})(t_{j+۱} - t_j) \qquad (۴)$$

فرم متناظر تصادفی برای تقریب (۴) می‌تواند چنین بیان شود (انتگرال استراتونوویچ):

$$\sum_{j=۰}^{N-۱} f(\frac{t_j + t_{j+۱}}{۲})(W(t_{j+۱}) - W(t_j)) \qquad (۵)$$

در حالتی که $f(t) = W(t)$ باشد، نیاز به محاسبه‌ی $W(t)$ در نقطه‌ی $t = (\frac{t_j+t_{j+1}}{2})$ داریم. می‌توان نشان داد که به ازای یک نمو مستقل $N(0, \frac{\Delta t}{4})$ و با $\frac{W(t_{j+1})+W(t_j)}{2}$، مقداری برای $W(\frac{t_j+t_{j+1}}{2})$ بدست می‌آید که در سه شرط بیان شده در ابتدای بخش (۲)، مطابقت دارد. به کمک همین روش، در برنامه‌ی $stint.m$، با تقریب انتگرالهای ایتو و نیز استراتونوویچ، مقدار خطا را در مورد انتگرال ایتو برابر $itoerr = 0/0158$ و خطا در مورد انتگرال استراتونوویچ برابر $straerr = 0/0186$ بدست آورده‌ایم.

```
randn('state',100)
T=input('Input final time:');
N=input('Input number of divisions:');
dt=T/N; dW=sqrt(dt)*randn(1,N);
W=cumsum(dW);
ito=sum([0,W(1:end-1)].*dW)
strat=sum((0.5*([0,W(1:end-1)]+W)+0.5*sqrt(dt)*randn(1,N)).*dW)
itoerr=abs(ito-0.5*(W(end)^2-T))
straterr=abs(strat-0.5*W(end)^2)
```

*** $stint.m$: *Approximate Stochastic Integrals* ***

در حالت انتگرال ایتو داریم:

$$\begin{aligned}\sum_{j=0}^{N-1} W(t_j)(W(t_{j+1}) - W(t_j)) &= \frac{1}{2}\sum_{j=0}^{N-1}\{(W(t_{j+1})^2 - W(t_j)^2) \\ &- (W(t_{j+1}) - W(t_j))^2\} \\ &= \frac{1}{2}(W(T)^2 - W(0)^2) \\ &- \frac{1}{2}\sum_{j=0}^{N-1}\{(W(t_{j+1}) - W(t_j))^2 \quad (6)\end{aligned}$$

حال جمله‌ی $\sum_{j=0}^{N-1}\{(W(t_{j+1}) - W(t_j))^2$ در (۶) دارای امید ریاضی یا مقدار مورد انتظار T و واریانس $O(\delta t)$ است. بنابراین برای δtی کوچک، این متغیر تصادفی به ثابت T نزدیک خواهد شد و بنابراین:

$$\int_0^T W(t)dW(t) = \frac{1}{2}\{(W^2(T) - T\} \qquad (7)$$

و معادل(۶) و (۷) برای انتگرال استراتونوویچ چنین می‌شود ($\Delta Z_j \sim N(۰, \frac{\Delta t}{۴})$):

$$\sum_{j=۰}^{N-۱}(\frac{W(t_{j+۱})+W(t_j)}{۲}+\Delta Z_j)(W(t_{j+۱})-W(t_j)) =$$

$$\frac{۱}{۲}(W(T)^۲-W(۰)^۲)+\sum_{j=۰}^{N-۱}\Delta Z_j\{(W(t_{j+۱})-W(t_j)) \qquad (۸)$$

جمله‌ی دوم در (۸)، دارای امید ریاضی ۰ و واریانس $O(\delta t)$ است. بنابراین داریم:

$$(Stra..) \quad \int_۰^T W(t)dW(t)=\frac{۱}{۲}W^۲(T)$$

هر کدام از انتگرالهای ایتو و استراتونوویچ کاربرد خاص خود را در علوم مختلف دارند و بسته به مواقع نیاز به کار گرفته می‌شوند، با این وجود، این انتگرالها قابل تبدیل به یکدیگر هستند و می‌توان با داشتن یکی از آنها، فرم دیگری را بدست آورد.

۴ روش اولر ـ مارویاما

معادله‌ی دیفرانسیل تصادفی زیر به فرم دیفرانسیلی آن را در نظر بگیرید:

$$dX(t)=f(X(t))dt+g(X(t))dW(t) \qquad X(۰)=X_۰, \quad ۰\le t\le T \quad (۹)$$

که می‌توان آنرا به فرم انتگرالی زیر هم نوشت:

$$X(t)=X(t_۰)+\int_۰^t f(X(s))ds+\int_۰^t g(X(s))dW(s) \qquad ۰\le t\le T(۱۰)$$

که در آن f و g توابع اسکالری و شرط اولیه‌ی $X_۰$ یک متغیر تصادفی می‌باشد. هنگام بررسی (۱۰)، انتگرال دوم در سمت راست آن، باید با توجه به فرایند براونی (وینر) که در بخش قبلی توضیح دادیم، برآورد گردد. در این مقاله ما از شکل ایتو استفاده خواهیم کرد. جواب $X(t)$ برای هر t، یک متغیر تصادفی خواهد بود. از توضیح بیشتر در مورد تمام خواص جواب $X(t)$ برای (۱۰) خودداری می‌نماییم و بجای آن روشهای عددی حل (۱۰) را معرفی می‌نماییم. بنابراین جواب $X(t)$ را متغیر تصادفی حاصل از میل دادن حد طول گام به صفر، در نظر می‌گیریم. برای راحتی کار، بیشتر با فرم (۹) کار خواهیم کرد. (توجه کنید چون فرایند وینر با احتمال ۱، هیچ جا مشتق‌پذیر نیست، اجازه نداریم از $\frac{dW}{dt}$ استفاده کنیم.)

حال اگر $g \equiv 0$ و X_0 ثابت باشد، آنگاه معادله از حالت تصادفی به حالت جبری تغییر فرم می‌دهد. اکنون برای اعمال یک روش عددی بر (۹) روی بازه‌ی $[0,T]$، ابتدا بازه را گسسته‌سازی می کنیم. فرض کنید $\Delta t = \frac{T}{N}$ (N عدد صحیح) و $t'_j = j\Delta t$ باشد. تقریب عددی برای $X(t'_j)$ را با X_j نمایش داده و روش اولر ـ مارویاما را چنین معرفی می‌نماییم: $(j = 1, 2, ..., N)$

$$X_j = X_{j-1} + f(X_{j-1})\Delta t + g(X_{j-1})(W(t'_j) - W(t'_{j-1})) \qquad (11)$$

در این مقاله، مسیرهای براونی گسسته‌سازی شده را محاسبه کرده و از آنها برای حصول نمو $W(t'_j) - W(t'_{j-1})$ که در (۱۱) لازم داریم، استفاده خواهیم کرد. برای راحتی کار در محاسبات کامپیوتری، تقریباً همه جا طول گام Δt برای نتایج عددی را، مضرب $R \geq 1$ از نمو δt مسیر براونی فرض خواهیم کرد. این کار به ما اطمینان می‌دهد که مجموعه نقاط $\{t_j\}$ که مبنای گسسته‌سازی مسیر براونی است، شامل مجموعه نقاط $\{t'_j\}$ که جواب روش اولر ـ مارویاما در آن محاسبه شده است، می‌باشد. به هرحال گاهی مسیر براونی به عنوان قسمتی از داده‌ی مسئله بیان شده است اما اگر مسیر تحلیلی بکار گرفته شود، Δt به اندازه‌ی دلخواه کوچک می‌تواند بکار گرفته شود. حالا روش اولر ـ مارویاما را برای حل معادله‌ی اسکالری ذیل بکار می‌بندیم:

$$dX(t) = \lambda X(t)dt + \mu X(t)dW(t), \qquad X(0) = X_0 \qquad (12)$$

فرض می‌کنیم که λ و μ ثابتهای حقیقی‌اند. این نوع معادله‌های اسکالری در بسیاری از علوم کاربرد دارند و من جمله می‌توان از آن به عنوان مدل ارزش سرمایه در ریاضیات مالی یا مدل رشد سهام در بازار بورس استفاده کرد [۴]. جواب واقعی این نوع معادلات را می‌توان به صورت زیر بیان کرد [۸]:

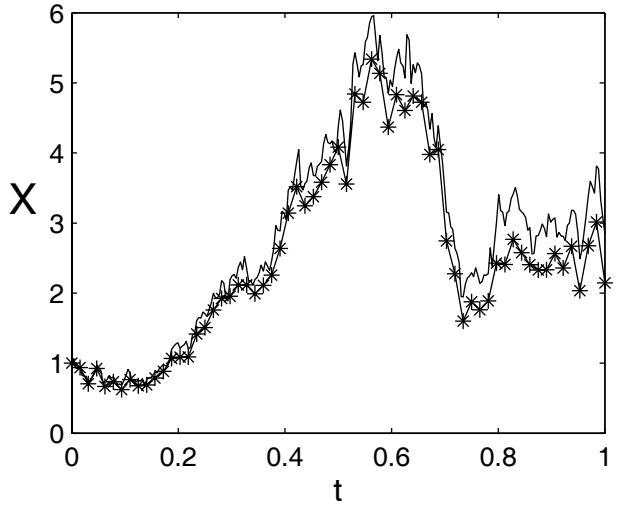
$$X(t) = X(0)e^{(\lambda - \frac{1}{2}\mu^2)t + \mu W(t))} \qquad (13)$$

در برنامه‌ی ذیل، $\lambda = 2$ و $\mu = 1$ و $X_0 = 1$ در نظر گرفته‌ایم. مسیر براونی گسسته‌سازی شده حول $[0,1]$ با $\delta t = 2^{-8}$ محاسبه شده است. بعلاوه جواب واقعی [یعنی (۱۳)] را در $Xtrue$ قرار داده‌ایم. ضریب نمو یعنی $R = 4$ قرار داده شده است.

```
randn('state',100)
λ=2; μ=1; X0=1; T=1;
N=2^8; dt=T/N;
dW=sqrt(dt)*randn(1,N);
W=cumsum(dW);
Xtrue=X0*exp((lambda-0.5*mu^2)...
*([dt:dt:T])+mu*W);
subplot(1.4,1.4,1);
plot([0:dt:T],[X0,Xtrue],'m-'),...
hold on
R=4; Dt=R*dt; l=N/R;
```

```
Xem=zeros(1,l); Xtemp=X0;
for j=1:l
Winc=sum(dW(R*(j-1)+1:R*j));
Xtemp=Xtemp+Dt*lambda*Xtemp+mu*Xtemp*Winc;
Xem(j)=Xtemp;
end
subplot(1.4,1.4,1);
plot([0:Dt:T],[X0,Xem],'r-*'), hold off
xlabel('t','FontSize',12), ylabel('X','FontSize',16,'Rotation',0,)
emerr=abs(Xem(end)-Xtrue(end))
```

**** em.m : Euler − Maroyama method for Scalar SDE ***



با تابع $emerr$ مقدار خطا را در نقطه‌ی انتهایی ارزیابی کرده‌ایم و در اینجا مقدار آنرا ۰٫۶۹۰۷ بدست آورده‌ایم. با اخذ مقادیر کوچکتری مثل $R = ۲$ و $R = ۱$ خطا را بصورت ۰٫۱۵۹۵ و ۰٫۰۸۲۱ بدست می‌آوریم. توجه کنید که در گام کلی، روش اولر ـ مارویاما به نمو $W(t'_j) - W(t'_{j-۱})$ نیاز دارد که با عبارت زیر محاسبه شده است:

$$W(t'_j) - W(t'_{j-۱}) = W(jR\delta t) - W((j-۱)R\delta t) = \sum_{k=jR-R+۱}^{jR} dW_k$$

در برنامه‌ی نوشته شده دستور $Winc = sum(dW(R*(j-۱)+۱ : R*j))$ این کار را انجام می‌دهد.

۵ همگرایی قوی و ضعیف روش اولر ـ مارویاما

اگر برنامه‌ی روش اولر ـ مارویاما (یعنی $em.m$) را با Δtهای کوچکتری بیازماییم، در می‌یابیم که هر قدر Δt کوچکتر می‌شود، جواب روش اولر ـ مارویاما سازگاری و نزدیکی بیشتری با جواب واقعی پیدا می‌کند. با توجه به اینکه، $X(t'_n)$ و X_n متغیرهای تصادفی هستند، برای بیان دقیق و واضح مفهوم همگرایی باید تصمیم بگیریم که چگونه تفاضل بین آنها را اندازه‌گیری کنیم. اگر از $E|X_n - X(t'_n)|$، که E امید ریاضی را می‌نمایاند، استفاده شود به مفهوم همگرایی قوی منجر می‌شود.
گوییم یک روش مرتبه‌ی قوی همگرایی γ دارد هرگاه ثابتی همانند C چنان باشد که برای Δt های بقدرکافی کوچک داشته باشیم:

$$E|X_n - X(t')| \leq C\Delta t^{\gamma}, \quad (t' = n\delta t \in [0, T]) \qquad (۱۴)$$

به عنوان مثال اگر توابع f و g توابعی هموار بوده و در شرایط خاصی صدق کنند، ثابت می‌شود که روش اولر ـ مارویاما همگرایی قوی مرتبه‌ی $\frac{1}{2}$ دارد (این مطلب را با توجه به شکل برنامه $em.m$ می‌توان دید).
در بررسیهای عددی که انجام خواهیم داد، بر روی خطا در آخرین نقطه تمرکز می‌نماییم. پس فرض کنیم:

$$e^S_{\Delta t} = E|X_N - X(T)|, \quad N\Delta t = T \qquad (۱۵)$$

نمایانگر خطای نقطه‌ی انتهایی روش اولر ـ مارویاما(به مفهوم قوی همگرایی) باشد. اگر کران $C\Delta t^{\gamma}$ در (۱۴) با $\gamma = \frac{1}{2}$، برای هر نقطه‌ی ثابت در $[0, T]$ برقرار باشد، قطعاً در نقطه‌ی انتهایی هم برقرار است. لذا برای Δtهای بقدرکافی کوچک می‌توان داشت:

$$e^S_{\Delta t} \leq C\Delta t^{\frac{1}{2}} \qquad (۱۶)$$

در برنامه‌ی ذیل بدنبال همگرایی روش اولر ـ مارویاما برای معادله‌ای به شکل (۱۲) هستیم. به ازای $\lambda = 2$ و $\mu = 1$ و $X_0 = 1$ و $T = 1$ و $N = 2^9$، ۱۰۰۰ مسیر براونی گسسته‌سازی شده را بکار برده‌ایم. در هر مسیری روش اولر ـ مارویاما، به ازای ۵ طول گام متفاوت $\Delta t = 2^{p-1}\delta t (1 \leq p \leq 5)$ محاسبه شده است. مقدار خطا در sامین مسیر برای pامین طول گام، با عبارت $Xerr(s, p)$ بررسی شده است. تابع $mean$ میانگین مسیرها را محاسبه می‌کند طوریکه pامین مولفه از $mean(Xerr)$ تقریبی از $e^S_{\Delta t}$ به ازای $\Delta t = 2^{p-1}\delta t$ است.

```
randn('state',100)
lambda=input(' drift cofficient:');
mu=input(' diffusion cofficient:');
X0=input(' initial value:');
```

```
T=input(' final time:');
N=input(' number of divisions:');
dt=T/N; M=1000; Xerr=zeros(M,5);
for s=1:M
dW=sqrt(dt)*randn(1,N);
W=cumsum(dW);
Xtrue=X0*exp((lambda-0.5*mu^2)+mu*W(end));
for p=1:5
R=2^(p-1); Dt=R*dt; l=N/R;
Xtemp=X0;
for j=1:l
Winc=sum(dW(r*(j-1)+1:r*j));
Xtemp=Xtemp+Dt*lambda*Xtemp+mu*Xtemp*Winc;
end
Xerr(s,p)=abs(Xtemp-Xtrue);
end
end
Dtvals=dt*(2.^([0:4])); subplot(221);
loglog(Dtvals,mean(Xerr),'b*-'), hold on
loglog(Dtvals,(Dtvals.^(.5)),'r–'), hold off
axis([1e-3 1e-1 1e-4 1])
xlabel('\Delta t'), ylabel('Sample average of |X(T) - X_l| ')
title('emstrong.m','FontSize',10)
•• Least squares fit of  error=C*Dt^q••
A=[ones(5,1), log(Dtvals)']; rhs=log(mean(Xerr)');
sol=A\rhs, q=sol(2), resid=norm(A*sol-rhs)
```

$***\ ems.m : Strong\ convergence\ of\ Euler - Maroyama\ method\ ***$

حالا اگر فرض کنیم نامساوی (۱۶) بطور تقریبی به تساوی برقرار باشد، داریم: $e^S_{\Delta t} \approx C\Delta t^{\frac{1}{2}}$، که اگر از طرفین آن لگاریتم بگیریم داریم:

$$\log e^S_{\Delta t} \approx \log C + \frac{1}{2}\log \Delta t \qquad (۱۷)$$

برای رسم لگاریتمی تقریبی از $e^S_{\Delta t}$ از دستور $loglog(Dtvals, mean(Xerr)$ استفاده کرده‌ایم. اگر در شکل رسم شده (شکل ۱ بالای گوشه‌ی سمت چپ) دقت شود، می‌توان دید که دو منحنی با هم به خوبی سازگارند که بیانگر دقیق و خوب بودن فرمول (۱۷) است. برای بررسی بیشتر کارایی فرمول (۱۷)، فرض کنیم که قاعده‌ی توانی $e^S_{\Delta t} = C\Delta t^q$ برای ثابتهایی همانند C و q برقرار باشد طوریکه:

$$\log e^S_{\Delta t} = \log C + q \log \Delta t$$

به کمک تقریب کمترین مربعات برای $\log C$ و q در انتهای برنامه‌ی $ems.m$، مقدار $q = ۰٫۵۳۸۴$ و باقیمانده‌ی ۰٫۰۲۶۶ بدست می‌آید که خود بیانگر اعتبار نتایج حاصله برای مرتبه‌ی قوی $\frac{1}{2}$

است.
ذکر این نکته لازم است که ما در بررسی خطای $e^S_{\Delta t}$، از منابع دیگر خطا که قابل چشم پوشی بودند، گذشته‌ایم. این منابع عبارتند از:

۱) خطای نمونه: یعنی خطایی که موقع تقریب میانگین نمونه با امید ریاضی حاصل می‌شود.

۲) خطای اریب اعداد تصادفی: یعنی خطای ذاتی مولد اعداد تصادفی

۳) خطای گرد کردن: که همان خطای گرد کردن اعداد ممیز شناور می‌باشد.

در محاسبات اصلی از میان سه مرجع خطای ذکر شده‌ی فوق، خطای نمونه از بقیه‌ی خطاها مهمتر است. از اینرو در بعضی از برنامه‌های این مقاله برای حصول مرتبه‌ی همگرایی پیش‌بینی شده مجبور شده‌ایم حجم نمونه‌ها را بسیار بزرگ و طول گام را تا حد ممکن کوچک اختیار کنیم (اگر M حجم نمونه باشد، خطای نمونه متناسب با $\frac{۱}{\sqrt{M}}$ کاهش می‌یابد.). البته یک بررسی در [۷] نشان داده است که مراجع دیگر ذکر شده‌ی خطا قبل از اینکه در مسئله اهمیت پیدا کنند از بین می‌روند.
همگرایی قوی ذکر شده در (۱۴) میزان سرعت کاهش میانگین خطا را وقتی $\Delta t \to ۰$، اندازه گیری می‌کند اما ممکن است یک الگوی کم کاربرد دیگر هم گاهی مد نظر باشد و آن عبارتست از اندازه‌گیری سرعت کاهش خطای میانگین‌ها و این منجر به مفهوم همگرایی ضعیف می‌شود.
گوییم روشی دارای مرتبه‌ی ضعیف همگرایی γ است هرگاه ثابت C چنان باشد که برای Δtهای بقدر کافی کوچک و هر تابع g داشته باشیم:

$$|Eg(X_n) - Eg(X(t'))| \leq C\Delta t^{\gamma}, \quad (t' = n\delta t \in [۰, T]) \qquad (۱۸)$$

البته تابع g باید تابعی بطور مناسب هموار بوده و در شرایط رشد چند جمله‌ایها صدق کند. در این مقاله توجه ما به حالتی معطوف خواهد بود که در آن g یک تابع یکّه (همانی) است. در [۱۱] ثابت شده که برای توابع f و g بطور مناسب هموار، روش اولر ـ مارویاما مرتبه‌ی همگرایی ضعیف ۱ دارد. به پیروی از آنچه در مورد همگرایی قوی گفتیم، فرض کنیم:

$$e^W_{\Delta t} = |EX_N - EX(T)|, \quad N\Delta t = T$$

نمایانگر خطای نقطه‌ی انتهایی روش اولر ـ مارویاما (به مفهوم ضعیف همگرایی) باشد. حال اگر $g(x) = x$ و $\gamma = ۱$ باشد، از (۱۸) نتیجه می‌شود که برای Δtهای بقدر کافی کوچک، داریم:

$$e^W_{\Delta t} \leq C\Delta t$$

در برنامه‌ی $emw.m$ ذیل همگرایی مرتبه‌ی ضعیف روش اولر ـ مارویاما را اجرا کرده‌ایم. در این برنامه معادله‌ی (۱۲) را به ازای $\lambda = ۲$ و $\mu = ۰/۱$ و $X_۰ = ۱$ اجرا کرده‌ایم. مطابق مطالبی که قبلاً بیان شد، حدود ۵۰۰۰۰ مسیر براونی گسسته شده را آزموده‌ایم تا نتیجه‌ی لازم حاصل

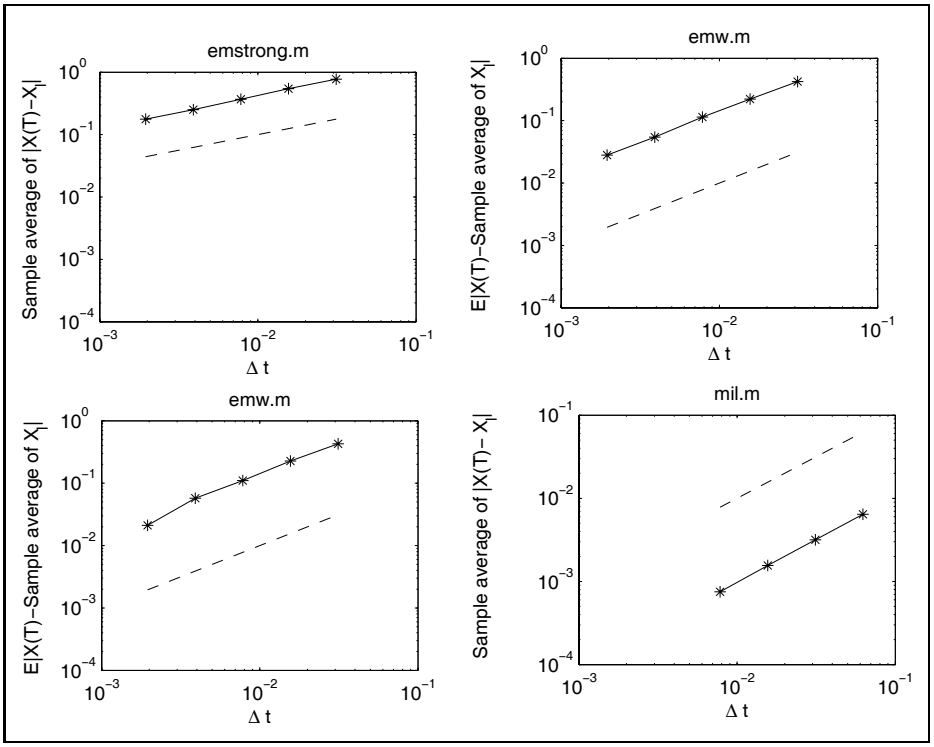
شود. $(1 \le p \le 5)$,$\Delta t = 2^{p-10}$ فرض شده است. این برنامه در مقایسه با $ems.m$ یک سطح اضافی برداری شدن در بر دارد یعنی ما شبیه‌سازی را با تمام ۵۰۰۰۰ مسیر انجام داده‌ایم. این کار زمان اجرای انجام برنامه را بهبود می‌بخشد اما حافظه‌ی زیادتری اشغال می‌کند. برای جبران این اشکال مسیرهای مختلفی را استفاده کرده‌ایم، طوریکه بجای تمامی مسیرها، فقط کافیست نموهای جاری ذخیره شوند. علاوه بر این برای افزایش کارایی برنامه $\Delta t = \delta t$ فرض شده است. تقریبهای میانگین نمونه برای EX_N در Xem و خطای روش در $Xerr$ ذخیره شده‌اند. با توجه به (۱۲) داریم: $E(X(t)) = e^{\lambda t}$ و از این مطلب جواب واقعی را بدست آورده‌ایم. در اینجا، مطابق قانون توانی که قبلاً گفتیم، مقدار $q = 0.9858$ و باقیمانده ۰٫۰۵۰۸ بدست آمده است که موید همگرایی ضعیف مرتبه‌ی ۱ روش است. (شکل ۱ بالای گوشه‌ی سمت راست)

```
randn('state',100)
lambda=input(' drift cofficient:');
mu=input(' diffusion cofficient:');
X0=input(' initial value:');
T=input(' final time:');
M=50000; Xem=zeros(5,1);
for p=1:5
Dt=2^(p-10); L=T/Dt;
Xtemp=X0*ones(M,1);
for j=1:L
Winc=sqrt(Dt)*randn(M,1);
••Winc=sqrt(Dt)*sign(randn(M,1));  Weak E-M method ••
Xtemp=Xtemp+Dt*lambda*Xtemp+mu*Xtemp.*Winc;
end
Xem(p)=mean(Xtemp);
end
Xerr=abs(Xem-exp(lambda));
Dtvals=2.^([1:5]-10);
subplot(222);
loglog(Dtvals,Xerr,'b*-'), hold on
loglog(Dtvals,Dtvals,'r–'), hold off
axis([1e-3 1e-1 1e-4 1])
xlabel('\ Delta t'), ylabel('E|X(T) – Sample average of X_l| ')
title('emw.m','FontSize',10)
A=[ones(p,1), log(Dtvals)']; rhs=log(Xerr);
sol=A\rhs; q=sol(2)
resid=norm(A*sol-rhs)
```

$*** emw.m : Weak\ convergence\ of\ Euler - Maroyama\ method ***$

لازم است تأکید کنیم که ما در $emw.m$ برای هر طول گام Δt، مسیرهای گوناگونی را استفاده کرده‌ایم و این کار هم معقولانه است زیرا همگرایی ضعیف تنها با میانگین جوابها سروکار دارد و ما مختار هستیم که هر نمونه‌ی $W(t'_j) - W(t'_{j-1})$ را برای نمو $\sqrt{\Delta t}N(0,1)$ در (۱۱) در هر

گامی قرار دهیم و در واقع حتی اگر نمو را با متغیر تصادفی دو نقطه‌ای مستقل $\sqrt{\Delta t}V_j$ (V_j مقادیر $+1$ و -1 را با احتمال مساوی می‌پذیرد.) جانشین نماییم، مرتبه‌ی همگرایی ضعیف همچنان ثابت باقی می‌ماند. (توجه کنید که V_j دارای میانگین و واریانسی همانند $N(0,1)\sqrt{\Delta t}$ است.). اگر در این مرحله نمونه‌های وینر را با $\sqrt{\Delta t}V_j$ عوض کنیم، روش منسوب به روش ضعیف اولر ـ مارویاما بدست می‌آید که مرتبه‌ی ضعیف ۱ دارد اما چون از اطلاعات مسیر استفاده نمی‌کند، همگرایی قوی در بر ندارد. نکته‌ی اصلی در استفاده از روش ضعیف اولر ـ مارویاما در این است که اعداد تصادفی که توسط مولد اعداد تصادفی از V_j تولید می‌شوند، نسبت به آنهایی که از $N(0,1)$ تولید می‌شوند، معمولاً کاراتر واقع می‌شوند. ما این موضوع را در برنامه‌ی $emw.m$ در خطی که بین دو • قرار دارد، بیان کرده‌ایم. اگر این خط هم در اجرای برنامه قرارگیرد، مقدار $q = 1/0671$ و باقیمانده برابر $0/2096$ بدست می‌آید.(شکل ۱ پایین سمت چپ)



شکل ۱

۶ روش ملشتاین با مرتبه‌ی بالاتر

در بخش گذشته دیدیم که روش اولر ـ مارویاما مرتبه‌ی قوی همگرایی $\frac{1}{2}$ و ضعیف ۱ دارد. این در حالیست که روش اولر برای معادلات جبری، مرتبه‌ی قوی ۱ دارد. امکان بالا بردن مرتبه‌ی قوی روش اولر ـ مارویاما با اضافه کردن یک تصحیح در نمو آن وجود دارد. این کار منجر به روش منسوب به ملشتاین[1] می‌گردد. امکان این افزایش با توجه به اینکه بسط روش قدیمی تیلور باید با توجه به فرمول ایتو تصحیح شود، وجود دارد. بسط تصادفی ایتو ـ تیلور یاریگر ما در این زمینه خواهد بود. هرگاه این بسط را از نقطه‌ی معینی برش بزنیم و باقی جملات را بعنوان خطا فرض کنیم، روش ملشتاین چنین بدست می‌آید:

$$\begin{aligned} X_j &= X_{j-1} + f(X_{j-1})\Delta t + g(X_{j-1})(W(t'_j) - W(t'_{j-1})) \\ &+ \frac{1}{2}g'(X_{j-1})g(X_{j-1})\{(W(t'_j) - W(t'_{j-1}))^2 - \Delta t\} \end{aligned}$$

در برنامه‌ی $mil.m$ ذیل، ما روش ملشتاین را برای معادله‌ی زیر اجرا کرده‌ایم:

$$dX(t) = rX(t)(k - X(t))dt + \beta X(t)dW(t),\ X(0) = X_0 \qquad (۱۹)$$

این معادله که در آن β, k, r مقادیر ثابتی هستند، یکی از معادلات خاص رشد جمعیتی است [۹].
در اجرای برنامه $X_0 = 0.5, \beta = 0.25, k = 1.5, r = 3$ فرض شده است. مسیر حول $[0, 1]$ با $\delta t = 2^{-11}$ گسسته‌سازی شده است.
جواب معادله‌ی (۱۹) را می‌توان به شکل انتگرالی نوشت. برای راحتی، جواب ملشتاین را با $\Delta t = \delta t$ در نظرگرفته‌ایم تا تقریب خوبی برای جواب واقعی باشد و و آنرا با جواب ملشتاین در ازای $\Delta t = 16\delta t, \Delta t = 32\delta t, \Delta t = 64\delta t, \Delta t = 128\delta t$ با ۵۰۰ مسیر مقایسه کرده‌ایم.
در مقایسه با $ems.m$، این برنامه یک سطح اضافی دارد یعنی بجای استفاده از حلقه‌ی for برای تغییر حول مسیرهای نمونه، آنرا با تمامی مسیرها بطور همزمان محاسبه کرده‌ایم. در این برنامه $q = 1.0316$ و باقیمانده برابر ۰٫۰۰۹۵ شده است. (شکل ۱ پایین سمت راست)

```
randn('state',100)
r=2; k=1; beta=0.25; X0=0.5;
T=1; N=2^(11); dt=T/N;
M=500;
R=[1; 16; 32; 64; 128];
dW=sqrt(dt)*randn(M,N);
Xmil=zeros(M,5);
for p=1:5
Dt=R(p)*dt; L=N/R(p);
Xtemp=X0*ones(M,1);
```

1) Milstein

```
for j=1:L
Winc=sum(dW(:,R(p)*(j-1)+1:R(p)*j),2);
Xtemp=Xtemp+Dt*r*Xtemp.*(k-Xtemp)+beta*Xtemp.*Winc...
+0.5*beta^2*Xtemp.*(Winc.^2-Dt);
end
Xmil(:,p)=Xtemp;
end
Xref=Xmil(:,1);
Xerr=abs(Xmil(:,2:5)-repmat(Xref,1,4));
mean(Xerr);
Dtvals=dt*R(2:5);
subplot(2.3,2.3,4);
loglog(Dtvals,mean(Xerr),'b*-'), hold on
loglog(Dtvals,Dtvals,'r–'), hold off
axis([1e-3 1e-1 1e-4 1])
xlabel('\Delta t'), ylabel('Sample average of |X(T) − X_l| ')
title('mil.m','FontSize',10)
A=[ones(4,1), log(Dtvals)]; rhs=log(mean(Xerr)');
sol=A\rhs; q=sol(2)
resid=norm(A*sol-rhs)
```

$***$ *mil.m : Milstien method with Strong convergence* $***$

۷ قانون زنجیره‌ای تصادفی

در بخش ۳ دیدیم که بیش از یک راه برای تعریف انتگرالها به مفهوم تصادفی وجود دارد. در این بخش به طور خلاصه به یکی دیگر از تفاوتهای محاسبات جبری و تصادفی اشاره می‌نماییم. در حالت جبری، اگر $dX/dt = f(X)$، آنگاه برای هر تابع هموار U بر طبق قانون زنجیره‌ای می‌توان نوشت:

$$\frac{dU(X(t))}{dt} = \frac{dU(X(t))}{d(X(t))}.\frac{d(X(t))}{dt} = \frac{dU(X(t))}{d(X(t))}.f(X(t)) \qquad (۲۰)$$

حالا فرض کنید که $X(t)$ در معادله‌ی ایتوی (۹) صدق می‌کند. در مقایسه با (۲۰)، در مورد $U(X)$ در این وضع چه می‌توان گفت؟ یک حدس معقول و مستدل این است که: $dU = (\frac{dU}{dX})dX$ طوریکه به کمک (۹)؛

$$dU(X(t)) = \frac{dU(X(t))}{d(X)}(f(X(t))dt + g(X(t))dW(t))$$

تجزیه و تحلیلی کمی مشکل، به کمک محاسبات ایتو، وجود یک جمله‌ی اضافی را آشکار ساخته و فرمول درست چنین بدست می‌آید:

$$dU(X(t)) = \frac{dU(X(t))}{dX}dX + \frac{۱}{۲}g(X(t))^۲\frac{d^۲U(X(t))}{dX^۲}dt$$

که با کمک (۹) از آن نتیجه می‌شود:

$$\begin{aligned} dU(X(t)) &= (f(X(t))\frac{dU(X(t))}{dX} + \frac{۱}{۲}g(X(t))^۲\frac{d^۲U(X(t))}{dX^۲})dt \\ &+ g(X(t))\frac{dU(X(t))}{dX}dW(t)) \end{aligned} \quad (۲۱)$$

با توجه به اینکه جزییات کامل این مطالب در [۹] و [۱۱] و [۶] وجود دارد، ما در پی اثبات یا توجیه روابط فوق نیستیم، اما یک آزمایش عددی را با آن شکل می‌دهیم.
معادله‌ی دیفرانسیل تصادفی ذیل را در نظر می‌گیریم:

$$dX(t) = (\alpha - X(t))dt + \beta\sqrt{X(t)}dW(t),\ X(۰) = X_۰ \quad (۲۲)$$

که α, β پارامترهای مثبت و ثابت‌اند. این یک معادله‌ی دیفرانسیل ریشه‌ی مربعی است که ارزش سرمایه‌ی گذاشته شده را مدلسازی می‌کند [۸]. می‌توان نشان داد که اگر $X(۰) \geq ۰$ (با احتمال ۱)، آنگاه این مثبت بودن برای تمام $t > ۰$، باز هم باقیست. با فرض $U(X) = \sqrt{X}$، اجرای روش (۲۱) نتیجه می‌دهد:

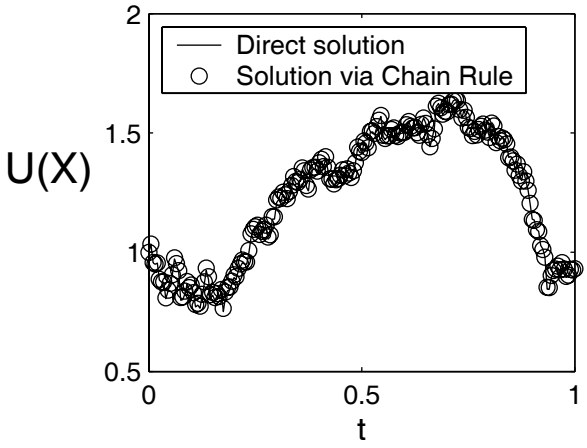
$$dU(t) = (\frac{۴\alpha - \beta^۲}{۸U(t)} - \frac{۱}{۲}U(t))dt + \frac{۱}{۲}\beta W(t) \quad (۲۳)$$

در برنامه‌ی $chain.m$ ذیل، روش اولر ـ مارویاما را برای معادله‌ی (۲۲) با $\alpha = ۲, \beta = ۱, X_۰ = ۱$ اجرا کرده‌ایم. مسیر براونی حول $[۰, ۱]$ با $\delta t = \frac{۱}{۲۰۰}$ و $\Delta t = \delta t$ را اجرا کرده‌ایم. همچنین (۲۳) را برای U حل کرده‌ایم و آنرا هم رسم کرده‌ایم. با توجه به شکل در می‌یابیم که تطابق خوبی بین این دو جواب وجود دارد، زیرا تفاضل بین آنها $Xdiff = ۰/۰۱۵۱$ شده است.

```
 randn('state',100)
alpha=2; beta=1; X0=1; T=1;
N=200; dt=T/N; X01=1/sqrt(X0);
Dt=dt;
Xem1=zeros(1,N); Xem2=zeros(1,N);
Xtemp1=X0; Xtemp2=X01;
for j=1:N
Winc=sqrt(dt)*randn;
f1=(alpha-Xtemp1);
```

```
g1=beta*sqrt(abs(Xtemp1));
Xtemp1=Xtemp1+Dt*f1+Winc*g1;
Xem1(j)=Xtemp1;
f2=(4*alpha-beta^2)/(8*Xtemp2)-Xtemp2/2;
g2=beta/2;
Xtemp2=Xtemp2+Dt*f2+Winc*g2;
Xem2(j)=Xtemp2;
end
subplot(1.2,1.2,1)
plot([0:Dt:T],[sqrt([X0,abs(Xem1)])],'b-',[0:Dt:T],[X0,Xem2],'ro')
legend('Direct solution','Solution via Chain Rule',2)
xlabel('t','FontSize',12)
ylabel('U(X)','FontSize',16,'Rotation',0,'HorizontalAlignment','right')
Xdiff=norm(sqrt(Xem1)-Xem2,'inf')
```

$***$ *chain.m : Test Stochastic Chain Rule* $***$



۸ نتیجه‌گیری

در این مقاله ضمن آنکه ازکمترین جزییات بهره گرفته شده است، سعی شده یک تکنیک کاربردی و عملی برای شبیه‌سازی عددی معادلات دیفرانسیل تصادفی در اختیار خواننده قرارگیرد. بنابراین از بسیاری از روشهای پیشرفته‌تر صرفنظر شده و اثبات مطالب اصلی ارجاع گردید. به عنوان مثال بیان نکردیم که چه شرایطی باید بر f, g اعمال شود تا معادله‌ی دیفرانسیل ایتوی (۹) جواب تضمینی و منحصر بفرد داشته باشد یا در جای دیگر، ما بیشتر توجه خود را به معادلات دیفرانسیل اسکالری معطوف کردیم. روش اولر ـ مارویاما تفاوت چندانی در حالت اسکالری یا غیر اسکالری

ندارد اما روش ملشتاین در حالت غیر اسکالری بسیار پیچیده می‌شود. به هر حال خواننده می‌تواند بدنبال روشهای عددی کاراتر و با مرتبه‌ی همگرایی بالاتر هم باشد.

مراجع

[1] T. C. GARD, Introduction to Stochastic Differential Equations, Marcel Dekker, New York (1988).

[2] D. J. HIGHAM, Mean-Square and asymptotic stability of the stochastic theta method, SIAM J. Numer. Anal., 38 (2000), pp. 753-769.

[3] D. J. HIGHAM AND N. J. HIGHAM, MATLAB Guide, SIAM, Philadelphia, 2000.

[4] J. C. HULL, Options, Futures, and Other Derivatives, 4th ed., Prentice-Hall, Upper Saddle River, NJ, 2000.

[5] I. KARATZAS AND S. E. SHREVE, Brownian Motion and Stochastic Calculus, 2th ed., Springer-Verlag, Berlin, 1991.

[6] P. E. KLOEDEN AND E. PLATEN, Numerical Solution of Stochastic Differentil Equations, Springer-Verlag, Berlin, 1999.

[7] Y. KOMORI, Y. SAITO, AND T. MITSUI, Some issues in discrete approximate solution for stochastic differential equations, Comput. Math. Appl., 28 (1994), pp. 269-276.

[8] X. MAO, Stochastic Differential Equations, Hoewood, Chichester, 1977.

[9] B. ØKSENDAL, Stochastic Differential Equations, 5th ed., Springer-Verlag, Berlin, 1998.

[10] H. C. ÖTTINGER, Stochasti Processes in polymeric Fluids, Springer-Verlag, Berlin, 1996.

[11] E. PLATEN, An introdution to numerical methods for stochastic differential equations, Acta Numer., 8 (1999), pp. 197-246.

[12] Y. SAITO, AND T. MITSUI,Stability analysis of numerical schemes for stochastic differential equations, SIAM J. Numer. Anal., 33(1996), pp. 2254-2267.

[13] THE MATH WORKS, INC, MATLAB User's Guide, Natick, Massachusetts, 1992.

[14] L. N. TREFETHEN, Spectral Methods in MATLAB, SIAM, Philadelphia, 2000.

رگرسیون ژرفا در حالت چند گانه

حمید شریف

بانک مرکزی جمهوری اسلامی ایران

چکیده: در این مقاله به بیان نظریه ژرفا در رگرسیون می‌پردازیم. رگرسیون ژرفا[1] که عدد صحیحی بین ۰ و n است به عنوان خاصیتی از برازش رگرسیونی در نظر گرفته می‌شود که رتبه برازش رگرسیونی تعبیر می‌شود و معیاری برای مقایسه برازشهای مختلف رگرسیونی می‌باشد. در این مقاله با بیان رگرسیون ژرفا در حالت چندگانه به بررسی خواص آن پرداخته و روش ژرفترین رگرسیون را که نسبت به تبدیلات یکنوا بر متغیر پاسخ هم‌ورداست، ارائه می‌دهیم. با معرفی معیار مقدار فرو ریزش نشان می‌دهیم که این روش، روشی استوار می‌باشد. در پایان برآوردکننده ژرفترین رگرسیون را با دیگر برآوردکننده‌های رگرسیونی با محاسبه رگرسیون ژرفای آنها برای داده‌های نمونه‌ای موجود مقایسه می‌کنیم.

واژه‌های کلیدی: ژرفا، نابرازا[2]، ژرفترین رگرسیون، مقدار فرو ریزش، رگرسیون ژرفای ماکسیمال، هم‌وردایی

۱ مقدمه

وقتی رگرسیون کمترین توانهای دوم را با استفاده از n مشاهده برای یک مدل پارامتری $Y = X\beta + \varepsilon$ به کار می‌بریم فرضهایی را در مورد بردار خطاها ε در نظر می‌گیریم. یکی از این فرضها نرمال بودن توزیع خطاهاست. در عمل انحرافاتی از این فرضها رخ می‌دهد. مثلاً ممکن است توزیع زیربنائی خطاها متقارن اما غیر نرمال باشد یا تیزتر از نرمال بوده و دمهای کوتاهتری داشته باشد یا کشیدگی کمتر از نرمال با دمهای پهن‌تر داشته باشد و یا ممکن است توزیع به صورت نرمال باشد ولی دارای دور افتاده‌هایی باشد.

در این موارد به جای روش کمترین توانهای دوم از روشهای رگرسیون استوار استفاده می‌شود که در مقایسه با روش کمترین توانهای دوم نسبت به این انحرافات حساسیت کمتری دارند. یک روش رگرسیونی استوارکه اخیرا توسط روسیف و هوبرت[3] (۱۹۹۹) مطرح شده است روش ژرفترین رگرسیون نام دارد که بر پایه نظریه رگرسیون ژرفا بنا شده است. رگرسیون ژرفا کیفیت برازش را اندازه‌گیری کرده و میزان دوری آن را از هر نابرازا اندازه می‌گیرد و بیان می‌کند که ابرصفحهٔ برازشی در توصیف داده‌ها به چه اندازه خوب عمل می‌کند. بنابراین برازشی با ژرفای بزرگ نسبت به

1) regression depth 2) nonfit 3) Rousseeuw and Hubert

داده‌ها متعادلتر است و لذا یک برازش خوب، ژرفای بزرگتری نسبت به یک برازش بد دارد. در این مقاله با مرور رگرسیون ژرفا در حالت چندگانه به بررسی خواص آن پرداخته و روش ژرفترین رگرسیون را بیان می‌کنیم.

در بخش دوم این مقاله، بطور مختصر رگرسیون ژرفا را در حالت ساده مرور می‌کنیم. در بخش سوم رگرسیون ژرفا را در حالت چندگانه معرفی می‌کنیم. رگرسیون ژرفای ماکسیمال و ژرفترین رگرسیون به ترتیب در بخشهای چهارم و پنجم بیان می‌شود و در بخش ششم خواص ژرفترین رگرسیون در حالت چندگانه بررسی می‌گردد. این خواص شامل هم‌وردایی و استواری ژرفترین رگرسیون است و در انتها با ارائه مثال برآوردکننده ژرفترین رگرسیون را با دیگر برآوردکننده‌های رگرسیونی مقایسه می‌کنیم.

۲ مروری بر رگرسیون ژرفا در حالت ساده

هدف در رگرسیون ساده برازش یک خط راست $y = \theta_1 x + \theta_2$ به یک مجموعه داده $Z_n = \{(x_i, y_i); i = 1, ..., n\} \subseteq R^2$ است. یک برازش را به صورت $\theta = (\theta_1, \theta_2)$ نشان می‌دهیم که مؤلفه اول آن برآورد شیب و مؤلفه دوم آن جمله عرض از مبداء است. مانده‌های مجموعه داده Z_n متناسب با برازش θ را به صورت $r_i(\theta) = r_i = y_i - \theta_1 x_i - \theta_2$ نشان می‌دهیم. برای معرفی ژرفای یک برازش ابتدا یک نابرازا را تعریف می‌کنیم.

تعریف ۱: برازش $\theta = (\theta_1, \theta_2)$ برای مجموعه داده Z_n نابرازا نامیده می‌شود اگر و فقط اگر یک عدد حقیقی $v_\theta = v$ مخالف همهٔ x_i ها وجود داشته باشد به طوری که

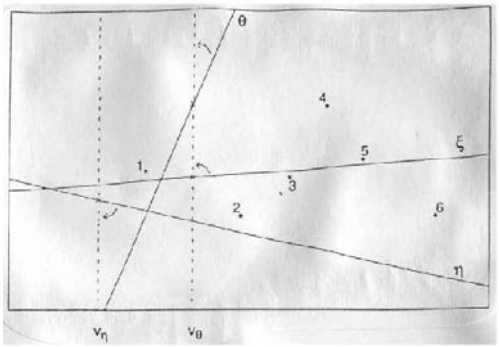
$$\forall\, x_i > v \;,\quad r_i(\theta) > 0 \qquad , \qquad \forall\, x_i < v \;,\quad r_i(\theta) < 0$$

یا

$$\forall\, x_i > v \;,\quad r_i(\theta) < 0 \qquad , \qquad \forall\, x_i < v \;,\quad r_i(\theta) > 0$$

خطی که بالا یا پایین همهٔ مشاهدات واقع می‌شود همیشه یک نابرازا است. بعداً خواهیم دید که نابرازاها در واقع برازش‌هایی با ژرفای صفر هستند.

شکل ۱ یک مجموعه داده را با ۶ مشاهده و θ و η نابرازا نشان می‌دهد. مقادیر مربوطه v_θ و v_η نشان داده شده‌اند. با توجه به شکل، وجود v متناظر با وجود یک ضربدر روی خط عمودی که از نقطه v رسم شده، است. چون $\forall\, x_i < v_\theta$، $r_i(\theta) > 0$ و همچنین $\forall\, x_i > v_\theta$، $r_i(\theta) < 0$ لذا با توجه به تعریف فوق θ، نابرازا است. همچنین چون خط η در پایین همهٔ مشاهدات قرار دارد این خط نیز نابرازا است. از آن جا که تعریف فوق برای خط ξ صادق نیست این خط نابرازا نیست. همان طور که ذکر شد با مشخص شدن v، نقطه متناظر با آن که محل تلاقی خط عمودی ترسیم شده از نقطه v با خط رگرسیونی است و با ضربدر نشان داده شده است، مشخص می‌گردد. به عبارت دیگر محل تلاقی خط برازش داده شده با خط عمودی که از نقطه v متناظر با آن خط ترسیم می‌گردد با ضربدر مشخص شده است. از این



شکل ۱: مجموعه داده دو متغیره با نابرازاهای θ و η و یک برازش ξ با رگرسیون ژرفای ۲

رو می‌توان خطوط نابرازا را از روی شکل ۱ مشخص کرد. به این ترتیب که خط برازش داده شده را حول نقطه‌ای که با ضربدر بر روی آن خط مشخص شده است می‌گردانیم تا عمودی شود. اگر در حین چرخاندن، خط از هیچ مشاهده‌ای عبور نکند نابرازا است. بنابراین خطوط θ و η، نابرازا هستند ولی خط ξ، نابرازا نیست چون وقتی آن را حول نقطه مذکور (مشخص شده با ضربدر) می‌گردانیم تا عمودی شود از مشاهدات ۴ و ۵ عبور می‌کند.
به طورکلی ژرفای یک برازش θ برای یک مجموعه داده Z_n به حجم n به صورت زیر بیان می‌شود:

تعریف ۲: رگرسیون ژرفای $(rdepth)$ یک برازش θ متناسب با مجموعه داده Z_n عبارت است از کمترین تعداد مشاهداتی که باید برداشته شود تا θ نابرازا شود. به طور معادل، $rdepth(\theta, Z_n)$ عبارت است از کوچکترین تعداد مانده‌هایی که می‌بایست علامتشان تغییر کند تا θ نابرازا شود.

برای مثال خط ξ را در شکل ۱ در نظر می‌گیریم. این خط با حذف کردن مشاهدات ۴ و ۵ نابرازا می‌شود (زیرا با قرار دادن v_ξ مساوی با v_θ و حذف مشاهدات ۴ و ۵ می‌توان خط ξ را به صورت عمودی درآورد بدون آن که از مشاهده‌ای عبور کند). چون خط ξ با حذف حداقل ۲ مشاهده نابرازا می‌شود لذا $rdepth(\xi, Z_n) = ۲$.

توجه ۱: تعاریف ۱ و ۲ در مواقعی که در x_i ها تکرار وجود داشته باشد (x_i ها با هم مساوی باشند) نیز برقرار است. و x_i ها به هیچ فرض توزیعی نیاز ندارند.

مرتبه زمانی: گوییم $f(n)$ از مرتبه زمانی $g(n)$ است و آن را با نماد $f(n) = O(g(n))$ نشان می‌دهیم اگر و تنها اگر اعداد صحیح و مثبتی مانند $n_۰$ و c وجود داشته باشد بطوریکه $f(n) \leq c\, g(n)$ برای تمام nهای $n \geq n_۰$ برقرار باشد.

برای محاسبه $rdepth(\theta, Z_n)$ ابتدا مشاهدات را به صورت $x_۱ \leq x_۲ \leq \cdots \leq x_n$ در

مرتبه زمانی $O(nlogn)$ مرتب می‌کنیم. تعریف می‌کنیم:

$$L^+(v) = \#\{j; x_j \le v, r_j \ge ۰\}$$

و

$$R^-(v) = \#\{j; x_j > v, r_j \le ۰\}$$

L^- و R^+ به صورت مشابه تعریف می‌شوند. سپس $L^+(x_i)$، $L^-(x_i)$، $R^-(x_i)$ و $R^+(x_i)$ برای هر $i = ۱, \cdots, n$ مشابه تعاریف فوق، محاسبه می‌شوند. $rdepth(\theta, Z_n)$ در $O(n)$ عمل به صورت زیر محاسبه می‌شود.

$$rdepth(\theta, Z_n) = \min_{۱\le i\le n} (\min\{L^+(x_i)+R^-(x_i)\ ,\ R^+(x_i)+L^-(x_i)\}) \quad (۱)$$

مثال زیر نحوه محاسبه رگرسیون ژرفای یک خط را با استفاده از رابطه (۱) نشان می‌دهد.

مثال ۱: شکل ۲، ۷ مشاهده با خط ξ را نشان می‌دهد همانطور که دیده می‌شود این خط نابرازا نیست. با استفاده از رابطه (۱) رگرسیون ژرفای این خط به صورت زیر محاسبه می‌شود.

$$i = ۱ : \min\{L^+(x_۱) + R^-(x_۱), R^+(x_۱) + L^-(x_۱)\} = \min\{۱ + ۳, ۳ + ۱\} = ۴$$
$$i = ۲ : \min\{L^+(x_۲) + R^-(x_۲), R^+(x_۲) + L^-(x_۲)\} = \min\{۱ + ۲, ۳ + ۲\} = ۵$$
$$i = ۳ : \min\{L^+(x_۳) + R^-(x_۳), R^+(x_۳) + L^-(x_۳)\} = \min\{۲ + ۲, ۲ + ۲\} = ۴$$
$$i = ۴ : \min\{L^+(x_۴) + R^-(x_۴), R^+(x_۴) + L^-(x_۴)\} = \min\{۲ + ۱, ۲ + ۳\} = ۳$$
$$i = ۵ : \min\{L^+(x_۵) + R^-(x_۵), R^+(x_۵) + L^-(x_۵)\} = \min\{۳ + ۱, ۱ + ۳\} = ۴$$
$$i = ۶ : \min\{L^+(x_۶) + R^-(x_۶), R^+(x_۶) + L^-(x_۶)\} = \min\{۳ + ۰, ۱ + ۴\} = ۳$$
$$i = ۷ : \min\{L^+(x_۷) + R^-(x_۷), R^+(x_۷) + L^-(x_۷)\} = \min\{۴ + ۰, ۰ + ۴\} = ۴$$

بنابراین

$$rdepth(\xi, Z_n) = \min(۴, ۵, ۴, ۳, ۴, ۳, ۴) = ۳$$

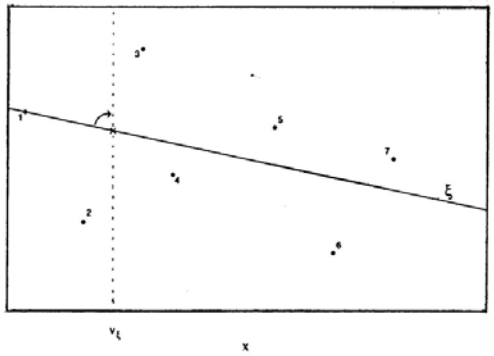
از روی شکل ۲ نیز دیده می‌شود که با برداشتن مشاهدات ۱ و ۴ و ۶ می‌توان خط ξ را بدون این که از مشاهدات دیگر عبور کند حول نقطه متناظر با v_ξ که با ضربدر روی آن خط مشخص شده است دوران داد تا به صورت عمودی در آید. به عبارت دیگر با برداشتن مشاهدات ۱ و ۴ و ۶ خط ξ، نابرازا می شود. رگرسیون ژرفای خط ξ برابر با کران پایین رگرسیون ژرفای ماکسیمال یعنی $۳ = \lceil \frac{۷}{۳} \rceil = \lceil \frac{n}{۳} \rceil$ است که در بخش‌های بعد بیان می‌شود.

توجه ۲: $\lceil \lambda \rceil$ کوچکترین عدد صحیح بزرگتر یا مساوی λ است.

۳ رگرسیون ژرفا در حالت چندگانه

در رگرسیون چندگانه مجموعه داده Z_n به صورت

$$Z_n = \{(x_{i۱}, \ldots, x_{i,p-۱}, y_i); i = ۱, \ldots, n\} \subset \mathbb{R}^p$$



شکل ۲: مجموعه داده دو متغیره و برازش ξ با رگرسیون ژرفای ۳.

است. X را به عنوان قسمتی از مختصات هر نقطه به صورت
$$X_i = (x_{i1}, \ldots, x_{i,p-1}) \in \mathbb{R}^{p-1}$$
در نظر می‌گیریم. اکنون می‌خواهیم y_i را به وسیله

$$\theta_1 x_{i1} + \cdots + \theta_{p-1} x_{i,p-1} + \theta_p = (X_i, 1)\theta'$$

یعنی به وسیله یک ابرصفحه آفین در $\mathbb{R}^p$، که $\theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$ است، برازش دهیم. در اینجا هیچ فرض توزیعی وجود ندارد.

قبل از بیان تعاریف و مطالب این بخش ابتدا ابرصفحه و ابرصفحه آفین را تعریف می‌کنیم.

تعریف ۳: فرض کنید $[f = \alpha]$ نشان دهنده مجموعه سطح $\{x; f(x) = \alpha\}$ باشد. یک ابرصفحه مجموعه‌ای به شکل $[f = \alpha]$ است که f تابعک خطی غیرصفر روی فضای برداری X و α عددی حقیقی است. به عبارت دیگر مجموعه $H \subseteq \mathbb{R}^n$ یک ابرصفحه است اگر اعداد حقیقی $a_1, \ldots, a_n$ و C (با $a_i \neq 0$ برای حداقل یک i) وجود داشته باشند به طوری که H متشکل از همهٔ نقاط $x = (x_1, \ldots, x_n)$ ایی باشدکه در رابطه $\sum_i \alpha_i x_i = C$ صدق می‌کنند.

تعریف ۴: یک ابرصفحه آفین در فضای برداری X مجموعه‌ای مانند $\mathcal{M}$ است به طوری که برای هر x_0 در $\mathcal{M}$، $\mathcal{M} - x_0$ یک ابرصفحه باشد. به عبارت دیگر $\mathcal{M} \subseteq X$ یک ابرصفحه آفین است اگر یک تابع خطی غیرصفر $f : X \longrightarrow \mathbb{F}$ و یک α در $\mathbb{F}$ وجود داشته باشد به طوری‌که $\mathcal{M} = \{x \in X\,;\, f(x) = \alpha\}$ ($\mathbb{F}$ میدان حقیقی $\mathbb{R}$، یا میدان مختلط $\mathbb{C}$ است).

مثال ۲: فضای برداری $X = \mathbb{R}^2$ و تابع $f(x, y) = y$ را در نظر می‌گیریم. در این صورت مجموعه $[f = 0] = \{(x, y); f(x, y) = 0\} = \{(x, 0); x \in \mathbb{R}\}$ ک ابرصفحه است که

فضای $\mathbb{R}^۲$ را به دو نیم‌فضای بسته $\{(x,y); f(x,y) \geq ۰\}$ و $\{(x,y); f(x,y) \leq ۰\}$ افراز می‌کند. مجموعه

$$C = \{(x,y); f(x,y) = ۱\} = \{(x, ۱); x \in \mathbb{R}\}$$

یک ابرصفحه آفین است زیرا بازای هر $(x_۰, ۱) \in C$

$$C - (x_۰, ۱) = \{(x - x_۰, ۰); x \in R\} = \{(x - x_۰, ۰); f(x - x_۰, ۰) = ۰\}$$

یک ابر صفحه است.

تعریف ۵: بردار $\theta = (\theta_۱, \ldots, \theta_p)$ یک نابرازا نامیده می‌شود اگر وفقط اگر یک ابرصفحه آفین V در فضای x وجود داشته باشد به طوری که هیچ یک از x_i ها متعلق به V نباشند و برای همهٔ x_i ها در یکی از نیم‌فضاهای باز آن داشته باشیم $r_i(\theta) > ۰$ و همچنین برای همهٔ x_i ها در نیم‌فضای باز دیگر $r_i(\theta) < ۰$ باشد.

شکل ۳ مثالی از یک صفحه نابرازاست. در این مثال $p = ۳$ است، بنابراین θ به یک صفحه متعلق است. فضای X به صورت صفحه افقی با $y \equiv ۰$ که شامل خط V است، دیده می شود. (با این تعریف، هر η به طوری که همهٔ مانده‌ها مثبت یا همهٔ مانده‌ها منفی باشند یک نابرازاست، زیرا کافی است V به صورتی انتخاب شود که همهٔ x_i ها در یک طرف آن واقع شوند).

رگرسیون ژرفای هر $\theta \in \mathbb{R}^p$ متناسب با $Z_n \subset \mathbb{R}^p$ همانند تعریف ۲ به صورت زیر تعریف می‌شود:

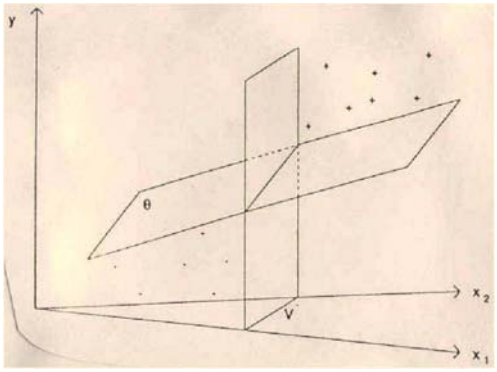
تعریف ۶: رگرسیون ژرفای یک برازش $\theta \in \mathbb{R}^p$ متناسب با مجموعه داده $Z_n \subset \mathbb{R}^p$ عبارت است از کوچکترین تعداد مشاهداتی که وقتی از مجموعه داده‌ها خارج شود θ، نابرازا گردد. به عبارت دیگر $rdepth(\theta, Z_n)$ عبارت است از کمترین تعداد مانده‌هایی که باید تغییر علامت داده شوند تا θ نابرازا شود.

$rdepth(\theta, Z_n)$ همانند رابطه (۱) بیان می‌شود. بدین صورت که هر ابرصفحه آفین V در فضای x مشاهدات را به دو مجموعه تقسیم می‌کند که آن‌ها را با $L(V)$ و $R(V)$ نشان می‌دهیم. مجموعه‌های $L^+(V)$ و $L^-(V)$ به ترتیب تعداد مشاهدات در $L(V)$ با مانده‌های مثبت و تعداد مشاهدات در $L(V)$ با مانده‌های منفی است. $R(V)$ نیز بطور مشابه به دو مجموعه $R^+(V)$ و $R^-(V)$ تقسیم می‌گردد. بنابراین

$$rdepth(\theta, Z_n) = \min_V(\min\{L^+(V) + R^-(V)\ ,\ R^+(V) + L^-(V)\})$$

که V شامل همهٔ ابرصفحه‌های آفین در فضای x است. برای محاسبه رگرسیون ژرفای یک برازش، جستجو را به مجموعه‌ای متناهی از ابرصفحه‌های V محدود می‌کنیم. الگوریتم‌های مناسب برای محاسبه رگرسیون ژرفا در فصل بعد ارائه می‌شود.

لازم به ذکر است که حالت تعمیم یافته قضیه ۱.۲ یعنی در حالت چندگانه (p بعدی) زمانی که $Z_n = \{(X_i, y_i); i = ۱, \ldots, n\}$ که $X_i = (x_{i۱}, \ldots, x_{i,p-۱}) \in \mathbb{R}^{p-۱}$ و



شکل ۳: مثالی از نابرازا $\theta \in \mathbb{R}^3$. ابرصفحه آفین V در فضای X ($\mathbb{R}^2$)، مشاهدات با مانده‌های مثبت را از مشاهدات با مانده‌های منفی جدا می‌کند.

$\theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$ باشد نیز برقرار است.

قضیه ۱: خاصیت دقیق ـ برازش. اگر تعداد مشاهداتی که روی θ قرار گیرند k باشد (که $0 \le k \le n$) آنگاه

$$k \le rdepth(\theta\ ,\ Z_n) \le \left[\frac{n+k}{2}\right]$$

برای $k = n$ داریم $rdepth(\theta, Z_n) = n$.

□

توضیح ۱: از آن‌جاکه برآوردکننده L^1 همیشه از حداقل p مشاهده می‌گذرد، لذا رگرسیون ژرفای آن حداقل p است (بلومفیلد و استیگر[4]، ۱۹۸۳، را ببینید). همچنین این خاصیت برای برآوردکننده کمترین توان‌های دوم پیراسته (LTS) که به وسیله روسیف مطرح شده، برقرار است (روسیف ۱۹۸۴).

توضیح ۲: برآورد کننده کمترین توان‌های دوم (LS) هرگز نابرازا نیست. در حقیقت اگر Z_n به صورتی باشد که $X_n = \{x_1, x_2, ..., x_n\}$ رتبه کامل داشته باشد و $n \ge 2p$، سپس $rdepth(\theta_{LS}, Z_n) \ge 1$. □

۴ رگرسیون ژرفای ماکسیمال

کران‌های بالا برای رگرسیون ژرفای ماکسیمال در قضیه زیر ارائه می‌شود. زیر مجموعه‌ای از $\mathbb{R}^p$ در موقعیت عمومی قرار دارد هرگاه بیشتر از p مشاهده در هر زیرفضای آفین $(p-1)$ بعدی قرار نگیرد.

4) Bloomfield and Steiger

قضیه ۲: اگر (x_i, y_i) در موقعیت عمومی باشند، آنگاه

$$\max_{\theta} rdepth(\theta, Z_n) \leq \left\lceil \frac{n+p}{۲} \right\rceil$$

□

حدس ۱: برای هر مجموعه داده $Z_n \subset \mathbb{R}^p$ داریم

$$\max_{\theta} rdepth(\theta, Z_n) \geq \left\lceil \frac{n}{p+۱} \right\rceil$$

۵ ژرفترین رگرسیون

ژرفترین رگرسیون T_r^* بعنوان θایی که $rdepth(\theta, Z_n)$ را بیشینه می‌کند، تعریف شده است. ژرفترین رگرسیون T_r^* "متعادل‌ترین" برازش است. T_r^* را می‌توان به وسیله محاسبه رگرسیون ژرفای همهٔ برازش‌هایی که از میان p مشاهده می‌گذرد، به دست آورد.

تعریف ۷: برآوردکننده ژرفترین رگرسیون $T_r^*(Z_n)$ عبارت است از برازش θ با بزرگترین رگرسیون ژرفا، یعنی

$$T_r^*(Z_n) = arg_{\theta} \max rdepth(\theta, Z_n)$$

توجه ۳: در رگرسیون چندگانه، ژرفترین رگرسیون برخلاف $L^۱$ و LTS برای تبدیلات یکنوای y، هم‌وردا است.

مثال ۳: می‌خواهیم رابطه بین میزان بخار مصرفی ماهانه توسط یک ماشین بخار (Y) با متوسط دمای ماهانه به فارنهایت $(X_۱)$ و تعداد روزهای کار در ماه $(X_۲)$ را مورد بررسی قرار می‌دهیم. بدین منظور ۲۵ مشاهده جمع‌آوری شده است که در جدول ۱ نشان داده شده‌اند (به دراپر و اسمیت[5] ،۱۹۹۸، مراجعه شود).

با استفاده از روش‌های رگرسیونی مختلف، ضرایب رگرسیونی را برآورد کرده و رگرسیون ژرفای هر یک از این برازش‌ها را به دست آورده‌ایم. نتایج در جدول ۲ آمده است. دیده می‌شود که برآوردکننده ژرفترین رگرسیون، رگرسیون ژرفای بالاتری نسبت به سایر برآوردکننده‌ها دارد و درنتیجه متعادل‌ترین برازش است.

۶ خواص بهینگی ژرفترین رگرسیون

در این بخش بعضی از خواص هم‌وردایی ژرفترین رگرسیون را بیان می‌کنیم. همچنین با معرفی کردن مقدار فروریزش به‌عنوان معیار پایداری برآوردکننده در برابر نقاط پرت، نشان می‌دهیم که روش ژرفترین رگرسیون با توجه به این معیار روشی استوار است.

5) Draper and Smith

جدول ۱: داده‌های ماشین بخار

Y	$X_۲$	$X_۱$	Y	$X_۲$	$X_۱$
۹٫۵۷	۱۹	۳۹٫۱	۱۰٫۹۸	۲۰	۳۵٫۳
۱۰٫۹۴	۲۳	۴۶٫۸	۱۱٫۱۳	۲۰	۲۹٫۷
۹٫۵۸	۲۰	۴۸٫۵	۱۲٫۵۱	۲۳	۳۰٫۸
۱۰٫۰۹	۲۲	۵۹٫۳	۸٫۴۰	۲۰	۵۸٫۸
۸٫۱۱	۲۲	۷۰٫۰	۹٫۲۷	۲۱	۶۱٫۴
۶٫۸۳	۱۱	۷۰٫۰	۸٫۷۳	۲۲	۷۱٫۳
۸٫۸۸	۲۳	۷۴٫۵	۶٫۳۶	۱۱	۷۴٫۴
۷٫۶۸	۲۰	۷۲٫۱	۸٫۵۰	۲۳	۷۶٫۷
۸٫۴۷	۲۱	۵۸٫۱	۷٫۸۲	۲۱	۷۰٫۷
۸٫۸۶	۲۰	۴۴٫۱	۹٫۱۴	۲۰	۵۷٫۵
۱۰٫۳۶	۲۰	۳۳٫۴	۸٫۲۴	۲۰	۴۶٫۴
۱۱٫۸۸	۲۲	۲۸٫۶	۱۲٫۱۹	۲۱	۲۸٫۹
			۱۱٫۸۸	۲۱	۲۸٫۱

جدول ۲: مقایسه رگرسیون ژرفای برآوردکننده‌های رگرسیونی با برآوردکنندهٔ ژرفترین رگرسیون

روش	LS	LAD	برآورد هوبر	برآوردکننده ژرفترین رگرسیون
$\hat{\beta}_۰$	۹٫۲۱۳۰	۹٫۷۳۷۰	۹٫۳۸۲۰	۴٫۰۹۰۰
$\hat{\beta}_۱$	-۰٫۰۷۴۸	-۰٫۰۷۳۸	-۰٫۰۷۶۱	-۰٫۰۸۲۴
$\hat{\beta}_۲$	۰٫۲۰۶۰	۰٫۱۹۲۰	۰٫۲۰۴۰	۰٫۴۷۶۰
رگرسیون ژرفا	۵	۵	۶	۱۱

۱.۶ هم‌وردایی ژرفترین رگرسیون

بررسی خواص فوق در قضیه زیر آمده است:

قضیه ۳.۴: مجموعه داده $Z_n = \{(X_i, Y_i); i = ۱, \ldots, n\}$ که در آن

$$X_i = (x_{i,۱}, \ldots, x_{i,p-۱}) \in \mathbb{R}^{p-۱}$$

و برآوردکننده ژرفترین رگرسیون $T_r^*(Z_n) = arg_\theta \max\ rdepth(\theta, Z_n)$ را در نظر می‌گیریم:

(a) برآوردکننده ژرفترین رگرسیون، رگرسیون هم‌وردا است، یعنی برای هربردار ستونی V

$$T_r^*(Z_n') = T_r^*(Z_n) + V$$

که $Z_n' = \{(X_i, y_i + X_i V);\ i = ۱, \ldots, n\}$ و $T_r^*(Z_n') = arg_\theta \max\ rdepth(\theta, Z_n')$

(b) برآوردکننده ژرفترین رگرسیون، مقیاس هم‌وردا است، یعنی برای هر c ثابت

$$T_r^*(Z_n') = c\ T_r^*(Z_n)$$

(c) $T_r^*(Z_n') = arg_\theta \max\ rdepth(\theta, Z_n')$ و $Z_n' = \{(X_i, cy_i);\ i = ۱, \ldots, n\}$ که

برآوردکننده ژرفترین رگرسیون، آفین هم‌وردا است، یعنی برای هر ماتریس مربع ناویژه A

$$T_r^*(Z_n') = A^{-۱}\ T_r^*(Z_n)$$

که $Z_n' = \{(X_i A, y_i);\ i = ۱, \ldots, n\}$ و $T_r^*(Z_n') = arg_\theta \max\ rdepth(\theta, Z_n')$.

۲.۶ استواری ژرفترین رگرسیون

یک معیار معروف برای پایداری برآورد کننده در برابر نقاط پرت، مقدار فروریزش است. بنا به تعریف داناهو و هوبر[6] (۱۹۸۳)، مقدار فروریزش هر برآوردکننده T_n به صورت زیر است:

$$\varepsilon_n^*(T_n, Z_n) = min\left\{\frac{k}{n};\ \sup_{Z_n'} ||T_n(Z_n') - T_n(Z_n)|| = \infty\right\}$$

که Z_n' شامل همهٔ مجموعه داده‌های بدست آمده با جایگزینی هر k مشاهده دلخواه Z_n با مقادیر اختیاری است. بنابراین مقدار فروریزش، کوچکترین کسری از مشاهدات تبدیل شده (مغشوش شده)ای است که می‌تواند برآوردکننده را به دلخواه منحرف کند. لازم به ذکر است که

6) Donoho and Huber

این اغتشاش فقط به نقاط پرت در y_i محدود نمی‌شود، بلکه Z'_n می‌تواند شامل نقاط پرت در x_i ها نیز باشد.

حدس ۲: برای هر مجموعه داده $Z_n \subset \mathbb{R}^p$ با x_i ها در موقعیت عمومی باشند آنگاه

$$\varepsilon_n^*(T_r^*, Z_n) \geq \frac{1}{n}\left(\left\lceil \frac{n}{p+1} \right\rceil - p + 1\right) \approx \frac{1}{p+1} \qquad (۲)$$

رابطه (۲) بیان می‌کند که مقدار فروریزش ژرفترین رگرسیون همواره مثبت است. این مقدار، زمانی برابر با کران پایین $\frac{1}{p+1}$ می‌شود که داده‌های اصلی خود، ویژه باشند مثلاً وقتی آن‌ها روی منحنی گشتاور قرار گیرند.

بنابراین ژرفترین رگرسیون در برابر نقاط نافذ همانند نقاط دورافتاده عمودی، استوار است. بعلاوه ژرفترین رگرسیون متفاوت از رگرسیون L^1 تعریف شده به صورت $L^1(Z_n) = arg\min_\theta \sum_{i=1}^{n} |r_i(\theta)|$ است. لازم به ذکر است که مقدار فروریزش L^1 صفر است که در نتیجه آسیب‌پذیری آن در برابر نقاط نافذ می‌باشد.

۷ مثال‌ها

در این بخش برآوردکننده ژرفترین رگرسیون را برای داده‌های موجود بدست آورده و نشان می‌دهیم که در مقایسه با برآوردکننده‌های دیگر دارای بیشترین رگرسیون ژرفاست. این مقایسه با محاسبه رگرسیون ژرفای برآوردکننده‌ها انجام می‌گیرد. بدین منظور از الگوریتمهایی که برنامه مربوط به آنها با استفاده از نرم‌افزار Splus و زبان برنامه‌نویسی فرترن نوشته شده است و از وب سایت http://win-www.uia.ac.be/u/statis/ قابل دسترسی است، استفاده می‌گردد.

۱.۷ مثال ۱

در یک تحقیق در سال ۱۹۷۶ میزان آلودگی ۲۰ رودخانه در ایالت نیویورک بررسی شده است. هدف، بررسی سهم هر نوع از زمین‌های مورد استفاده در اطراف رودخانه بر روی آلودگی آب رودخانه‌هاست. این آلودگی براساس میانگین غلظت نیتروژن (میلی‌گرم) اندازه‌گیری می‌شود. نوع زمین‌های مورد استفاده در اطراف رودخانه‌ها و میزان آلودگی آن‌ها در زیر بیان شده است. داده‌ها در جدول ۳ نشان داده شده‌اند (به چاترجی و همکاران[7]، ۲۰۰۰ مراجعه شود).

Y ــ میانگین غلظت نیتروژن (میلی‌گرم بر لیتر) براساس نمونه‌هایی که در فواصل منظم بهار، تابستان و ماههای پاییز گرفته شده است.

7) Chatterjee and et al.

جدول ۳: داده‌های رودخانه‌های نیویورک

Y	$X_۴$	$X_۳$	$X_۲$	$X_۱$	Y	$X_۴$	$X_۳$	$X_۲$	$X_۱$
۱٫۰۱	۰٫۰۹	۰٫۷	۵۷	۲۹	۱٫۱۵	۰٫۲۹	۱٫۲	۶۳	۲۶
۱٫۰۰	۱٫۹۸	۱٫۹	۸۴	۲	۱٫۹۰	۰٫۵۸	۱٫۸	۲۶	۵۴
۱٫۴۲	۰٫۵۶	۳٫۴	۶۱	۱۹	۱٫۹۹	۳٫۱۱	۲۹٫۴	۲۷	۳
۱٫۶۵	۰٫۲۴	۱٫۳	۴۳	۴۰	۲٫۰۴	۱٫۱۱	۵٫۶	۶۰	۱۶
۱٫۲۱	۰٫۲۳	۰٫۹	۶۰	۲۶	۱٫۰۱	۰٫۱۵	۱٫۱	۶۲	۲۸
۰٫۷۵	۰٫۱۶	۰٫۷	۷۵	۱۵	۱٫۳۳	۰٫۱۸	۰٫۹	۵۳	۲۶
۰٫۸۰	۰٫۳۵	۰٫۸	۸۱	۳	۰٫۷۳	۰٫۱۲	۰٫۵	۸۴	۶
۰٫۸۷	۰٫۱۵	۰٫۵	۸۲	۶	۰٫۷۶	۰٫۳۵	۰٫۷	۸۹	۲
۰٫۸۷	۰٫۱۸	۰٫۴	۷۵	۴	۰٫۸۰	۰٫۲۲	۰٫۹	۷۰	۲۲
۱٫۲۵	۰٫۱۳	۱٫۱	۴۹	۴۰	۰٫۶۶	۰٫۱۳	۰٫۵	۵۶	۲۱

جدول ۴: برآورد پارامترهای مدل و رگرسیون ژرفای برآوردکننده‌های مختلف رگرسیونی

روش	LS	LAD	برآورد هوبر	برآوردکننده ژرفترین رگرسیون
$\hat{\beta}_۰$	۱٫۷۲۲۲	۲٫۷۴۰۰	۲٫۶۵۶۰	۲٫۹۹۰۰
$\hat{\beta}_۱$	۰٫۰۰۵۸	-۰٫۰۰۵۲	-۰٫۰۰۵۳	-۰٫۰۱۰۶
$\hat{\beta}_۲$	-۰٫۰۱۳۰	-۰٫۰۲۴۳	-۰٫۰۲۳۲	-۰٫۰۲۸۹
$\hat{\beta}_۳$	-۰٫۰۰۷۲	-۰٫۰۲۱۶	-۰٫۰۲۲۳	۰٫۰۴۵۶
$\hat{\beta}_۴$	۰٫۳۰۵۰	۰٫۱۷۸۹	۰٫۲۱۶۰	۰٫۴۴۰۰
رگرسیون ژرفا	۳	۴	۴	۹
$\sum e_i^۲$	۱٫۰۵۲۷	۱٫۱۸۰۰	۱٫۱۲۷۰	۹٫۱۷۶۰
$\sum \lvert e_i \rvert$	۳٫۲۱۴۶	۲٫۸۱۸۰	۲٫۹۲۴۰	۵٫۳۵۷۷

$X_۱$ ــ کشاورزی: درصد مساحت زمین مورد استفاده برای کشاورزی

$X_۲$ ــ جنگلی: درصد زمین جنگلی

$X_۳$ ــ مسکونی: درصد مساحت زمین مورد استفاده برای سکونت

$X_۴$ ــ تجاری/ صنعتی: درصد مساحت زمین مورد استفاده تجاری یا صنعتی.

با بررسی داده‌های فوق دیده می‌شود که مشاهدات ۴ و ۵ دور افتاده و نافذ و در عین حال مؤثر هستند. بنابراین انتظار می‌رود که روش‌های LS، LAD و برآورد هوبر تحت تأثیر این نقاط قرار گیرند و بطور قابل ملاحظه رگرسیون ژرفای کوچکتری نسبت به برآوردکننده ژرفترین رگرسیون داشته باشند. با نگاه کردن به جدول ۴ مشخص می‌شود که برآوردکننده رگرسیونی LAD و برآورد هوبر نزدیک به هم هستند و دارای رگرسیون ژرفای ۴ هستند. چون برآوردکننده ژرفترین رگرسیون نسبت به نقاط دورافتاده و نافذ استوار است لذا تحت تأثیر این نقاط قرار نمی‌گیرد و به نظر می‌رسد که برازشی مناسب برای داده‌هاست. با توجه به اینکه دارای بیشترین رگرسیون ژرفا یعنی ۹ است.

جدول ۵: جدول ۵: برآورد پارامترهای مدل و رگرسیون ژرفای برآوردکننده‌های مختلف رگرسیونی

روش	LS	LAD	برآورد هوبر	برآوردکننده ژرفترین رگرسیون
$\hat{\beta}_۰$	-۹٫۶۹۵۵	-۹٫۴۹۷۴	-۹٫۶۶۳۳	-۹٫۹۵۰
$\hat{\beta}_۱$	۱٫۱۲۰۸	۱٫۱۱۴۵	۱٫۱۳۷۲	۱٫۱۳۰
$\hat{\beta}_۲$	۱٫۹۶۴۳	۱٫۸۸۵۹	۱٫۹۹۸۲	۲٫۰۱۰
$\hat{\beta}_۳$	-۴۹٫۹۱۱۲	-۴۹٫۸۵۱۳	-۴۹٫۸۸۵۹	-۵۰٫۱۴۰
$\hat{\beta}_۴$	۰٫۵۲۸۸	۰٫۴۷۶۲	۰٫۴۸۷۲	۰٫۴۹۶
رگرسیون ژرفا	۱۳	۱۰	۱۴	۱۸
$\sum e_i^۲$	۳۶٫۴۷۲۰	۳۸٫۶۳۱۰	۳۶٫۶۹۸۰	۳۹٫۹۴۱
$\sum \lvert e_i \rvert$	۳۱٫۳۲۷۰	۲۹٫۸۸۹۰	۳۰٫۸۹۱۰	۳۱٫۷۱۸

۲.۷ مثال ۲ـ شبیه‌سازی

برای بررسی مناسب بودن برآوردکننده ژرفترین رگرسیون، ۴۰ نقطه در ۵ بعد مطابق با مدل $y = -۱۰ + x_۱ + ۲x_۲ - ۵۰x_۳ + ۰٫۵x_۴ + e$ تولید می‌کنیم که $x_۱$، $x_۲$، $x_۳$، $x_۴$ و e از توزیع نرمال استاندارد گرفته شده‌اند. پس از محاسبه y با استفاده از مقادیر تولید شدهٔ $x_۱$، $x_۲$، $x_۳$، $x_۴$ و e برآوردکننده‌های رگرسیونی بدست آمده از روش‌های مختلف و برآوردکننده ژرفترین رگرسیون را برای این مقادیر بدست می‌آوریم. نتایج بدست آمده در جدول زیر آمده‌اند.

با مقایسه ضرایب برآورد شده بدست آمده از روش‌های مختلف دیده می‌شود که ضرایب بدست آمده از روش ژرفترین رگرسیون نسبت به سایر روش‌ها به ضرایب مدل مفروض نزدیک‌تر است و در عین حال دارای بیشترین رگرسیون ژرفا است. به عبارت دیگر برآوردکننده ژرفترین رگرسیون

$$y = -۹٫۹۵ + ۱٫۱۳x_۱ + ۲٫۰۱x_۲ - ۵۰٫۱۴x_۳ + ۰٫۴۹۶x_۴$$

دارای رگرسیون ژرفای تقریبی ۱۸ است.

۸ نتیجه‌گیری

وقتی که استواری و تشخیص نقاط دورافتاده اهمیت دارند روش LMS روشی مناسب است. اما زمانی که نقاط دورافتادهٔ زیادی وجود ندارد (می‌توان به وسیله روش‌های LMS یا LTS بررسی کرد) و یکنوایی اهمیت دارد روش ژرفترین رگرسیون T_r^* انتخابی مناسب است.

مراجع

[1] Bloomfield, P., and Steiger, W. (1983). Least Absolute Deviations: Theory, Applications, and Algorithms, Boston:Birkhauser.

[2] Chatterjee, S., and Hadi, A.S., and Price, B. (2000), Regression Analysis by Example, New York: John Wiley.

[3] Donoho, D.L., and Huber, P.J. (1983), "The notion of breakdown point", in A Festschrift for Erich Lehmann, eds. Bickel, P., Doksum, K., and Hodges, J.L., Belmont: Wadsworth.

[4] Draper, N., Smith, H. (1998) Applied Regression Analysis, third edition, John Wiley & Sons.

[5] Rousseeuw, P.J., (1984) "Least Median of Squares Regression", Journal of the American Statistical Association, **79**, 871-880.

[6] Rousseeuw, P.J., and Hubert, M. (1999), "Regression Depth", Journal of the American Statistical Association, **94**, 388-402.

استفاده از معیارهای کولبک ـ لیبلر و فاصلهٔ چرنوف برای آنالیز ممیزی سری‌های زمانی چند متغیره

سارا شفیعی بابایی[۱]، رحیم چینی‌پرداز[۲]

[۱] دانشگاه شهید چمران، دانشکدهٔ علوم ریاضی و کامپیوتر، گروه آمار
[۲] دانشگاه شهید چمران، دانشکدهٔ علوم ریاضی و کامپیوتر، گروه آمار

چکیده: هدف از آنالیز ممیزی خطی در سری‌های زمانی بدست آوردن تابعی خطی می‌باشد که مطابق با آن تابع یک مشاهدهٔ سری زمانی را بتوان به یکی از دو یا چند جامعهٔ مستقل سری زمانی تخصیص داد. روش‌های کلاسیک برای بدست آوردن تابع ممیزی مبتنی بر ماکزیمم کردن فاصلهٔ ماهالونوبیس است. در سالهای اخیر فاصله‌های کولبک ـ لیبلر و چرنوف نیز بوسیلهٔ آماردانان مورد توجه قرار گرفته است. در این مقاله فرض شده است که جوامع سری زمانی به صورت مستقل و دارای توزیع نرمال هستند. این جوامع ممکن است دارای میانگین‌های برابر و ماتریس‌های کواریانس برابر و یا متفاوت از هم باشند. تابع ممیزی خطی با استفاده از ماکزیمم کردن فاصلهٔ کولبک ـ لیبلر و چرنوف بدست آمده است. با توجه به اینکه استفاده از حوزهٔ زمانی این تابع منجر به روش‌های پیچیده و محاسبات مشکل می‌شود از تقریب طیفی سری‌های زمانی در حوزهٔ فرکانس استفاده شده است. سپس تابع ممیزی خطی با همان روش ولی برای سری‌های زمانی چند متغیره گسترش داده شده است. در انتها با روش شبیه‌سازی برای مدلهای $ARMA(p,q)$ نتایج حاصل مورد بررسی و مقایسه قرار داده شده‌اند.

واژه‌های کلیدی: آنالیز ممیزی، تابع ممیزی خطی، آنالیز طیفی سری‌های زمانی، معیار ممیزی کولبک ـ لیبلر، معیار ممیزی چرنوف

۱ مقدمه

مساله آنالیز ممیزی و رده‌بندی از جمله مسایل پر اهمیت آماری است که به شکل کاربردهای مختلف و در بسیاری از علوم اهمیت خاصی پیدا کرده است. یکی از مهمترین و اساسی‌ترین کاربردهای آنالیز ممیزی در روش‌های کلاسیک تشخیص الگو در داده‌های آزمایشی سری‌های زمانی است که در موارد مختلف قابل توجه می‌باشد.
به عنوان مثال گوینز و دیگران (۱۹۷۵) کاربردهای آنالیز ممیزی را در داده‌های دریافت شده از دستگاه نوار مغز بررسی کردند. در این زمینه همچنین می‌توان به گرش و دیگران (۱۹۷۹) آلاگان (۱۹۸۶، ۱۹۸۹) راجی (۱۹۸۵) اشاره نمود. در زمینه کاربرد توابع ممیزی در داده‌های مربوط

به گویش‌های مختلف گزارش شده می‌توان به ولف (۱۹۷۶) مارکل و دیگران (۱۹۷۶) و دویت (۱۹۸۵) اشاره کرد.
در زمین‌شناسی نیز مساله آنالیز ممیزی بین داده‌های لرزه‌ای مربوط به دو منبع مختلف ایجاد موج؛ یعنی زمین لرزه و انفجار هسته‌ای بسیار مورد توجه واقع بوده است. چنانچه در این راستا می‌توان به تی جی استیم (۱۹۷۶) درگاهی ـ نوبری و لیکاک (۱۹۸۱) درگاهی ـ نوبری (۱۹۹۵) شاموی (۱۹۹۶ ؛ ۱۹۸۸) بلند فورد (۱۹۹۳) کاکیزاوا و دیگران (۱۹۹۸) اشاره نمود. پیکالو (۱۹۹۰) نیز از آنالیز ممیزی در داده‌های اقتصادی استفاده کرده است. همچنین چینی‌پرداز (۲۰۰۴) با استفاده از بردارهای اتوکواریانس تابع ممیزی را برای داده‌های مربوط به زمین لرزه و انفجار هسته‌ای پیشنهاد کرده است.

هدف این مقاله؛ تعیین و بدست آوردن توابع ممیزی خطی در دامنه فرکانسی سری‌های زمانی است. توابع ممیزی در حالت یک متغیره و چند متغیره با استفاده از معیار اطلاع ممیزی کولبک ـ لیبلر و معیاراطلاع ممیزی چرنوف بدست آورده می‌شوند و سپس تقریب‌های طیفی مناسب برای این توابع پیشنهاد می‌گردند. در پایان با یک مثال عددی نتایج استفاده از معیارهای ممیزی مقایسه می‌شوند.

۱.۱ توابع ممیزی در دامنه فرکانس

اگر مشاهدات سری زمانی به صورت $x_t = (x_0, x_1, ..., x_{T-1})$ تحت دو فرضیه H_j برای $j = 1, 2$ باشد به طوریکه:

$$H_j \ : x \sim N_T(0, R_j)$$

در آن صورت تابع ممیزی بر اساس نسبت درستنمایی چگالی دو جامعه به صورت زیر خواهد بود:

$$d_Q(x) = x'(R_2^{-1} - R_1^{-1})x$$

و تخصیص مشاهده x به یکی از دو جامعه به این صورت می‌باشد که هرگاه $d_Q(x) \geq 0$ مشاهده x به جامعه اول تخصیص داده می‌شود و هرگاه $d_Q(x) < 0$ مشاهده x به جامعه دوم تخصیص داده می‌شود. اما در این حالت توزیع $d_Q(x)$ تحت هر یک از فرضیه‌ها بسیار پیچیده می‌باشد؛ بنابراین با استفاده از تبدیلات فوریه گسسته بحث آنالیز ممیزی به شکل بهتر و ساده‌تری در دامنه فرکانس دنبال می‌شود. در این حالت متغیرهای تولید شده در دامنه فرکانس؛ متغیرهای تقریبا نرمال و ناهمبسته با ماتریس‌های کواریانس قطری می‌باشند؛ بطوریکه بین مشاهدات پشت سر هم همبستگی ناچیزی وجود دارد.
فرض کنید $X_t = (X(0), X(1), ..., X(T-1))'$ بردار تبدیل فوریه متناهی از مشاهدات x_t در دامنه فرکانس باشد؛ که به صورت زیر تعریف می‌شود:

$$X(k) = \frac{1}{\sqrt{T}} \sum_{k=0}^{T-1} x_t \exp\{i\lambda_k t\} \qquad \lambda_k = 2\pi k T^{-1}$$

با بردار میانگین:

$$M(k) = \frac{1}{\sqrt{T}} \sum_{k=0}^{T-1} \mu_j(t) \exp\{i\lambda_k t\} \qquad \lambda_k = 2\pi k T^{-1}$$

و تابع چگالی طیفی؛ یا توان طیف برابر با:

$$f_x(\lambda_k) = \sum_{k=0}^{T-1} R_k \exp\{i\lambda_k t\} \qquad \lambda_k = 2\pi k T^{-1}$$

باشد. در حقیقت دو رابطه اخیر تبدیل فوریه متناهی از میانگین $\mu_j(t)$ و ماتریس کواریانس R_j برای $j = 1, 2$ در دامنه فرکانس می‌باشد.
لگاریتم تقریبی تابع چگالی توأم تمام مقادیر تبدیل فوریه گسسته به صورت زیر خواهد بود:

$$\ln p_j(x) = \frac{1}{2} \sum_{k=0}^{T-1} \ln f_j(\lambda_k) - \frac{1}{2} \sum_{k=0}^{T-1} \frac{|\, X(k) - M_j(k) \,|^2}{f_j(\lambda_k)}$$

اگر فرض شودکه میانگین‌های جامعه‌ها نابرابر هستند؛ آنگاه تحت این فرض‌که $f_j(\lambda_k) = f(\lambda_k)$ برای $j = 1, 2$ تابع ممیزی خطی زیر تابع ممیزی بین دو جامعه خواهد بود:

$$d_L(x) = \mu_j' R^{0-1} x - \frac{1}{2} \mu_j' R^{0-1} \mu_j$$

$$= \sum_{k=0}^{T-1} \frac{\overline{M_j(k) X(k)}}{f(\lambda_k)} - \frac{1}{2} \sum_{k=0}^{T-1} \frac{|\, M_j(k) \,|^2}{f(\lambda_k)}$$

که در آن $R_j^0 = \left\{ r_j^0(s-t), t, s = 0, 1, \ldots T-1 \right\}$ و R^{0-1} دارای عناصری به فرم

$$r^{0-1}(s-t) = \frac{1}{\sqrt{T}} \sum_{k=0}^{T-1} \exp\{i\lambda_k(s-t)\} f(\lambda_k)$$

می‌باشند. [لیگت (۱۹۷۱)؛ شاموی و آنگر (۱۹۷۴) و شاموی (۱۹۸۲)].
به علاوه زمانیکه توابع کواریانس (مقادیر طیف) دو جامعه نابرابر باشند نیز؛ تابع رده‌بندی درجه دوم به صورت زیر بدست می‌آید:

$$d_Q^0(x) = x'(R_2^{0-1} - R_1^{0-1})x$$

$$= \sum_{k=0}^{T-1} \mid X(k)^2 \mid \left((f_2^{-1}(\lambda_k) - f_1^{-1}(\lambda_k) \right)$$

و بر اساس اینکه $d_Q^0(x) \geq 0$ یا $d_Q^0(x) < 0$ باشد؛ سری زمانی x_t به جامعه اول و یا دوم تخصیص داده می‌شود.

۲ آنالیز ممیزی بر اساس معیارهای جداسازی

در این بخش با بکارگیری تبدیلات فوریه گسسته و تقریب‌های طیفی؛ آنالیز ممیزی را در دامنه فرکانسی سری‌های زمانی بر اساس معیارهای ممیزی دنبال می‌کنیم.

۱.۲ آنالیز ممیزی بر اساس معیارهای ممیزی کولبک ـ لیبلر و چرنوف در حالت یک متغیره

فرض کنید مساله رده‌بندی؛ تخصیص یک فرایند ایستای نرمال $T \times 1$ بعدی $\mathbf{x}_t = (x_1, x_2, ..., x_t)'$ با بردار میانگین $\mu_j = (\mu_{j0}, \mu_{j1}, ..., \mu_{jT-1})'$ و ماتریس کواریانس $T \times T$ بعدی $R_j = \{\sigma_{j(s-t)}, t, s = 0, 1, ...T-1\}$ تحت فرضیه‌های H_j برای $j = 1, 2$ به یکی از دو جامعه نرمال باشد.
یکی از معیارهای کلاسیک جداسازی بین دو جامعه چند متغیره؛ معیارکولبک ـ لیبلر می‌باشد که به صورت زیر تعریف می‌شود: کولبک ـ لیبلر (۱۹۵۲)؛ کولبک (۱۹۷۸).

$$I(1, 2; x) = \frac{1}{T} E_1 \left\{ \log \frac{p_1(x)}{p_2(x)} \right\} \qquad (1)$$

که در آن $p_1(x)$ و$p_2(x)$ به ترتیب توابع چگالی مربوط به جامعه‌های نرمال تحت فرضیه‌های H_j برای $j = 1, 2$ در نظر گرفته می‌شود و T طول بردار مشاهدات سری زمانی است. به سادگی می‌توان رابطه (۱) را به صورت رابطه زیر بدست آورد:

$$I(1, 2; x) = \frac{1}{2T} \left(\left(tr(R_1 R_2^{-1}) \log \frac{\mid R_2 \mid}{\mid R_1 \mid} - T \right) + \delta' R_2^{-1} \delta \right) \qquad (2)$$

که در آن $\delta_t = \mu_{2t} - \mu_{1t}$ برای $t = 0, 1, ...T-1$ می‌باشد.
به دلیل نامتقارن بودن این معیار؛ با تعریف رابطه زیر مقدار متقارنی از معیارکولبک ـ لیبلر حاصل می‌شود:

$$J(1, 2; x) = I(1, 2; x) + I(2, 1; x) \qquad (3)$$

همچنین معیار جداسازی دیگری که بین دو جامعه چند متغیره به عنوان یک معیار ممیزی بکار برده می‌شود؛ معیار چرنوف [چرنوف (۱۹۵۲)؛ رینی (۱۹۶۱)] می‌باشد که پارزن (۱۹۹۰) آن را در ممیزی بین سری‌های زمانی پیشنهاد نموده و به صورت زیر تعریف می‌گردد.

$$Q_r(1,2;x) = -\frac{1}{2T}\log E_1\left\{\log\left(\frac{p_1(x)}{p_2(x)}\right)^r\right\} \qquad 0 < r < 1$$

$$= \frac{1}{2T}\left(\log\frac{\mid rR_1 + (1-r)R_2 \mid}{\mid R_2 \mid} - r\log\frac{\mid R_1 \mid}{\mid R_2 \mid}\right) \qquad (۴)$$

با استفاده از این خاصیت که $Q_r(1,2;x) = Q_{1-r}(2,1;x)$ می‌توان معیار متقارنی از $Q_r(1,2;x)$ به صورت زیر تعریف نمود:

$$J_{Q_r}(1,2;x) = Q_r(1,2;x) + Q_r(1,2;x) \qquad (۵)$$

با توجه به اینکه در دامنه زمان استفاده از این معیارها مشکل می‌باشد؛ لذا برای ممیزی بین دو جامعه نرمال از روش‌های طیفی استفاده می‌گردد. بدین ترتیب فرض کنید سری زمانی x_t تحت فرضیه‌های H_j برای $j = 1, 2$ دارای میانگین μ_j و ماتریس کواریانس $T \times T$ بعدی $R_j = \{\sigma_{j(s-t)}, t, s = 0, 1, ...T-1\}$ باشد. آنگاه رابطه معکوس تبدیل فوریه زیر برای تابع کواریانس وجود دارد و به شکل زیر است:

$$\sigma_j(s-t) = \int_{-\pi}^{\pi} f_j(\lambda) e^{i\lambda(s-t)}\left(\frac{d\lambda}{2\pi}\right)$$

که در آن $f_j(\lambda)$ طیف دو جامعه می‌باشد و فرض می‌شود که بر روی فاصله $[-\pi, \pi]$ مثبت؛ پیوسته و به طور مطلق انتگرال پذیر است. فولر (۱۹۹۶).
در این صورت فرض می‌شود دنباله $\delta_t = \mu_2 t - \mu_1 t$ برای $t = 0, 1, ...T-1$ در شرایط زیر صدق کند:

$$(I) \quad \sup_t \mid \delta_t \mid = c < \infty$$

$$(II) \quad \rho_T(\tau) = \frac{1}{T}\sum_{t=0}^{T-1-|\tau|} \delta_{t+|\delta|}\delta_t$$

که $\rho_t(\tau)$ تابع خود همبستگی سری زمانی x_t می‌باشد و دارای حد داده شده در رابطه زیر است:

$$\rho(\tau) = \lim_{T\to\infty} \rho_T(\tau) = \int_{-\pi}^{\pi} e^{i\lambda\ tau}\left(\frac{dM\lambda}{2\pi}\right)$$

که $M(\lambda)$ تابع یکنوای اکیداً غیر نزولی است. به گونه‌ای که $M(-\pi)$ و از راست پیوستگی دارد. همچنین $\rho(\circ) > \circ$. فولر (۱۹۹۶).

iii) اگر $x_t = (x_1, x_2, ..., x_t)'$ فرایند سری زمانی ایستای نرمال با میانگین μ_j و ماتریس کواریانس R_j تحت فرضیه‌های H_j برای $j = 1, 2$ باشد؛ آنگاه اگر تابع کواریانس $\sigma_j(s-t)$ دارای تابع چگالی مطلقا انتگرال‌پذیر و مثبت و پیوسته $f_j(\lambda)$ در فاصله $[-\pi, \pi]$ باشد. همچنین تابع δ_t در شرایط (i) و (ii) صدق کند؛ در این صورت حد اطلاع ممیزی رابطه (۲) و (۴) به صورت زیر داده می‌شود: [شاموی و آنگر (۱۹۷۴)؛ کاکیزاوا و دیگران (۱۹۹۸)].

$$I(1, 2; x) = \frac{1}{2T} \int_{-\pi}^{\pi} \left(\frac{f_1(\lambda)}{f_2(\lambda)} - \log \frac{f_1(\lambda)}{f_2(\lambda)} - 1 \right) \frac{d\lambda}{2\pi} + \frac{1}{2} \int_{-\pi}^{\pi} \frac{1}{f_2(\lambda)} \frac{dM\lambda}{2\pi}$$

و اگر میانگین‌های دو جامعه برابر با صفر در نظرگرفته شوند؛ قسمت دوم از سمت راست رابطه بالا برابر با صفر خواهد بود.

لذا:

$$I(1, 2; x) = \frac{1}{2T} \int_{-\pi}^{\pi} \left(\frac{f_1(\lambda)}{f_2(\lambda)} - \log \frac{f_1(\lambda)}{f_2(\lambda)} - 1 \right) \frac{d\lambda}{2\pi} \qquad (۶)$$

همچنین:

$$Q_r(1, 2; x) = \frac{1}{2T} \int_{-\pi}^{\pi} \left(\log \frac{\mid r f_1(\lambda) + (1-r) f_2(\lambda) \mid}{\mid f_2(\lambda) \mid} - r \log \frac{\mid f_1(\lambda) \mid}{\mid f_2(\lambda) \mid} \right) \frac{d\lambda}{2\pi} \qquad (۷)$$

در نتیجه:

$$I(1, 2; x) = \lim_{T \to \infty} I_T(1, 2, x)$$

و به همین صورت:

$$I(2, 1; x) = \lim_{T \to \infty} I_T(2, 1, x)$$

و بنابراین:

$$J(1, 2; x) = \lim_{T \to \infty} J_T(1, 2, x)$$

بدست می‌آید.

بنابراین حد اطلاع ممیزی که به عنوان تقریب طیفی برای معیارهای ممیزی کولبک ـ لیبلر و چرنوف در نظرگرفته می‌شود؛ به معیارهای اصلی ممیزی همگرا می‌باشد. یعنی اینکه صورت‌های حدی برای تقریب معیارهای ممیزی و معیار واگرایی J با فرم‌های معادل برای فرایند اصلی x_t مشابه می‌باشند.

پس می‌توان نوشت: [شاموی و آنگر (۱۹۷۴)؛ کاکیزاوا و دیگران (۱۹۹۸)؛ کازاکاس ـ پاپ‌آنتونی کازاکس (۱۹۸۰)].

$$I(1, 2; x^{\circ}) = \frac{1}{2T} \sum_{k=\circ}^{T-1} \left[tr \left(f_1(\lambda_k) f_2^{-1}(\lambda_k) \right) - \log \frac{f_1(\lambda_k)}{f_2(\lambda_k)} - 1 \right] \qquad (۸)$$

و همچنین:

$$Q_r(1,2;x^\circ)=\frac{1}{2T}\sum_{k=0}^{T-1}\left(\log\frac{|\,rf_1(\lambda_k)+(1-r)f_2(\lambda_k)\,|}{|\,f_2(\lambda_k)\,|}-r\log\frac{|\,f_1(\lambda_k)\,|}{|\,f_2(\lambda_k)\,|}\right) \quad (9)$$

ملاحظه می‌شود که این روابط؛ حتی برای مقادیر بزرگ T نیز به سادگی قابل محاسبه می‌باشد. بنابراین:

$$\lim_{T\to\infty} I_T(1,2,x^\circ)=I(1,2;x)$$

که در آن $I(1,2;x)$ در رابطه (۶) داده شده است. همچنین:

$$\lim_{T\to\infty} Q_{rT}(1,2,x^\circ)=Q_r(1,2;x)$$

به همین ترتیب می‌توان رابطه‌های بالا را برای $I(2,1;x^\circ)$ و $J(1,2;x^\circ)$ بدست آورد.
همچنین آنگر (۱۹۷۳) با استفاده از نتایجی که در واهبا (۱۹۶۸) و لیگت (۱۹۷۱) اثبات شده است نشان داده است که:

$$|\,I(1,2,x)-I(1,2,x^\circ)\,|=O(T^{-1})$$

به شرطی که:

$$\sum_{-\infty}^{+\infty}|\,t\,|^{1+\beta}|\,\sigma_j(t)\,|\leqslant M_j^\beta<\infty \quad 0<\beta<1 \quad j=1,2$$

که در آن $M>0$ یک عدد ثابت می‌باشد.
اکنون اگر $f_T(\lambda)$ به عنوان برآوردیاب نااریب ماتریس طیفی هر سری برداری نمونه‌ای در نظر گرفته شود که به صورت ناپارامتری بدست آمده است و $f_{(\lambda)}$ به عنوان طیف جامعه jام برای $j=1,2$ در نظر گرفته شود. در آن صورت تقریب زیر برای معیارهای ممیزی کولبک ـ لیبلر و چرنوف به ترتیب مطابق با رابطه‌های (۸) و (۹) به صورت زیر بکار برده می‌شوند:

$$I(f_T,f_j)=\frac{1}{2T}\sum_{k=0}^{T-1}\left[\left(tr(f_T(\lambda_k)f_j^{-1}(\lambda_k)\right)-\log\frac{f_T(\lambda_k)}{f_j(\lambda_k)}-1\right] \quad (10)$$

و همچنین:

$$Q_r(f_T,f_j)=\frac{1}{2T}\sum_{k=0}^{T-1}\left(\log\frac{|\,rf_T(\lambda_k)+(1-r)f_j(\lambda_k)\,|}{|\,f_j(\lambda_k)\,|}-r\log\frac{|\,f_T(\lambda_k)\,|}{|\,f_j(\lambda_k)\,|}\right) \quad (11)$$

لذا با مینیمم کردن اطلاع ممیزی بین چگالی‌های طیفی هر یک از دو نمونه سری زمانی با جامعه؛ قاعده ممیزی به صورت زیر فراهم می‌شود:
سری زمانی x_t به ترتیب به جامعه اول و یا دوم تخصیص داده می‌شود؛ اگر به ترتیب $I(f_T, f_2) \geq I(f_T, f_1)$ و $I(f_T, f_2) < I(f_T, f_1)$ حال اگر گروه ماتریس طیفی بوسیله متوسط ماتریس طیفی نمونه‌ای برای جامعه jام به صورت زیر برآورد شود:

$$\bar{f}_j(\lambda_k) = \frac{1}{n_j} \sum_{l=1}^{n_j} f_T^{(l)}(\lambda_k)$$

که در آن $f_T^{(l)}(\lambda_k)$ برآورد ناپارامتری ماتریس طیفی lامین نمونه از جامعه jام می‌باشد؛ می‌توان با جایگزین کردن $\bar{f}_j$ با f_j قاعده ممیزی بالا را به صورت زیر بدست آورد:
سری زمانی x_t به ترتیب به جامعه اول و یا دوم تخصیص داده می‌شود؛ اگر به ترتیب $I(f_T, \bar{f}_2) \geq I(f_T, \bar{f}_1)$ و $I(f_T, \bar{f}_2) < I(f_T, \bar{f}_1)$ برقرار باشد.

۲.۲ آنالیز ممیزی بر اساس معیارهای ممیزی کولبک ـ لیبلر و چرنوف در حالت چند متغیره

فرض کنید مساله رده‌بندی؛ تخصیص سری‌های زمانی بردای $pT \times 1$ بعدی $x = (x_0', x_1', ..., x_{T-1}')'$ که شامل p بردار سری زمانی یک متغیره $p \times 1$ بعدی $\mathbf{x}_t = (x_0, x_1, ..., x_{T-1})'$ با ماتریس کواریانس $pT \times pT$ بعدی $R_j(s-t)$ برای $s, t = 0, 1, ..., T-1$ که خود شامل ماتریس‌های $p \times p$ بعدی $R_j = \{\sigma_{j(s-t)}, t, s = 0, 1, ...T-1\}$ تحت فرضیه‌های جدای H_j برای $j = 1, 2$ باشد. در این حالت معیارهای ممیزی کولبک ـ لیبلر و چرنوف به صورت زیر می‌باشند.

$$I(1, 2; x) = \frac{1}{2T} \left(tr(R_1 R_2^{-1}) - \log \frac{\mid R_1 \mid}{\mid R_2 \mid} - pT \right) \qquad (12)$$

$$Q_r(1, 2; x) = -\frac{1}{2T} \left(\log \frac{\mid rR_1 + (1-r)R_2 \mid}{\mid R_2 \mid} - r \log \frac{\mid R_1 \mid}{\mid R_2 \mid} \right) \qquad (13)$$

مانند حالت یک متغیره می‌توان معیار واگرایی J را به دلیل نامتقارن بودن معیارهای ممیزی کولبک ـ لیبلر و چرنوف به صورت زیر تعریف کرد:
کاکیزاوا و دیگران (۱۹۹۸).

$$J(1, 2; x) = I(1, 2; x) + I(2, 1; x)$$

$$J_{Qr}(1, 2; x) = Q_r(1, 2; x) + Q_r(2, 1; x)$$

ملاحظه می‌گردد که برای محاسبه رابطه‌های (۱۲) و (۱۳) لازم به محاسبه وارون و دترمینان ماتریس‌هایی با ابعاد بزرگ می‌باشد؛ و مخصوصاً زمانی که مقدار T بزرگ باشد این کار بسیار دشوار است. در این حالت نیز استفاده از تقریب‌های طیفی بسیار مفید می‌باشد؛ چنانچه در تبدیل ماتریس‌های کواریانس به ماتریس‌های طیفی نه تنها هیچ اطلاعی از دست نخواهد رفت بلکه بعد مجموعه ماتریس‌های از مرتبه $pT \times pT$ به ماتریس‌های از مرتبه $p \times p$ کاهش می‌یابد. در این زمینه می‌توان به هانن (۱۹۷۰)؛ پینسکر (۱۹۶۴)؛ شاموی و آنگر (۱۹۷۴) کاکیزاوا و دیگران (۱۹۹۸) مراجعه نمود. کازاکس ـ پاپ‌آنتونی کازاکس (۱۹۸۰) نیز خواص حدی رابطه‌های (۱۵) و (۱۶) را به صورت زیر فراهم نمودند:

$$\lim_{T\to\infty} \frac{1}{T} I(1, 2; x_t) \tag{۱۴}$$

$$= I(f_1, f_2; x_t) = \frac{1}{2T} \int_{-\pi}^{\pi} \left(tr\left(f_1(\lambda) f_2^{-1}(\lambda)\right) - \log \frac{f_1(\lambda)}{f_2(\lambda)} - p \right) \frac{d\lambda}{2\pi}$$

$$\lim_{T\to\infty} \frac{1}{T} Q_R(1, 2; x_t) \tag{۱۵}$$

$$= Q_r(f_1, f_2; x_t) = \frac{1}{2T} \int_{-\pi}^{\pi} \left(\log \frac{\mid r f_1(\lambda) + (1-r) f_2(\lambda) \mid}{\mid f_2(\lambda) \mid} - r \log \frac{\mid f_1(\lambda) \mid}{\mid f_2(\lambda) \mid} \right) \frac{d\lambda}{2\pi}$$

که در آن $f_j(\lambda_k)$ ماتریس‌های طیفی $p \times p$ بعدی جامعه‌ها برای $j = 1, 2$ می‌باشند.
با استفاده از تبدیل انتگرال به مجموع ریمان رابطه‌های (۱۴) و (۱۵) را می‌توان به صورت زیر بدست آورد:

$$I(1, 2; x) = \frac{1}{2T} \sum_{k=0}^{T-1} \left[tr\left(f_1(\lambda_k) f_2^{-1}(\lambda_k)\right) - \log \frac{f_1(\lambda_k)}{f_2(\lambda_k)} - p \right] \tag{۱۶}$$

و همچنین:

$$Q_r(1, 2; x) = \frac{1}{2T} \sum_{k=0}^{T-1} \left(\log \frac{\mid r f_1(\lambda_k) + (1-r) f_2(\lambda_k) \mid}{\mid f_2(\lambda_k) \mid} - r \log \frac{\mid f_1(\lambda_k) \mid}{\mid f_2(\lambda_k) \mid} \right) \tag{۱۷}$$

ملاحظه می‌شود که رابطه‌های (۱۶) و (۱۷) هر دو تابعی از ماتریس $tr\left(f_1(\lambda_k) f_2^{-1}(\lambda_k)\right)$ می‌باشند و در موارد خاصی به شکل عمومی‌تر برای بعضی از مقادیر تابع $H(.)$؛ که یک اندازه جداسازی عمومی‌تر به شکل زیر است می‌توان تعریف نمود: [کاکیزاوا و دیگران (۱۹۹۸)].

$$D_H(f_1, f_2) = \frac{1}{2} \int_{-\pi}^{\pi} H\left(f_1(\lambda) f_2^{-1}(\lambda)\right) \frac{d\lambda}{2\pi} \tag{۱۸}$$

به طوریکه:

$$H(\lambda) = f_1(\lambda) f_2^{-1}(\lambda)$$

لازم به ذکر است که فاصله عمومی $D_H(f_1, f_2)$ در رابطه (۱۸) دارای خاصیت شبه فاصله‌ای است. همچنین تابع $H(.)$ را می‌توان به شکل‌های مختلفی برگزید به طوریکه خاصیت شبه فاصله‌ای آن حفظ گردد. در ادامه تنها برای دو رابطه (۱۲) و (۱۳) این تابع در نظرگرفته می‌شود که می‌توان آن را به صورت زیر نوشت: [کاکیزاوا و دیگران (۱۹۹۸)].

$$H_I(Z) = tr(Z) - \log \mid Z \mid -p \qquad (۱۹)$$

$$H_Q(Z) = \log \mid rZ + (1-r) I_p \mid -r \log \mid Z \mid$$

نکته‌ای که باید توجه کرد این است که عموماً فاصله $D_H(f_1, f_2)$ متقارن نیست و با تعریف رابطه زیر مقدار متقارنی از $D_H(f_1, f_2)$ به صورت زیر بدست می‌آید:

$$\widetilde{H(z)} = H(Z) + H(Z^{-1}) \qquad (۲۰)$$

و با قرار دادن رابطه‌های (۱۹) در رابطه (۲۰) معیار متقارنی از معیارهای کولبک ـ لیبلر و چرنوف فراهم می‌شود. کاکیزاوا و دیگران (۱۹۹۸). با جمع بستن بر روی فرکانس‌هایی به شکل $\lambda_k = 2\pi k T^{-1}$؛ $k = 0, 1, ..., T-1$ رابطه زیر بدست می‌آید:

$$D_H(f_1, f_2) \approx \frac{1}{2T} \sum_k \Big(f_1(\lambda_k) f_2^{-1}(\lambda_k) \Big) \qquad (۲۱)$$

اکنون اگر $f_T(\lambda)$ به عنوان برآورد ماتریس طیفی هر سری برداری نمونه‌ای باشد که به صورت ناپارامتری بدست آمده است و $f_j(\lambda)$ به عنوان طیف جامعه jام برای $j = 1, 2$ در نظرگرفته شود. در آن صورت تقریب زیر برای معیارهای ممیزی کولبک ـ لیبلر و چرنوف به ترتیب مطابق با رابطه‌های (۱۶) و (۱۷) به صورت زیر بکار برده می‌شوند:

$$I(f_T, f_j) = \frac{1}{2T} \sum_{k=0}^{T-1} \left[\Big(tr(f_T(\lambda_k) f_j^{-1}(\lambda_k) \Big) - \log \frac{f_T(\lambda_k)}{f_j(\lambda_k)} - p \right] \qquad (۲۲)$$

و همچنین:

$$Q_r(f_T, f_j) = \frac{1}{2T} \sum_{k=0}^{T-1} \left(\log \frac{\mid r f_T(\lambda_k) + (1-r) f_j(\lambda_k) \mid}{\mid f_j(\lambda_k) \mid} - r \log \frac{\mid f_T(\lambda_k) \mid}{\mid f_j(\lambda_k) \mid} \right) \qquad (۲۳)$$

با این تعریف این رابطه‌ها؛ یک روش عمومی بر اساس رابطه‌های (۱۵) و (۱۶) می‌توان بدست آورد. با این روش اندازه جداسازی بین طیف نمونه‌ای x_t و جامعه jام بوسیله $D_H(f_T, f_j)$ اندازه‌گیری می‌شود و این اندازه به عنوان تابع ممیزی در نظرگرفته می‌شود: بر اساس این معیار تابع ممیزی به صورت زیر می‌باشد:

$$D_H = D_H(f_T, f_۲) - D_H(f_T, f_۱) \geq ۰ \qquad (۲۴)$$

سری زمانی $\mathbf{x}_t$ به ترتیب به جامعه اول و یا دوم تخصیص داده می‌شود؛ اگر به ترتیب $D_H \geq ۰$ و $D_H < ۰$ برقرار باشد.
به ازای $H_I(Z)$ داده شده در رابطه (۱۹) مقدار $D_H(f_T, f_j)$ شامل مینیمم اطلاعات ممیزی می‌باشد که اولین بار توسط کولبک (۱۹۵۱) بیان شده است.
می‌توان آنالیز ممیزی را با استفاده از معیار ممیزی کولبک ـ لیبلر نیز انجام داد که با مینیمم کردن اطلاع ممیزی بین چگالی‌های طیفی هر یک از دو نمونه سری زمانی با جامعه؛ تابع ممیزی با استفاده از رابطه‌های (۲۲) و (۲۴) به صورت زیر بدست آورده می‌شود:

$$D_H(I) = I(f_T, f_۲) - I(f_T, f_۱)$$

و قاعده ممیزی بدین صورت می‌باشد که سری زمانی x_t به ترتیب به جامعه اول و یا دوم تخصیص داده می‌شود؛ اگر به ترتیب $D_H(I) \geq ۰$ و $D_H(I) < ۰$ برقرار باشد.
هچنین بر اساس معیار چرنوف نیز می توان تابع ممیزی را با استفاده از رابطه‌های (۲۳) و (۲۴) به صورت زیر بدست آورد:

$$D_H(I) = Q_r(f_T, f_۲) - Q_r(f_T, f_۱)$$

و قاعده ممیزی بدین صورت می‌باشد که سری زمانی x_t به ترتیب به جامعه اول و یا دوم تخصیص داده می‌شود؛ اگر به ترتیب $D_H(Q_r) \geq ۰$ و $D_H(Q_r) < ۰$ برقرار باشد.
حال اگر گروه ماتریس طیفی بوسیله متوسط ماتریس طیفی نمونه‌ای برای جامعه jام به صورت زیر برآورد شود:

$$\bar{f}_j(\lambda_k) = \frac{۱}{n_j} \sum_{l=۱}^{n_j} f_T^{(l)}(\lambda_k)$$

که در آن $f_T^{(l)}(\lambda_k)$ برآورد ناپارامتری ماتریس طیفی lامین نمونه از جامعه jام می‌باشد؛ می‌توان با جایگزین کردن $\bar{f}_j$ با f_j قاعده ممیزی بالا را به صورت زیر بدست آورد:
سری زمانی x_t به ترتیب به جامعه اول و یا دوم تخصیص داده می‌شود؛ اگر به ترتیب $I(f_T, \bar{f}_۲) - I(f_T, \bar{f}_۱) \geq ۰$ و $I(f_T, \bar{f}_۲) - I(f_T, \bar{f}_۱) < ۰$ برقرار باشد.

همچنین بر اساس معیار چرنوف سری زمانی x_t به ترتیب به جامعه اول و یا دوم تخصیص داده می‌شود؛ اگر به ترتیب $Q_r(f_T, \bar{f}_۲) - Q_r(f_T, \bar{f}_۱) \geq ۰$ و

$Q_r(f_T, \bar{f}_۲) - Q_r(f_T, \bar{f}_۱) < ۰$ برقرار باشد.
کاکیزاوا و دیگران (۱۹۹۸) استفاده از روش‌های درستنمایی را در یافتن تابع ممیزی بکار بردند. این روش یکی از روش‌های معمول در یافتن تابع ممیزی بین دو سری زمانی ایستای نرمال با میانگین صفر ماتریس‌های طیفی $f_j(\lambda)$ برای $j = ۱, ۲$ می‌باشد و به صورت زیر است:

$$L(f_i) = \log p_j(x) = \frac{۱}{۲T} \sum_{k=۰}^{T-۱} \left[j(\lambda_k) - \frac{|X(k)|^۲}{f_j(\lambda_k)} \right]$$

که در آن:

$$| X(k) |^۲ = I_T(\lambda) = \frac{۱}{۲\pi T} \left\{ \sum_{k=۰}^{T-۱} x_t \exp\{-i\lambda_k t\} \right\} \left\{ \sum_{k=۰}^{T-۱} x_t \exp\{-i\lambda_k t\} \right\}^*$$

اگر فرایند نرمال باشد: در نتیجه تقریب طیفی لگاریتم راستنمایی x_t به صورت زیر ساده می‌شود: [وایتل (۱۹۵۴)].

$$L(f_j) = \frac{۱}{۲T} \sum_{k=۰}^{T-۱} \left[\log f_j(\lambda_k) + tr \left\{ I_T(\lambda_k) f_j^{-۱}(\lambda_k) \right\} \right]$$

و بوسیله تقریب مجموع به حالت انتگرال رابطه بالا به صورت رابطه زیر خلاصه می‌شود:

$$L(f_j) = \frac{۱}{۲T} \int_{-\pi}^{\pi} \left[\log f_j(\lambda_k) + tr \left\{ I_T(\lambda_k) f_j^{-۱}(\lambda_k) \right\} \right] \frac{d\lambda}{۲\pi}$$

برای $\lambda_k = ۲\pi k T^{-۱}$, $k = ۱, ۲, ..., T-۱$.
قاعده ممیزی بر اساس نسبت راستنمایی به صورت زیر می‌باشد که سری زمانی x_t به ترتیب به جامعه اول و یا دوم تخصیص داده می‌شود؛ اگر به ترتیب $L(f_۱) - L(f_۲) \geq ۰$ و $L(f_۱) - L(f_۲) < ۰$ برقرار باشد.

۳ مقایسه روش‌های ممیزی برای دو فرایند $AR(۱)$

در این بخش به منظور مقایسه روش‌های ممیزی در استفاده از تابع ممیزی کولبک ـ لیبلر و تابع ممیزی چرنوف؛ در ممیزی بین دو فرایند $AR(۱)$ با دو پارامتر مختلف $\phi_۱$ و $\phi_۲$ به ترتیب برای فرضیه اول و دوم؛ نرخ‌های خطا یا رده‌بندی نادرست؛ محاسبه می‌شوند. برای این کار تعداد ۲۰۰ سری زمانی ایستای $AR(۱)$ هرکدام به طول ۱۰۰ مشاهده در نرم‌افزار $S-plus$ ۲۰۰۰ شبیه‌سازی شده است. جدول‌های (۱) و (۲) به ترتیب نشان دهنده نرخ‌های خطا یا رده‌بندی

نادرست به ترتیب برای تابع ممیزی کولبک ـ لیبلر و تابع ممیزی چرنوف می‌باشند.
چنانچه در جداول (۱) و (۲) دیده می‌شود: هر دو تابع ممیزی در ممیزی بین دو جامعه به طور مناسب و خوبی عمل کرده‌اند. چنانچه مقادیر خطای رده‌بندی نادرست برای آنها تقریباً کم می‌باشد. همچنین همان طورکه دیده می‌شود تابع ممیزی کولبک ـ لیبلر در مقایسه با تابع ممیزی چرنوف در ممیزی بین دو جامعه بهتر عمل کرده است؛ چنانچه خطای رده‌بندی نادرست برای آن کمتر می‌باشد.

جدول ۱: مقادیر نرخ خطا (رده‌بندی نادرست) برای دو فرایند $AR(1)$ با پارامترهای مختلف ϕ_1 و ϕ_2 با استفاده از تابع ممیزی چرنوف

۰٫۸	۰	۰	۰	۰	۰	۰	۰	۰	۰	۷	۶	۴	۷	۸	۲۵	۳۳	*	-۰٫۸	
۰٫۷	۰	۰	۰	۰	۰	۰	۰	۰	۱	۷	۴	۲۰	۱۹	۲۰	۴۱	*	۲۷	-۰٫۷	
۰٫۶	۰	۰	۰	۰	۰	۰	۰	۲	۱۲	۱۵	۲۷	۳۳	۳۷	*	۳۰	۱۹	-۰٫۶		
۰٫۵	۰	۰	۰	۰	۰	۳	۲	۱۰	۱۰	۲۰	۲۵	۴۴	*	۳۷	۱۱	۹	-۰٫۵		
۰٫۴	۰	۰	۰	۰	۲	۳	۴	۱۲	۱۸	۲۸	۳۹	*	۴۱	۲۹	۱۲	۳	-۰٫۴		
۰٫۳	۰	۰	۰	۰	۳	۵	۷	۱۵	۳۰	۴۱	*	۳۷	۲۹	۱۱	۵	۱	-۰٫۳		
۰٫۲	۰	۰	۱	۱	۶	۱۰	۱۶	۱۶	۳۸	*	۳۹	۱۸	۱۲	۶	۴	۰	-۰٫۲		
۰٫۱	۰	۰	۱	۲	۴	۸	۲۳	۲۷	*	۴۴	۲۵	۱۵	۱۱	۵	۰	۰	-۰٫۱		
-۰٫۱	۰	۰	۳	۱۴	۱۶	۲۸	۴۱	*	۳۲	۱۳	۹	۳	۴	۰	۰	۰	۰		
-۰٫۲	۰	۴	۸	۱۸	۳۶	۳۳	*	۴۵	۲۳	۱۸	۵	۵	۲	۰	۰	۰	۰		
-۰٫۳	۲	۸	۱۹	۲۵	۳۶	*	۳۹	۲۹	۹	۱۶	۳	۱	۰	۰	۰	۰	۰		
-۰٫۴	۹	۱۳	۲۲	۳۵	*	۳۵	۲۸	۲۳	۱۳	۷	۳	۰	۰	۰	۰	۰	۰		
-۰٫۵	۷	۱۵	۴۱	*	۳۶	۳۶	۲۷	۲۰	۴	۳	۳	۰	۰	۰	۰	۰	۰		
-۰٫۶	۱۵	۳۴	*	۳۲	۲۵	۱۷	۱۶	۷	۲	۰	۲	۱	۰	۰	۰	۰	۰		
-۰٫۷	۲۸	*	۳۷	۳۰	۱۶	۹	۵	۹	۴	۱	۰	۰	۰	۰	۰	۰	۰		
-۰٫۸	*	۳۲	۱۹	۱۳	۱۱	۸	۶	۳	۰	۰	۰	۰	۰	۰	۰	۰	۰		
	-۰٫۸	-۰٫۷	-۰٫۶	-۰٫۵	-۰٫۴	-۰٫۳	-۰٫۲	-۰٫۱	۰٫۱	۰٫۲	۰٫۳	۰٫۴	۰٫۵	۰٫۶	۰٫۷	۰٫۸				

جدول ۲: مقادیر نرخ خطا (رده‌بندی نادرست) برای دو فرایند $AR(1)$ با پارامترهای مختلف ϕ_1 و ϕ_2 با استفاده از تابع ممیزی کولبک ـ لیبلر

۰/۸	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲۲	*
۰/۷	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲۴	*	۱۴
۰/۶	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۷	۲۲	*	۲۰	۱
۰/۵	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۲	۸	۲۷	*	۲۳	۳	۰	۰
۰/۴	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۹	۳۲	*	۱۳	۳	۰	۰	۰
۰/۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۹	۲۷	*	۲۳	۷	۱	۰	۰	۰
۰/۲	۰	۰	۰	۰	۰	۰	۰	۳	۲۸	*	۱۸	۳	۱	۰	۰	۰	۰	۰	۰
۰/۱	۰	۰	۰	۰	۰	۱	۳	۷	*	۲۷	۱۱	۷	۰	۰	۰	۰	۰	۰	۰
-۰/۱	۰	۰	۰	۰	۰	۸	۳۰	*	۹	۵	۰	۰	۰	۰	۰	۰	۰	۰	۰
-۰/۲	۰	۰	۰	۱	۶	۲۳	*	۲۱	۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
-۰/۳	۰	۰	۲	۷	۲۲	*	۱۸	۷	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
-۰/۴	۰	۰	۵	۲۱	*	۱۸	۷	۴	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
-۰/۵	۰	۱	۱۴	*	۲۴	۵	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
-۰/۶	۱	۱۹	*	۲۶	۶	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
-۰/۷	۷	*	۲۴	۳	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
-۰/۸	*	۱۸	۴	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
	-۰/۸	-۰/۷	-۰/۶	-۰/۵	-۰/۴	-۰/۳	-۰/۲	-۰/۱	۰/۱	۰/۲	۰/۳	۰/۴	۰/۵	۰/۶	۰/۷	۰/۸			

مراجع

[1] Alogon, J.(1989), Spectral Discrimination for Tow Groups of Time Series, J. of Time Series Analysis, vol 10, 203-214.

[2] Blandford, R. R. (1993), Discrimination of Earthquakes and Explosions at Regional Distances Using Complexity, AFTACTR-93-044, HQ. Air Force Technical Applications Center, Patrich Air Force Base, FL.

[3] Chernoff, H. (1952), A Measure of Asymptotic Efficiency for Tests a Hypothesis Based on The Sum of The Observations, Ann. Math. Stat, vol 25, 573-578.

[4] Chinipardaz, R. (2004), Discriminatin in Seismology Based On Autocovariances, Far East J, Theo. Stat. vol 12(1), 101-115.

[5] Dargahi- Noubary, G. R. and Laycock, P. J. (1981), Spectral Ratio Discrimination Theory, J. of Time Series Analysis, vol 2(2), 71-86.

[6] Dargahi- Noubary, G. R. (1992), Discrimination Between Gaussian Time Series Based on Their Spectral Differences, Comm. Stat. Theory and Method, vol 21(9), 2439-2458.

[7] Dewitt, P. E. (1985), His Master's (Digital) Voise, Time 125, No 13, 84-85.

[8] Geresh, W. Martinelli, F. Yonemoto, J. Lew, M. D. & McEwan, J. A. (1979), Automatic Classification of Electroencephalograms: Kullback- Leibler Nearest neighbor Rules, Science, vol 205, 193-195.

[9] Gevins, A. S. Veager, C. L. Diamond, S. Spire, J. Zeitlin, G. & Gevins, A. (1975), Automated Analysis of The Electerical Activity of The Human Brain (EEG), A Progress Report, Proc. IEEE 63, 1382-1399.

[10] Fuller, W. A. (1996), Introduction to Statistical Time Series, John Wiley, New York.

[11] Hannan, E. J. (1968), Multiple Time Series, Wiley, New York. (1992), Discrimination Between Gaussian Time Series Based on Their Spectral Differences, Comm. Stat. Theory and Method, vol 21(9), 2439-2458. (1992)

[12] Kakizava, Y., Shumway, R. H. and Taniguchi, M. (1998), Discrimination and Clustering for Multivariate Time Series, JASA, vol 93, NO, 441,theory and Methods.

[13] Kullback, S. & Leibler, R. A. (1951), An Information and Sufficiency, Ann. Math. Stat, vol 22, 76-86.

[14] Kullback, S. (1952), An Application of Information Theory to Multivariate Analysis, Ann. Math. Stat, vol 23, 88-102.

[15] Liggett, W. S. (1971), On The Asymptotic Optimality of Spectral Analysis for Testing Hypotheses about Time Series , Ann. Math. Stat, vol 42, 1348-1358.

[16] Kazakos, D. & Popantoni-Kazakos, P. (1980), Spectral Distance Measuring Between Gaussian Processes IEEE, Transactions on Automatic Control, AC-25, 950-959.

[17] Markel, J. D. & Gray, A. H., Jr. (1976), Linear Prediction of Speech, Springer, Berlin.

[18] Parzen, E. (1990), Time Series Statistics and Information, IMA.

[19] Piccolo, D. (1990), ADistance Measure for Classifying ARIMA Models, J. of Time Series Analysis, vol 11(21), 153-164.

[20] Pinsker, M. S. (1964), Information and Information Stability of Random Variables and Processes, Holden-Day, San Francisco.

[21] Renyi, A. (1961), On Measure on Entropy and Information, in Proceeding of 4^{th} Berkeley Sumposium on Mathmatical Statistics and Probability. Berkeley: University of California Press, PP. 547-561.

[22] Raji, E. (1986), Evidence for a Peripheral Origin of The Brainstem Auditory Evoked Potentials, Electronyogr. Clin. Neurophysiol, vol 26, 13-16.

[23] Shumway, R. H. & Unger, A. N. (1974), LInear Discrimination Function for Stationary Time Series, JASA, vol 69, 948-956.

[24] Shumway, R. H. (1982). Discriminant Analysis for Time Series, Handbook of Statistics, vol 2, 1-46, eds, Krishnaiah, P. R. & kanal, L. N. Amsterdam, North-Holland.

[25] Shumway, R. H. (1988), Applied Statistical Time Series Analysis, Prentice Hall.

[26] Shumway, R. H. (1996), Statistical Approaches to Seismic Discrimination, in Monitoring a Comprehensive Test Ban Treaty eds. Husebye, E. S. & Dainty, A. M. Doordrecht: Kluwer, PP, 791-803.

[27] Tjostheim, D. (1974), Autoregressive Representation of Seismic P-wave Signals with an Application to The Problem of Short Period Discriminants, Geophys. J. Roy. Astron. Soc. vol 43, 269-291.

[28] Wahba, G. (1968), On The Distribution of Some Statistic Useful in the Analysis of Jointly Stationary Time Series, Ann. Math. Stat, vol 38, 1849-1868.

[29] Whittle, P. (1954), Estimation and Information in Stationary Time Series, Arkiv for Matematik, vol 2, 423-434.

[30] Wolf, J. J. (1976), Speech Recognition and Underestanding Digital Pattern Recognition, K. S. Fu, Ed., Springer -Verlag, Berlin , 167-203.

مدل تاثیر تصادفی با داده‌هایی با پاسخهای آمیخته

علی صادقی[۱]، مجتبی گنجعلی[۲]

[۱] کارشناسی ارشد آمار اقتصادی ـ اجتماعی دانشگاه شهید بهشتی
[۲] عضو هیئت علمی دانشگاه شهید بهشتی

چکیده: در کاربردهای مختلف آمار چند متغیره در شاخه‌های مختلف علوم با مدلهایی تاثیر تصادفی مواجه‌ایم که متغیر پاسخ برداری از داده‌های آمیخته پیوسته و گسسته است. در چنین موارد شناسایی همبستگی میان متغیرهای پیوسته و گسسته اهمیت بسیاری دارد. هنگامی که تمامی متغیرهای پاسخ پیوسته باشند می‌توان از تحلیل عاملی استفاده نمود و هنگامی که کلیه متغیرها گسسته باشند استفاده از روشهای چند متغیره گسسته می‌تواند سودمند باشد. در این مقاله با استفاده از روش تجزیه تک عاملی روشی ارائه گردیده است که می‌توان آمیخته‌ای از متغیرهای پیوسته و گسسته را با به دست آوردن همبستگی‌ها در مدل اثرات تصادفی مورد بررسی قرار داد. در این مدل متغیرهای پیوسته از توزیع نرمال و متغیرهای گسسته از توزیع دودیی تبعیت می‌کنند، شده‌اند. با استفاده از روش تجزیه تک عاملی علاوه بر بدست آوردن همبستگی میان متغیرهای پاسخ، تاثیر عوامل محیطی و ژنتیکی را نیز در نظر می‌گیریم. در این مقاله روشی مبتنی بر استفاده از تاثیرهای تصادفی را برای مدل‌بندی پاسخهای آمیخته نرمال و دودویی معرفی و نحوه تحلیل تاثیر برخی متغیرهای تبیینی بر این پاسخها را نیز شرح می‌دهیم.

واژه‌های کلیدی: اثرات تصادفی، تجزیه عاملی، ،متغیر پنهان، پاسخهای آمیخته

۱ مقدمه

در بسیاری از تحقیقات آماری با حالتی مواجه هستیم که آمیخته‌ای از داده‌های پیوسته و گسسته به عنوان متغیر پاسخ مد نظر است و بعضی از متغیرها دارای اثرات تصادفی هستند این مدلها توسط لیتل[۱] و رابین [۲] (۲۰۰۲) بسیار مورد توجه قرارگرفته است. در چنین حالتی بر خلاف هنگامی که تمامی اثرات ثابت هستند نمی‌توان از روشهای معمولی مورد استفاده در علم آمار مدل مناسب را به داده‌ها برازش داد [اگرستی[۳] (۲۰۰۲)]. روشهایی که تاکنون برای برازش مدل‌های با اثرات تصادفی برای داده‌های آمیخته مورد بررسی قرار گرفته‌اند به دلیل استفاده از روشهای عددی در برازش مدل مناسب از تئوری پیچیده‌ای استفاده کرده‌اند و بسیاری از آنها تنها در شرایط بخصوصی به نتیجه مورد نظر می‌رسند. در این مقاله سعی گردیده است که با استفاده از تئوری ساده و با استفاده از روشهای عددی ساده‌تر این بحث مورد توجه قرار گیرد.

1) Little 2) Rubin 3) Agresti

۲ مدل اثرات تصادفی برای پاسخهای آمیخته

مانند مدل مورد استفاده سامل و ریان[4] (۱۹۹۶) یک مدل اثرات تصادفی را با M برآمد مشاهده شده که تعداد M_1 تا از این برآمدها دارای توزیع دودویی و $M_2 = M - M_1$ تا از برآمدهای باقی مانده از توزیع نرمال پیروی می‌کنند در نظر می‌گیریم. ذکر این نکته ضروری است که متغیرهای تصادفی در مدل مورد بحث مانند اثرات متغیرهای پنهان رفتار می‌کنند [بکر[5] (۱۹۹۲)]. برای این منظور مدل اثرات تصادفی را برای آزمودنی iـام به صورت زیر در نظر می‌گیریم:

$$\begin{aligned} y^*_{im} &= X'_{im}\beta + \varepsilon_{im} \qquad m = ۱, \ldots, M_۱ \\ z_{im} &= U'_{im}\beta + \varepsilon_{im} \qquad m = M_۱ + ۱, \ldots, M \end{aligned} \tag{۱}$$

که در آن X_{im} و U_{im} ماتریس متغیرهای تبیینی برای آزمودنی iـام در برآمد mـام است. در این مدل y_{im} مقدار مشاهده شده یک متغیر پنهان y^*_{im} با توزیع نرمال استاندارد است که فقط کوچکتر یا بزرگتر از عدد ۰ بودن آنرا در دست داریم و اگر $y^*_{im} > ۰$ آنگاه $y_{im} = ۱$ و در غیر این صورت $y_{im} = ۰$ خواهد بود.
ماتریس واریانس کواریانس مدل فوق دارای پارامترهای زیادی حتی با در نظرگرفتن مقادیر کوچک M خواهد بود. برای درک بهتر این مسئله حالتی را در نظر می‌گیریم که $M_1 = ۴$ و $M_2 = ۲$ است . در این صورت مدل به صورت زیر بیان می‌شود.

$$\begin{aligned} y^*_{im} &= X'_{im}\beta + \varepsilon_{im} \qquad m = ۱, ۲, ۳, ۴ \\ z_{im} &= U'_{im}\beta + \varepsilon_{im} \qquad m = ۵, ۶ \end{aligned} \tag{۲}$$

در این صورت ساختار ماتریس واریانس کواریانس غیرساختاری به صورت زیر در خواهد آمد

$$\Sigma = \begin{pmatrix} ۱ & \rho_{۱۲} & \rho_{۱۳} & \rho_{۱۴} & \sigma_۱\rho_{۱۵} & \sigma_۲\rho_{۱۶} \\ \rho_{۱۲} & ۱ & \rho_{۲۳} & \rho_{۲۴} & \sigma_۱\rho_{۲۵} & \sigma_۲\rho_{۲۶} \\ \rho_{۱۳} & \rho_{۲۳} & ۱ & \rho_{۳۴} & \sigma_۱\rho_{۳۵} & \sigma_۲\rho_{۳۶} \\ \rho_{۱۴} & \rho_{۲۴} & \rho_{۳۴} & ۱ & \sigma_۱\rho_{۴۵} & \sigma_۲\rho_{۴۶} \\ \sigma_۱\rho_{۱۵} & \sigma_۱\rho_{۲۵} & \sigma_۱\rho_{۳۵} & \sigma_۱\rho_{۴۵} & \sigma^۲_۱ & \sigma_۱\sigma_۲\rho_{۵۶} \\ \sigma_۲\rho_{۱۶} & \sigma_۲\rho_{۲۶} & \sigma_۲\rho_{۳۶} & \sigma_۲\rho_{۴۶} & \sigma_۱\sigma_۲\rho_{۵۶} & \sigma^۲_۲ \end{pmatrix}$$

با وجود اثرات تصادفی در مدل نمی‌توان از روشهای معمول در آمار به منظور برازش مدل با وجود آمیخته‌ای از برآمدهای پیوسته و گسسته استفاده کرد. برای بدست آوردن پارامترها از روشهای عددی استفاده می‌گردد اما یکی از مشکلات اصلی استفاده از روشهای عددی آن است که هنگاهی که تعداد کمی داده در اختیار داریم و مدل دارای تعداد زیادی برآمد است نمی‌توانند به خوبی عمل کنند. برای مثال هنگاهی که ۴ پاسخ گسسته و ۲ پاسخ پیوسته مانند مدل بالا در

4) Sammel and Ryan 5) Baker

دست باشد احتیاج به برآورد ۱۷ پارامتر در ماتریس کواریانس خواهیم داشت و تعداد کم داده‌ها می‌تواند مشکل ساز گردد و دقت مورد نظر در برآورد پارامترها حاصل نخواهد شد. اولین بار سامل، لوئیز[۶] و ریان (۱۹۹۷) این مدل را مورد بررسی قرار دادند اما روش مورد استفاده آنها از تئوری عددی پیچیده‌ای برای بدست آوردن پارامترهای مدل استفاده می‌کند و دیگر اشکال آن حساسیت بسیار زیاد مدل در برابر داده‌های کوچک است. یکی از راههای برخورد با چنین مشکلی کاهش تعداد پارامترهای مدل است که برای اولین بار در این مقاله برای چنین داده‌هایی مورد بررسی قرار گرفته است. برای این منظور استفاده از روش تجزیه تک عاملی که اولین بار توسط هکمن (۱۹۸۱) ارائه شده است می‌تواند بسیار سودمند باشد.

۳ مدل اثرات تصادفی با استفاده از تجزیه تک عاملی

همانگونه که بیان شد، روش تک عاملی می‌تواند در کاهش تعداد پارامترهای مدل مناسب باشد. برای توضیح این روش ابتدا جملات خطا ε_{im} را برای $m = ۱, \ldots, M$ به دو جزء با اثرات تصادفی (b_{im}) و (v_{im}) به صورت زیر تقسیم می‌کنیم:

$$\varepsilon_{im} = b_{im} + v_{im} \qquad m = ۱, \ldots, M$$

در این صورت مدل (۱) با بکار بردن روش تجزیه تک عاملی که توسط هکمن[۷] (۱۹۸۱) برای اثرات تصادفی مورد استفاده قرارگرفته است، به صورت زیر در خواهد آمد

$$\begin{aligned} y^*_{im} &= X'_{im}\beta_m + b_{im} + v_{im} \qquad m = ۱, \ldots, M_۱ \\ z_{im} &= U'_{im}\beta_m + b_{im} + v_{im} \qquad m = M_۱ + ۱, \ldots, M \end{aligned} \qquad (۳)$$

آنگاه برای قابل تشخیص بودن مدل خواهیم داشت:

$$var(\varepsilon_{im}) = var(b_{im} + v_{im}) = \sigma^۲_{b_m} + \sigma^۲_{v_{im}} \qquad m = ۱, \ldots, M$$

که در آن

$$var(\varepsilon_{im}) = var(b_{im} + v_{im}) = \sigma^۲_{b_m} + \sigma^۲_{v_{im}} = ۱ \qquad m = ۱, \ldots, M_۱$$

این نوع تعریف برای اثرات تصادفی، یک ساختار برای ماتریس واریانس، کواریانس همانند Σ ایجاد می‌کند.
می‌توان تعداد اثرات تصادفی را با لحاظ محدودیت روی ساختارکواریانس توسط تجزیه تک عاملی

6) Louis 7) Heckman

که توسط هکمن (۱۹۸۱) بیان شده است کاهش داد به این معنی که فرض کنیم $b_{im} = \lambda_m b_i$ برای $m = ۱, \ldots, M$ که برای برآمدهای دودویی $-۱ < \lambda_m < ۱$ و برای برآمدهای نرمال $-\infty < \lambda_m < +\infty$ خواهد بود که در آن b_i دارای توزیع نرمال استاندارد است و λ_m پارامترهای مقیاس هستند. در این حالت ساختار همبستگی به صورت زیر است:

$$cov(y_m, y_{m'}) = \lambda_m \lambda_{m'} \qquad m, m' = ۱, \ldots, M_۱ \qquad m \neq m'$$

و

$$cov(z_m, z_{m'}) = \lambda_m \lambda_{m'} \qquad m, m' = M_۱, \ldots, M_۲ \qquad m \neq m'$$

و همچنین

$$cov(z_{m'}, y_m) = \lambda_m \lambda_{m'} \qquad m = ۱, \ldots, M_۱ \ , \quad m' = M_۱, \ldots, M_۲$$

در این صورت تعداد پارامترهای مدل و ماتریس واریانس کواریانس به میزان قابل توجهی کاهش می‌یابد. برای مثال با $M = ۶$ و $M_۱ = ۴$ و $M_۲ = ۲$ ساختار ماتریس واریانس کواریانس به صورت زیر در خواهد آمد:

$$\Sigma = \begin{pmatrix} ۱ & \lambda_۱\lambda_۲ & \lambda_۱\lambda_۳ & \lambda_۱\lambda_۴ & \lambda_۱\lambda_۵ & \lambda_۱\lambda_۶ \\ \lambda_۱\lambda_۲ & ۱ & \lambda_۳\lambda_۲ & \lambda_۲\lambda_۴ & \lambda_۲\lambda_۵ & \lambda_۶\lambda_۲ \\ \lambda_۱\lambda_۳ & \lambda_۲\lambda_۳ & ۱ & \lambda_۳\lambda_۴ & \lambda_۳\lambda_۵ & \lambda_۳\lambda_۶ \\ \lambda_۱\lambda_۴ & \lambda_۲\lambda_۴ & \lambda_۳\lambda_۴ & ۱ & \lambda_۵\lambda_۴ & \lambda_۶\lambda_۴ \\ \lambda_۱\lambda_۵ & \lambda_۲\lambda_۵ & \lambda_۳\lambda_۵ & \lambda_۴\lambda_۵ & \sigma_۵^۲ + \lambda_۵^۲ & \lambda_۵\lambda_۶ \\ \lambda_۱\lambda_۶ & \lambda_۲\lambda_۶ & \lambda_۳\lambda_۶ & \lambda_۴\lambda_۶ & \lambda_۵\lambda_۶ & \sigma_۶^۲ + \lambda_۶^۲ \end{pmatrix}$$

با استفاده از روش تک عاملی برای $M = ۶$ تعداد پارامترهای مدل از ۱۷ پارامتر به ۸ پارامتر کاهش می‌یابد.

۴ مثال کاربردی

لیتل و شلوچتر[8] (۱۹۸۵) تحقیق روانشناسی بر روی نمونه‌ای از ۶۹ خانوار انجام دادند. هدف از این تحقیق شناسایی تاثیر ناهنجاری‌های والدین بر روی ناهنجاریهای فرزندان خانوار است. از آنجا که نمی‌توان اثر شدت ناهنجاری والدین بر روی فرزندان را مشاهده کرد، از مدل اثرات تصادفی در برازش داده‌های فوق استفاده شده است. در مدل فوق از ۳ برآمد که هر یک به نوعی نشان دهنده (انعکاس دهنده) ناهنجاریهای فرزندان خانوار هستند استفاده شده‌اند هدف از این تحقیق آن است که بدانیم شدت ناهنجاریهای والدین چه تاثیری بر روی هر یک از برآمدها که برآمدهای مشاهده شده از ناهنجاریهای فرزندان هستند دارد. در اینجا یک برآمد دودویی و دو برآمد باقی مانده دارای توزیع نرمال هستند استفاده شده است. که عبارتند از:

8) Little and Schluchter

جدول ۱: خصوصیات توصیفی برآمدها

متغیر	Low Risk		H/M Risk		Total	
	میانگین	انحراف معیار	میانگین	انحراف معیار	میانگین	انحراف معیار
y	۰٫۳۷۵	۰٫۵۱۷	۰٫۷۷۷	۰٫۴۲۷	۰٫۶۵۳	۰٫۴۸۵
z_1	۱۱۰٫۸۷۵	۱۵٫۹۷۷	۹۴٫۷۷۷	۱۱٫۸۱۵	۱۰۳٫۱۷۳	۱۵٫۳۶۲
z_2	۱۳۶٫۷۵	۳۰٫۳۴۹	۹۶٫۵۵۵	۳۱٫۳۲۴	۱۱۵٫۲۱۷	۳۲٫۷۶۴

متغیر y_1 که دارای توزیع دودویی است و مقدا ۱ را می‌پذیرد اگر فرزند خانواده دارای علائم ناهنجاری در رفتار باشد و مقدار ۰ را می‌پذیرد اگر فرزند خانواده دارای علائم ناهنجاری در رفتار نباشد. توجه به این نکته ضروری به نظر می‌رسد که این برآمد خود برآمد مشاهده شده از یک متغیر پنهان y_1^* با توزیع نرمال است.

متغیر z_1 که دارای توزیع نرمال است و نشان دهنده نمرات کسب شده فرزند خانوار در آزمون درک مطلب است.

متغیر z_2 که دارای توزیع نرمال بوده و نشان دهنده نمرات کسب شده فرزند خانوار در یک آزمون قدرت محاوره است.

در این مدل از یک متغیر کمکی X که یک متغیر دودویی است استفاده می‌کنیم. این متغیر کمکی مقدار ۱ را می‌پذیرد هنگامی که والدین دارای ناهنجاریهای روانی اندکی باشند (LR) و ۰ را می‌پذیرد هنگامی که حداقل یکی از والدین دارای سابقه بیماری روانی و یا ناهنجاری روانی در حد متوسط یا زیاد (H/MR) باشند. تحقیقات افراد دیگر بر روی داده‌های فوق نشان داده است که تاثیر دو گروه با ناهنجاری متوسط و زیاد یکسان است به همین دلیل این دو گروه را در این مقاله با هم ترکیب کرده‌ایم.(گنجعلی[9])

۵ تحلیل توصیفی داده‌ها

جدول ۱ خصوصیات توصیفی برآمدها را بیان می‌کند. همانگونه که در جدول ۱ اطلاعات توصیفی مربوط به برآمدها مشخص است، نمره درک مطلب در بین کودکانی که والدین آنها در معرض کمتر ناهنجاری روانی (LR) هستند از کودکانی که والدین آنها در معرض ناهنجاری متوسط و شدید روانی (H/MR) هستند بیشتر است. همچنین برای نمره آزمون قدرت محاوره، چنانکه در جدول اطلاعات توصیفی برآمدها مشخص است. نمره کسب شده توسط کودکان در گروه ناهنجاری کمتر (LR) بیشتر از گروه (H/MR) است.

این بدان معنی است که در صورت معنی‌داری، نمره در بین کودکانی که والدین آنها در معرض ناهنجاری متوسط و شدید روانی (H/MR) هستند نسبت به کودکانی که در معرض کمتر ناهنجاری والدین هستند (LR)، کمتر است و میزان تاثیر ناهنجاری والدین در این دو گروه متفاوت خواهد بود.

9) Gangali

جدول ۲: نتایج بدون در نظرگرفتن همبستگی‌ها

پارامتر	برآورد	سطح معنی‌داری
$\beta_{0\,1}$	۰/۷۶۵	--
$\beta_{1\,1}$	-۱/۰۸۳	۰/۳۵۳
$\beta_{0\,2}$	۹۴/۷۷۸	--
$\beta_{1\,2}$	۱۶/۱	۰/۰۰۸
$\beta_{0\,3}$	۹۶/۵۵۶	--
$\beta_{1\,3}$	۴۰/۱۹۴	۰/۰۰۶

۶ تجزیه و تحلیل مدل اثرات تصادفی

مدل اثرات تصادفی مطابق مدل (۱) به صورت زیر در خواهد آمد:

$$\begin{aligned} y^*_{i1} &= \beta_{0m} + \beta_{1m}X + \varepsilon_{im} \qquad m = 1 \\ z_{im} &= \beta_{0m} + \beta_{1m}X + \varepsilon_{im} \qquad m = 2,3 \end{aligned} \tag{۴}$$

در این مدل هنوز همبستگی میان متغیرها (λ ها) را در نظر نگرفته‌ایم. برآورد پارامترهای مدل در جدول ۲ داده شده است.

چنانچه در جدول ۲ مشخص است میان ناهنجاری والدین و ناهنجاری فرزندان رابطه معنی داری وجود ندارد و شدت ناهنجاری والدین تاثیری در ناهنجاری در بین فرزندان در سطح $\alpha = 0/05$ ندارد اما نمره درک مطلب و نمره آزمون محاوره در بین کودکانی که والدین آنها دارای ناهنجاری کمتری (LR) هستند بیشتر از فرزندانی است که والدین آنها دارای ناهنجاری متوسط یا زیاد (H/MR) هستند. حال همبستگی میان دو متغیر پیوسته را نیز با واردکردن اثرات تصادفی و استفاده از روش تک عاملی وارد مدل می‌کنیم در این صورت مدل به صورت زیر در خواهد آمد:

$$\begin{aligned} y^*_{i1} &= \beta_{0m} + \beta_{1m}X + v_{i1} \qquad m = 1 \\ z_{im} &= \beta_{0m} + \beta_{1m}X + \lambda_m b_i + v_{im} \qquad m = 2,3 \end{aligned} \tag{۵}$$

برآورد پارامترهای مدل در جدول ۳ داده شده است.

چنانچه در جدول ۳ نیز ملاحظه می‌گردد همان نتایج مدل قبل نیز در این مدل بدست می‌آید علاوه بر آن می‌توان مقدار همبستگی میان برآمدها را نیز بدست آورد. حال حالت کلی را یعنی با وارد کردن اثر تصادفی برای برآمد دودویی بیان می‌کنیم در حقیقت با در نظرگرفتن اثر تصادفی همبستگی میان متغیرگسسته دودویی و متغیرهای پیوسته را نیز در نظر می‌گیریم. در این صورت مدل نهایی به صورت زیر در خواهد آمد:

جدول ۳: نتایج تنها با در نظرگرفتن همبستگی میان متغیرهای پیوسته

پارامتر	برآورد	سطح معنی داری
$\beta_{\circ 1}$	۰/۷۶۵	--
β_{11}	-۱/۰۸۳	۰/۳۷۱
$\beta_{\circ 2}$	۹۳/۷۱۸	--
β_{12}	۱۴/۱	۰/۰۰۵
λ_2	۱۳/۴۷۹	۰/۰۰۰
$\beta_{\circ 3}$	۹۴/۹۳۷	--
β_{13}	۳۷/۱۳۷	۰/۰۰۶
λ_3	۴/۹۱۲	۰/۰۰۰
σ_{v_2}	۳/۰۷	--
σ_{v_3}	۲۳/۱۹۷	--

جدول ۴: نتایج کلی مدل

پارامتر	برآورد	سطح معنی داری
$\beta_{\circ 1}$	۰/۵۰۳	--
β_{11}	-۱/۹۴	۰/۰۰۵
λ_1	۰/۹۹	۰/۰۰۰
$\beta_{\circ 2}$	۹۳/۳۹۴	--
β_{12}	۱۴/۳۹۴	۰/۰۰۰
λ_2	۱۳/۸۰۳	۰/۰۰۰
$\beta_{\circ 3}$	۹۴/۴۹۷	--
β_{13}	۳۷/۶۶۶	۰/۰۰۰
λ_3	۲۰/۴۹۹	۰/۰۰۰
σ_{v_2}	۲/۹۰۷	--
σ_{v_3}	۲۳/۵۶۵	--

$$\begin{aligned} y^*_{i1} &= \beta_{\circ m} + \beta_{1m}X + \lambda_1 b_i + v_{i1} \qquad m = 1 \qquad (۶) \\ z_{im} &= \beta_{\circ m} + \beta_{1m}X + \lambda_m b_i + v_{im} \qquad m = 2,3 \end{aligned}$$

برآورد پارامترهای این مدل در جدول ۴ داده شده است.

این برآوردها با استفاده از برنامه NAG (۱۹۹۶) و زیر روال $E04UCF$ بدست آمده‌اند. چنانچه از سطح معنی‌داری پارامترهای برآورد شده نیز مشخص است تمامی پارامترها در سطح $\alpha = 0/05$ معنی‌دار هستند و این بدان معنی است که شدت ناهنجاری والدین بر روی ناهنجاری فرزندان تاثیر دارد و هر چه والدین دارای ناهنجاری بیشتری باشند خطر وجود ناهنجاری در فرزندان نیز بیشتر است همچنین شدت ناهنجاری در والدین در درک مطلب و قدرت محاوره فرزندان تاثیر گذار است به این صورت که هر چه والدین دارای ناهنجاری روانی بیشتری باشند قدرت درک

مطلب و محاوره فرزندان کاهش می‌یابد. با توجه به برآورد پارامترها ماتریس واریانس کواریانس به صورت زیر در می‌آید:

$$\Sigma = \begin{pmatrix} ۱ & ۱۳/۵۲۶ & ۲۰/۲۷۳ \\ ۱۳/۵۲۶ & ۱۹۸/۹۷۳ & ۲۸۲/۹۴۷ \\ ۲۰/۲۷۳ & ۲۸۲/۹۴۷ & ۹۷۵/۵۱۸ \end{pmatrix}$$

حال با توجه به ماتریس واریانس کواریانس بالا کواریانس میان برآمدهای گسسته دودویی و پیوسته نرمال با وجود اثرات تصادفی در اختیار است.

۷ نتیجه‌گیری و پیشنهادات

چنانکه در جدول ۱ مشاهده می‌گردد هنگامی که همبستگی میان متغیرها را در مدل در نظر نگرفته‌ایم، میان ناهنجاری والدین و ناهنجاری فرزندان رابطه معنی‌داری وجود ندارد. اما هنگامی که در مدل نهایی همبستگی میان متغیرها را وارد مدل می‌کنیم مدل دقیقتر عمل نموده و تاثیر ناهنجاری والدین بر روی ناهنجاری فرزندان مشخص می‌گردد. نتایج حاصل شده از این مدل دقیقتر از مدل قبل می‌باشد.

در مدل اثرات تصادفی برای داده‌های آمیخته ما تغییرات واریانس برآمدها را درون دو گروه با ناهنجاری کمتر والدین و ناهنجاری متوسط یا بیشتر والدین ثابت در نظر گرفتیم. پیشنهاد می‌گردد هنگاهی که تغییرات واریانس معنی‌دار است اثرات تغییرات واریانس را نیز وارد مدل کنیم به طور حتم این امر باعث برازش بهتر مدل خواهد شد.

مراجع

[1] Agresti, A. (2002). Categorical data Analysis. New York: Wiley.

[2] Baker, F. B. (1992). Item response Theory: Parameter Estimation Techniques. New York: Dekker.

[3] Gangali. M. (2003). A model for Mixed Continuous and Discrete Response with possible of Missing Response . Journal of Sciences, Islamic Republic of Iran, 14(1) , 53-60.

[4] Heckman, J. (1981). Statistical models for discrete panel data in Manski, C.& Mc Fadden D., strutural analysis of discrete data with econometric applications. 114-195, cambridge, mass: MIT press

[5] Little, R. J. and Schelucher M. (1985). Maximum liklihood estimation for mixed continuous and categorical data with missing value. Biometrika, 72(3) , 497-512.

[6] Little, R. J. A. and Rubin, D. B. (2002). Statistical analysis with Missing Data. New York: Wiley.

[7] NAG, (1996). Numerical Algoritms Group Manual, Mark 16, Oxford, UK.

[8] Sammel, M. D.and Ryan, L. M.(1996). Latent variable models with fixed efects. Biometrics, 52, 220-243.

[9] Sammel, M. D.and Louise. M. and Ryan, L. M.(1997) Latent variable models for Mixed Discrete and Continuous Outcomes. J. R. Statist. Soc. B, 59, No 3, 667-678.

آزمون تصادفی شده برای میانگین توزیع نرمال بر پایه‌ی داده‌های نادقیق

سید محمود طاهری، تکتم بزرگوار

دانشکده علوم ریاضی، دانشگاه صنعتی اصفهان

چکیده: در این مقاله یک روش آزمون تصادفی شده برای انجام آزمون‌های مربوط به میانگین توزیع نرمال بر پایه‌ی داده‌های فازی مطرح، تشریح و بررسی می‌شود. در این روش تابع عضویت آماره‌ی آزمون بر اساس مقادیر مختلف α برش‌ها ساخته می‌شود. بر پایه‌ی تابع عضویت این آماره، احتمال رد فرض صفر محاسبه می‌شود. نشان داده می‌شود که در حالت خاص، این روش به روش آزمون فرض کلاسیک تبدیل می‌شود. شیوه‌ی فوق با چند مثال عددی مبتنی بر داده‌های واقعی، که به صورت اعداد فازی مثلثی بیان شده‌اند، تشریح می‌گردد.

واژه‌های کلیدی: داده‌های نادقیق، آزمون تصادفی شده، عدد فازی ذوزنقه‌ای

۱ مقدمه و تاریخچه

آزمون یک فرض آماری عبارت از به کارگرفتن مجموعه‌ی قواعدی است برای آن که تصمیم بگیریم که آیا یک فرض را بپذیریم یا آن را به نفع یک فرض مقابل رد کنیم. مثلاً فرض کنید که آماردانی می‌خواهد فرض صفر $\theta = \theta_0$ را در برابر فرض مقابل $\theta = \theta_1$ آزمون کند. یک نمونه تصادفی از جامعه موردنظر انتخاب می‌کند و بر اساس یک تابع آزمون تصمیم به رد یا پذیرش فرض صفر می‌گیرد. در آزمون‌های غیر تصادفی هر تابع آزمون فضای نمونه‌ای را به دو مجموعه، یک ناحیه قبول و یک ناحیه رد (بحرانی)، افراز می‌کند.

یکی از مفروضات مهم در آزمون‌های آماری آن است که داده‌های مشاهده شده دقیق هستند. اما گاهی اوقات با مشاهدات نادقیق و مبهم روبرو هستیم. در این موارد نمی‌توان از روش‌های متداول آماری استفاده کرد. یکی از روش‌های جانشین در این موارد به کارگیری مجموعه‌های فازی در فرمول‌بندی داده‌های نادقیق است. بدین ترتیب با مسئله آزمون فرض‌های آماری بر پایه‌ی داده‌های فازی روبرو هستیم.

آزمون فرض بر پایه‌ی داده‌های فازی نخستین بار توسط کازالس[1] و همکاران [3,2] در سال ۱۹۸۶ مورد مطالعه قرار گرفت. آنها لم نیمن-پیرسن را برای حالت بالا تعمیم دادند و رهیافت بیزی را به این مسئله بررسی کردند [4]. سون[2] و همکاران [10] نیز بر پایه‌ی تعمیم لم نیمن-پیرسن،

1) Casals 2) Son

مفهوم تواناترین آزمون فازی را تعریف و یک کاربرد از آن را بررسی کردند. زگورزسکی[3] [6] رهیافتی را برای موضوع بالا پیشنهاد داده است که در آن، تصمیم درباره‌ی رد یا قبول فرض‌ها، به صورت فازی بیان می‌شود. طاهری و بهبودیان [11] مسئله آزمون فرض‌های فازی را بر پایه‌ی داده‌های فازی با رویکرد بیزی مطالعه نموده‌اند. رهیافت‌های دیگری نیز توسط رامر[4] و کاندل[5] [9] و واتانابه[6] [12] به مسئله‌ی آزمون فرض با داده‌های فازی ارائه شده است. همچنین مونتنگرو[7] و همکاران [8] آزمون مقایسه‌ی میانگین‌های دو جامعه را بر اساس داده‌های فازی مطالعه نموده‌اند. در این مقاله آزمون میانگین یک توزیع نرمال بر اساس داده‌های نادقیق بر پایه‌ی رهیافت لیو و شن[8] [7] مورد مطالعه قرار می‌گیرد. نخست به معرفی مفاهیم مقدماتی مورد نیاز در این مقاله پرداخته و سپس در بخش‌های سوم و چهارم به بیان مسئله و تشریح شیوه کار می‌پردازیم. آزمون میانگین جامعه را در حالتی که واریانس جامعه معلوم و مجهول است، به ترتیب در بخش‌های پنجم و ششم مورد بررسی قرار می‌دهیم.

۲ مفاهیم مقدماتی

در این بخش فرض می‌شود که خواننده آشنایی اولیه با مجموعه‌های فازی دارد. در ادامه به چند مفهوم و نتیجه اساسی که در این مقاله مستقیماً به کار گرفته می‌شوند، می‌پردازیم [14,1].

تعریف ۱.۲ یک مجموعه فازی نرمال، محدب و تک نمائی از $\mathbb{R}$ را که تابع عضویت آن قطعه به قطعه پیوسته باشد، عدد فازی گوییم. مجموعه تمام اعداد فازی از $\mathbb{R}$ را با $F(\mathbb{R})$ نشان می‌دهیم.

تعریف ۲.۲ عدد فازی $\tilde{N}$ را یک عدد فازی LR گوییم هرگاه تابع عضویت آن به شکل زیر باشد

$$\tilde{N}(x) = \begin{cases} L(\frac{m-x}{\alpha}) & x \leq m \\ R(\frac{x-m}{\beta}) & x > m \end{cases}$$

که در آن $L, R : \mathbb{R}^+ \to [0, 1]$ را توابع مرجع، m را مقدار میانه (نما) و α, β (اعداد حقیقی مثبت) را پهنای چپ و راست $\tilde{N}$ نامیم. $\tilde{N}$ را با نماد $\tilde{N} = (m, \alpha, \beta)$ نشان می‌دهیم. چنانچه برای عدد فازی $\tilde{N} = (m, \alpha, \beta)_{LR}$ داشته باشیم

$$L(x) = R(x) = \begin{cases} 1 - x & 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

آنگاه $\tilde{N}$ را عدد فازی مثلثی گوییم و با نماد $(m, \alpha, \beta)_T$ نشان می‌دهیم.

تعریف ۳.۲ عدد فازی $\tilde{M}$ را یک عدد فازی ذوزنقه‌ای LR گوییم هرگاه تابع عضویت آن به

3) Grzegorzewski 4) Romer 5) Kandel 6) Watanabe 7) Montenegro
8) Leu and Chen

شکل زیر باشد

$$\tilde{M}(x) = \begin{cases} L(\frac{m_1 - x}{\alpha}) & m_1 > x \\ 1 & m_1 \leq x \leq m_2 \\ R(\frac{x - m_2}{\beta}) & m_2 < x \end{cases}$$

که در آن Lو R توابعی هستند که در تعریف عدد فازی LR صدق می‌کنند. در این حالت $\tilde{M}$ با نماد $\tilde{M} = (m_1, m_2, \alpha, \beta)_{LR}$ نشان داده می‌شود. واضح است که عدد فازی مثلثی حالت خاصی از عدد فازی ذوزنقه‌ای است.

تعریف ۴.۲ مجموعه عناصری از X را که درجه عضویت آنها در مجموعه فازی $\tilde{A}$ حداقل به بزرگی α ($\alpha > 0$) باشد، α ـ برش (مجموعه تراز) $\tilde{A}$ گوئیم و با $\tilde{A}_\alpha$ نشان می‌دهیم، یعنی

$$\tilde{A}_\alpha = \{x \in X; \tilde{A}(x) \geq \alpha\}$$

تعریف ۵.۲ فرض کنید $(\mathbb{R}^n, F, P)$ یک فضای احتمال باشد. یک پیشامد فازی در $\mathbb{R}^n$ عبارتست از یک مجموعه فازی A از $\mathbb{R}^n$ که تابع عضویت آن اندازه پذیر بورل باشد.

تعریف ۶.۲ اگر $\tilde{A}$ یک پیشامد فازی باشد، احتمال آن به صورت زیر تعریف می‌شود [13]

$$P(\tilde{A}) = \int_{\mathbb{R}^n} \tilde{A}(x) dP = E(\tilde{A}(x))$$

که در حالت گسسته علامت انتگرال به علامت سیگما تبدیل می شود.

عملگرهای جبری بر اعداد فازی بر پایه اصل گسترش و به صورت زیر تعریف می‌شوند.

تعریف ۷.۲ اگر $\lambda \in \mathbb{R} - \{0\}$، آنگاه ضرب اسکالر λ در عدد فازی $\tilde{M}$ به صورت یک عدد فازی با تابع عضویت زیر تعریف می‌شود

$$\lambda \tilde{M}(x) = \tilde{M}(\frac{x}{\lambda})$$

تعریف ۸.۲ فرض کنید $\tilde{N}, \tilde{M} \in F(\mathbb{R})$ با توابع عضویت پیوسته باشند. حاصل $\tilde{M} \oplus \tilde{N}$ و $\tilde{M} \ominus \tilde{N}$ به صورت مجموعه‌های فازی با توابع عضویت زیر تعریف می‌شوند

$$(\tilde{M} \oplus \tilde{N})(z) = \sup_{z=x+y} \min[\tilde{M}(x), \tilde{N}(y)]$$

$$(\tilde{M} \ominus \tilde{N})(z) = \sup_{z=x-y} \min[\tilde{M}(x), \tilde{N}(y)]$$

قضیه ۱.۲ اگر $\tilde{M} = (m, \alpha, \beta)_{LR}$ و $\tilde{N} = (n, \delta, \gamma)_{LR}$ دو عدد فازی LR باشند و $\lambda \in \mathbb{R}$ یک اسکالر باشد، آنگاه

$$a)\lambda \otimes (m,\alpha,\beta)_{LR} = (\lambda m, \lambda\alpha, \lambda\beta)_{LR} \qquad \lambda > 0$$
$$b)\lambda \otimes (m,\alpha,\beta)_{LR} = (\lambda m, -\lambda\beta, -\lambda\alpha)_{RL} \qquad \lambda < 0$$
$$c)\tilde{M} \oplus \tilde{N} = (m+n, \alpha+\delta, \beta+\gamma)_{LR}$$
$$d)\tilde{M} \ominus \tilde{N} = (m-n, \alpha+\delta, \beta+\gamma)_{RL}$$

نکته ۱.۲ اگر $\tilde{M} = (m,\alpha,\beta)$ و $\tilde{N} = (n,\delta,\gamma)$ دو عدد فازی LR باشند، آنگاه ممکن است که ضرب یا تقسیم آن‌ها، یک عدد فازی LR نباشد.

۳ بیان مساله

فرض کنید $X_1, X_2, \ldots, X_n$ یک نمونه تصادفی از جامعه‌ای با توزیع نرمال با میانگین μ و واریانس σ^2 باشد. در حالتی که مقادیر X_iها به صورت اعداد دقیق باشند با شیوه‌های متداول برای آزمون فرض درباره‌ی میانگین جامعه آشنا هستیم. به طور متداول آزمون فرض درباره‌ی میانگین جامعه در حالتی که واریانس جامعه معلوم است، بر پایه‌ی آماره $Z = \frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ و در حالتی که واریانس جامعه مجهول است بر پایه‌ی آماره $Z = \frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}}$ انجام می‌شود [5].

اکنون فرض کنید که مشاهدات مربوط به X_iها دقیق نباشند (یا دقیق گزارش نشده باشند) بلکه به صورت اعداد فازی باشند. می‌خواهیم شیوه‌های متداول برای آزمون فرض درباره‌ی μ را به حالتی که گفته شد، به صورت یک آزمون تصادفی شده، تعمیم دهیم. برای سادگی محاسبات فرض می‌کنیم مشاهدات به صورت اعداد فازی ذوزنقه‌ای هستند. نخست حالتی را که σ معلوم است در نظر می‌گیریم، آنگاه به حالتی می پردازیم که σ مجهول است.

۴ تشریح شیوه‌ی کار

۱.۴ حالت σ معلوم

در این حالت آماره‌ی $Z = \frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ به شکل زیر تبدیل می‌شود

$$\tilde{Z} \approx \frac{\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i - \mu_0}{\frac{\sigma}{\sqrt{n}}} \qquad (1)$$

چنانچه مشاهدات اعداد فازی ذوزنقه‌ای باشند، با توجه به اصل گسترش، $\tilde{Z}$ یک عدد فازی ذوزنقه‌ای می‌باشد [14,7]. نخست تابع عضویت آماره فوق را بر اساس αـبرش‌های $\tilde{X}_i$ به

دست می‌آوریم، و سپس از آن برای انجام آزمون استفاده می‌کنیم. α-برش‌های $\tilde{X}_i$ را به صورت $(\tilde{X}_i)_\alpha = [(\tilde{X}_i)^L_\alpha, (\tilde{X}_i)^U_\alpha]$ نشان می‌دهیم که در آن

$$(\tilde{X}_i)^L_\alpha = [\min_{x_i}\{x_i \in X \quad s.t \quad \mu_{\tilde{X}_i}(x_i) \geq \alpha\}]$$

و

$$(\tilde{X}_i)^U_\alpha = [\max_{x_i}\{x_i \in X \quad s.t \quad \mu_{\tilde{X}_i}(x_i) \geq \alpha\}]$$

فرض کنید

$$(\tilde{Z})^L_\alpha = \frac{(\bar{\tilde{X}})^L_\alpha - \mu_\circ}{\frac{\sigma}{\sqrt{n}}} \quad , (\tilde{Z})^U_\alpha = \frac{(\bar{\tilde{X}})^U_\alpha - \mu_\circ}{\frac{\sigma}{\sqrt{n}}} \qquad (۲)$$

که در آن $(\bar{\tilde{X}})^L_\alpha = \frac{1}{n}\sum_{i=1}^n (\tilde{X}_i)^L_\alpha$ و $(\bar{\tilde{X}})^U_\alpha = \frac{1}{n}\sum_{i=1}^n (\tilde{X}_i)^U_\alpha$. بنابراین α-برش‌های $\tilde{Z}$ به صورت $\tilde{Z}_\alpha = [(\tilde{Z})^L_\alpha, (\tilde{Z})^U_\alpha]$ است، که $(\tilde{Z})^L_\alpha$ و $(\tilde{Z})^U_\alpha$ توابع خطی بر حسب α-برش‌های $\bar{\tilde{X}}$ هستند. بنابراین تابع عضویت $\tilde{Z}$ عبارت است از

$$\mu_{\tilde{Z}}(z) = \begin{cases} L(z) & \tilde{Z}^L_\circ \leq z \leq (\tilde{Z})^L_1 \\ 1 & (\tilde{Z})^L_1 \leq z \leq (\tilde{Z})^U_1 \\ R(z) & (\tilde{Z})^U_1 \leq z \leq (\tilde{Z})^U_\circ \end{cases} \qquad (۳)$$

که در آن توابع مرجع $L(z)$ و $R(z)$ به صورت زیر هستند

$$L(z) = \frac{z - (\tilde{Z})^L_\circ}{(\tilde{Z})^L_1 - (\tilde{Z})^L_\circ} \quad , \quad R(z) = \frac{(\tilde{Z})^U_\circ - z}{(\tilde{Z})^U_\circ - (\tilde{Z})^U_1}$$

تذکر ۱.۴ اگر تمام داده‌های فازی به مقادیر حقیقی تبدیل شوند، معادلات (۲) ساده شده و به معادله حالت کلاسیک تبدیل می‌شوند.

۲.۴ حالت σ مجهول

در این حالت، آماره‌ی $Z = \frac{\bar{X}-\mu_\circ}{\frac{s}{\sqrt{n}}}$ به صورت زیر تبدیل می‌شود

$$\tilde{T} \approx \sqrt{n}\frac{(\frac{1}{n}\sum_{i=1}^n \tilde{X}_i) - \mu_\circ}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (\tilde{X}_i - \frac{1}{n}\sum_{i=1}^n \tilde{X}_i)^2}} \qquad (۴)$$

در این جا نیز فرض می‌کنیم مشاهدات، به صورت اعداد فازی ذوزنقه‌ای باشند. ابتدا تابع عضویت آماره فوق را بر اساس αـبرش‌ها به صورت زیر به دست می‌آوریم، سپس از آن برای انجام آزمون استفاده می‌کنیم. فرض کنید

$$\tilde{T}_\alpha^L = \min_{(\tilde{X}_i)_\alpha^L \le X_i \le (\tilde{X}_i)_\alpha^U} \{\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^n X_i - \mu_0}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \frac{1}{n}\sum_{i=1}^n X_i)^2}}\} \qquad (1-5)$$

$$\tilde{T}_\alpha^U = \max_{(\tilde{X}_i)_\alpha^L \le X_i \le (\tilde{X}_i)_\alpha^U} \{\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^n X_i - \mu_0}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \frac{1}{n}\sum_{i=1}^n X_i)^2}}\} \qquad (2-5)$$

که $(\tilde{X}_i)_\alpha^U$ و $(\tilde{X}_i)_\alpha^L$ را بر پایه‌ی نمونه می‌توان محاسبه نمود. بنابراین αـبرش‌های $\tilde{T}$ به صورت $\tilde{T}_\alpha = [\tilde{T}_\alpha^L, \tilde{T}_\alpha^U]$ است، که در آن $\tilde{T}_\alpha^U$و$\tilde{T}_\alpha^L$ توابع غیرخطی بر حسب مشاهدات می‌باشند. برای محاسبه $\tilde{T}_\alpha$ می‌توان از نرم‌افزارهای محاسباتی استفاده کرد. بنابراین، با استفاده از مقادیر مختلف α، تابع عضویت $\tilde{T}$ که به صورت عدد $(\tilde{T}_0^L, \tilde{T}_1^L, \tilde{T}_1^U, \tilde{T}_0^U)_{LR}$ می‌باشد، به صورت زیر به دست می‌آید

$$\mu_{\tilde{T}}(t) = \begin{cases} 0 & t < \tilde{T}_0^L \ \ or \ \ t > \tilde{T}_0^U \\ L(t) & \tilde{T}_0^L \le t \le \tilde{T}_1^L \\ 1 & \tilde{T}_1^L \le t \le \tilde{T}_1^U \\ R(t) & \tilde{T}_1^U \le t \le \tilde{T}_0^U \end{cases} \qquad (6)$$

که $L(t), R(t)$ توابع مرجع می‌باشند.

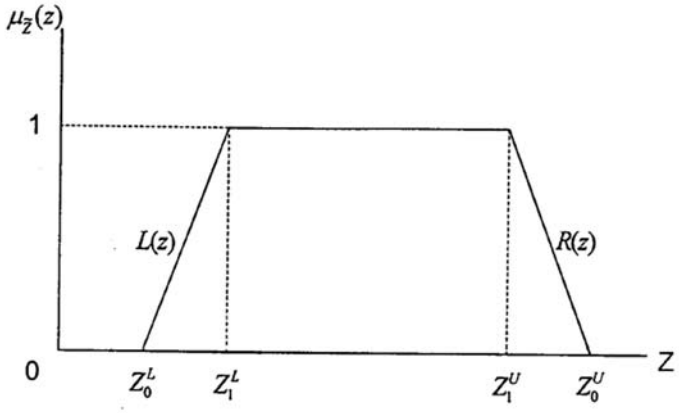
نکته ۱.۴ زمانی که همه‌ی داده‌ها مقادیر دقیق باشند روابط b, a یکی و به آماره حالت کلاسیک تبدیل می شوند.

۵ آزمون میانگین جامعه (σ معلوم)

همان طورکه اشاره شد در این حالت آماره‌ی آزمون به صورت زیر خواهد بود

$$\tilde{Z} = \frac{\frac{1}{n}\sum_{i=1}^n \tilde{X}_i - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

تحت این فرض که مشاهدات $\tilde{X}_i$ اعداد فازی ذوزنقه‌ای هستند، مقدار $\tilde{Z}$ یک عدد فازی ذوزنقه‌ایLR و به صورت $\tilde{Z} = (\tilde{Z}_0^L, \tilde{Z}_1^L, \tilde{Z}_1^U, \tilde{Z}_0^U)_{LR}$ است [14]، که در آن $\tilde{Z}_0^L = \frac{\tilde{\bar{X}}_0^L - \mu_0}{\frac{\sigma}{\sqrt{n}}}$



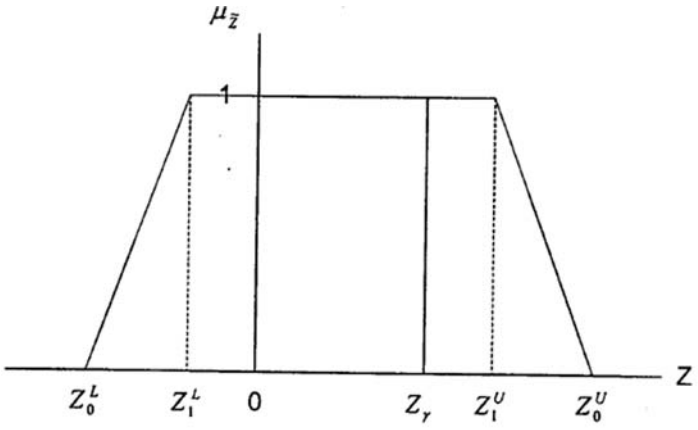
شکل ۱:

و $\tilde{Z}_1^L = \frac{\bar{\tilde{X}}_1^L - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ و $\tilde{Z}_0^U = \frac{\bar{\tilde{X}}_0^U - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ و $\tilde{Z}_1^U = \frac{\bar{\tilde{X}}_1^U - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ شکل ۱ تابع عضویت $\tilde{Z}$ را نشان می‌دهد.

تعریف ۱.۵ در شیوه‌ی آزمون تصادفی شده بر اساس داده‌های فازی، احتمال رد فرض صفر را به صورت احتمال قرار داشتن آماره آزمون در ناحیه بحرانی تقسیم بر احتمال مرتبط با آماره آزمون تعریف می‌کنیم و با P_0 نشان می‌دهیم. همچنین احتمال پذیرش فرض صفر را به صورت $1 - P_0$ در نظر می‌گیریم.
در حالت خاص، $P_0 = 0$ نشان دهنده‌ی عدم رد فرض صفر و $P_0 = 1$ نشان دهنده‌ی رد قطعی فرض صفر می باشد.
تعریف ۲.۵ احتمال مرتبط با آماره آزمون فازی $\tilde{Z}$ به صورت زیر تعریف می شود

$$\Delta = \int_{\tilde{Z}_0^L}^{\tilde{Z}_1^L} L(z)\phi(z)dz + \int_{\tilde{Z}_1^L}^{\tilde{Z}_1^U} \phi(z)dz + \int_{\tilde{Z}_1^U}^{\tilde{Z}_0^U} R(z)\phi(z)dz \qquad (۷)$$

که در آن $\phi(z)$ تابع چگالی احتمال توزیع نرمال استاندارد Z است.
فرض کنید که می‌خواهیم فرض صفر $H_0 : \mu \leq \mu_0$ در برابر فرض مقابل $H_a : \mu > \mu_0$ آزمون کنیم. شکل ۲ رابطه‌ی بین تابع عضویت آماره آزمون فازی $\tilde{Z}$ و ناحیه بحرانی $\{z : z > z_\gamma\}$ را، که γ سطح معنی داری آزمون و z_γ مقدار بحرانی است، نشان می‌دهد.



شکل ۲:

با توجه به شکل ۲ احتمال رد فرض صفر به صورت زیر به دست می‌آید

$$P_0 = \frac{1}{\Delta}[\int_{\tilde{Z}_\gamma}^{\tilde{Z}_1^U} \phi(z)dz + \int_{\tilde{Z}_1^U}^{\tilde{Z}_0^U} R(z)\phi(z)dz] \qquad (۸)$$

در شکل ۳، پنج نمونه ممکن از توابع عضویت $\tilde{Z}$، مربوط به آزمون یکطرفه راست فرض صفر $H_0 : \mu \leq \mu_0$ در برابر فرض مقابل $H_a : \mu > \mu_0$ نشان داده شده است. با در نظر گرفتن ناحیه بحرانی $\{\tilde{Z} \quad s.t \quad \tilde{Z} > \tilde{Z}_\gamma\}$ و بر اساس روابط (۷) و (۸) احتمال رد فرض صفر در حالت‌های a تا e عبارت است از

$$a) P_0 = 0$$

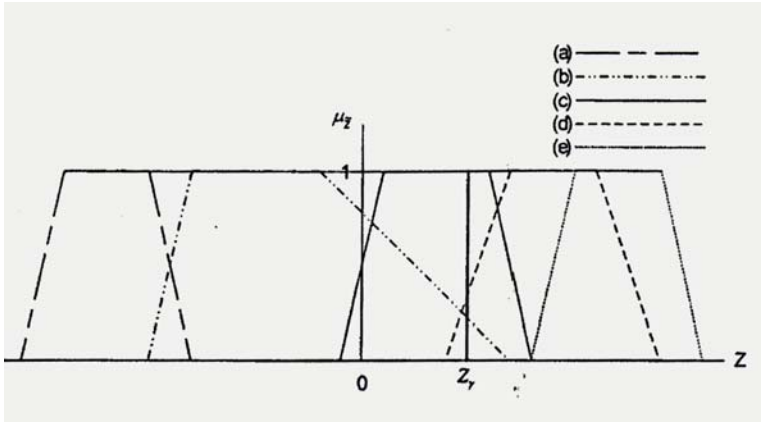
$$b) P_0 = \frac{1}{\Delta}[\int_{\tilde{Z}_\gamma}^{\tilde{Z}_1^U} R(z)\phi(z)dz]$$

$$c) P_0 = \frac{1}{\Delta}[\int_{\tilde{Z}_\gamma}^{\tilde{Z}_1^U} \phi(z)dz + \int_{\tilde{Z}_1^U}^{\tilde{Z}_0^U} R(z)\phi(z)dz]$$

$$d) P_0 = \frac{1}{\Delta}[\int_{\tilde{Z}_\gamma}^{\tilde{Z}_1^L} L(z)\phi(z)dz + \int_{\tilde{Z}_1^L}^{\tilde{Z}_1^U} \phi(z)dz + \int_{\tilde{Z}_1^U}^{\tilde{Z}_0^U} R(z)\phi z dz]$$

$$e) P_0 = 1$$

تذکر ۱.۵ تعریف مقدار احتمال P_0 در حالت یکطرفه چپ شبیه به حالت یکطرفه راست است.
تذکر ۲.۵ از روش فوق می‌توان برای آزمون فرض‌های ساده در برابر مرکب نیز استفاده نمود [7].



شکل ۳:

مثال ۱.۵ فرض کنید $\tilde{X_۸}, \ldots, \tilde{X_۲}, \tilde{X_۱}$، هشت مشاهده فازی از توزیع نرمال $N(\mu, ۱/۶۱۸)$ به صورت اعداد فازی مثلثی متقارن زیر باشند

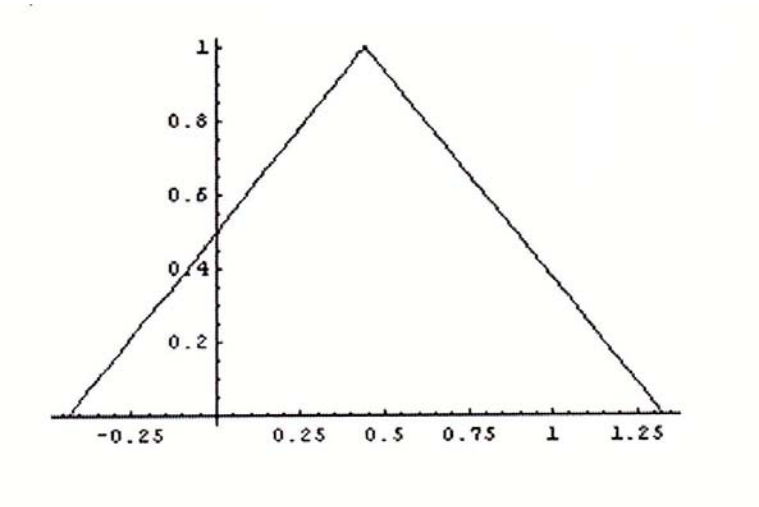
$$\tilde{X_۱} = (۳/۰۸, ۰/۳, ۰/۳)_T \quad , \quad \tilde{X_۲} = (۲/۸۶, ۰/۳, ۰/۳)_T$$
$$\tilde{X_۳} = (۶/۲۵, ۰/۶, ۰/۶)_T \quad , \quad \tilde{X_۴} = (۴/۱۱, ۰/۴, ۰/۴)_T$$
$$\tilde{X_۵} = (۲/۷۱, ۰/۲, ۰/۲)_T \quad , \quad \tilde{X_۶} = (۴/۴۵, ۰/۴, ۰/۴)_T$$
$$\tilde{X_۷} = (۵/۰۵, ۰/۵, ۰/۵)_T \quad , \quad \tilde{X_۸} = (۵/۲۳, ۰/۵, ۰/۵)_T$$

می‌خواهیم آزمون یکطرفه فرض $H_۰ : \mu \leq ۴$ را در برابر فرض مقابل $H_a : \mu > ۴$ انجام دهیم. با توجه به این‌که هر عدد فازی مثلثی حالت خاصی از یک عدد فازی ذوزنقه‌ای است بنابراین تمام مشاهدات فوق را می‌توان نوعی اعداد فازی ذوزنقه‌ای در نظرگرفت. در این مثال $\sigma^۲ = ۱/۶۱۸$، بنابراین آماره آزمون به شکل زیر می‌باشد

$$\tilde{Z} = \frac{\frac{۱}{n}\sum_{i=۱}^{n} \tilde{X_i} - \mu_۰}{\frac{\sigma}{\sqrt{n}}}$$

با کمک معادلات (۲) و (۳)، تابع عضویت $\tilde{Z}$ عبارت است از

$$\mu_{\tilde{Z}}(z) = \begin{cases} \frac{۱}{۰/۸۸}(z + ۰/۴۴) & -۰/۴۴ \leq z \leq ۰/۴۴ \\ \frac{۱}{۰/۸۹}(۱/۳۳ - z) & ۰/۴۴ \leq z \leq ۱/۳۳ \end{cases} \qquad (۹)$$



شکل ۴:

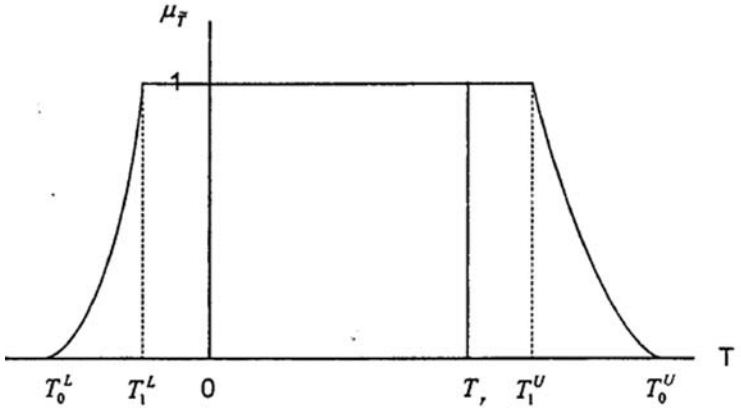
در شکل ۴ نمودار عدد فازی مثلثی فوق رسم شده است. تکیه گاه آماره آزمون فازی، $\tilde{Z}$، بین ۱/۳۳ و۰/۴۴− می‌باشد. این دامنه نشان دهنده این مطلب است که اگر چه آماره آزمون، یک عدد فازی است، اما مقدارش نمی‌تواند خارج از بازه $[۱/۳۳, ۰/۴۴-]$ باشد.
با توجه به اینکه ناحیه بحرانی این آزمون یکطرفه در سطح معنی‌داری ۰/۱ به صورت $\{\tilde{Z} \quad s.t \quad \tilde{Z} > \tilde{Z}_{۰/۱}\}$ می‌باشد، با کمک معادله (۷)، احتمال مربوط به آماره آزمون فازی، $\tilde{Z}$، به صورت زیر محاسبه می‌شود

$$\Delta = \int_{-۰/۴۴}^{۰/۴۴} \frac{۱}{۰/۸۸}(z + ۰/۴۴)\phi(z)dz + \int_{۰/۴۴}^{۱/۳۳} \frac{۱}{۰/۸۹}(۱/۳۳ - z)\phi(z)dz = ۰/۳۰۴۲$$

سرانجام، با کمک معادله (۸)، احتمال رد فرض صفر عبارت است از

$$P_{۰} = \frac{۱}{۰/۳۰۴۲}[\int_{۱/۲۸}^{۱/۳۳} \frac{۱}{۰/۸۹}(۱/۳۳ - z)\phi(z)dz] = ۰/۰۰۰۸$$

بنابراین فرض صفر با احتمال ۰/۰۰۰۸ رد و با احتمال ۰/۹۹۹۲ پذیرفته می‌شود.



شکل ۵:

۶ آزمون میانگین جامعه (σ مجهول)

در این حالت، با توجه به مطالب بخش چهارم، داریم

$$\tilde{T} \approx \sqrt{n}\frac{(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i) - \mu_0}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{X}_i - \frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i)^2}}$$

که در آن

$$S^2 = \frac{\sum_{i=1}^{n}(\tilde{X}_i - \frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i)^2}{n-1}$$

توجه کنید که، عطف به نکته ۱.۲، رابطه‌ی موجود بین آماره‌ی آزمون فازی و داده‌های فازی در آماره $\tilde{T}$ غیر خطی می‌باشد. بنابراین به دست آوردن تابع عضویت $\mu_{\tilde{T}}$ به طور دقیق تقریباً غیر ممکن است.
فرض کنید می‌خواهیم فرض صفر $H_0 : \mu \leq \mu_0$ را در برابر فرض مقابل $H_a : \mu > \mu_0$ آزمون کنیم. شکل ۵ رابطه‌ی بین تابع عضویت $\mu_{\tilde{T}}$ و ناحیه بحرانی $\{\tilde{T} \quad s.t \quad \tilde{T} > \tilde{T}_\gamma\}$ را توصیف می‌کند، که در آن γ سطح معنی داری آزمون و $\tilde{T}_\gamma$ مقدار بحرانی می‌باشد.

تعریف ۱.۶ احتمال مرتبط با آماره آزمون فازی، $\tilde{T}$، به صورت زیر تعریف می‌شود

$$\Delta = \int_{\tilde{T}_{\circ}^{L}}^{\tilde{T}_{1}^{L}} L(t)p(t)dt + \int_{\tilde{T}_{1}^{L}}^{\tilde{T}_{1}^{U}} p(t)dt + \int_{\tilde{T}_{1}^{U}}^{\tilde{T}_{\circ}^{U}} R(t)p(t)dt \qquad (۱۰)$$

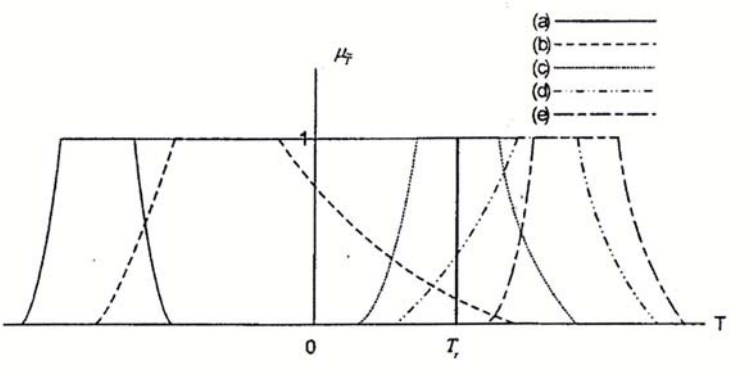
که در آن $p(t)$ تابع چگالی احتمال استیودنت می‌باشد.
در این صورت احتمال رد فرض صفر با توجه به شکل ۵ به صورت زیر تعریف می‌شود

$$P_{\circ} = \frac{1}{\Delta}[\int_{\tilde{T}_{\gamma}}^{\tilde{T}_{1}^{U}} p(t)dt + \int_{\tilde{T}_{1}^{U}}^{\tilde{T}_{\circ}^{U}} R(t)p(t)dt] \qquad (۱۱)$$

تذکر ۱.۶ احتمال رد فرض صفر برای آزمون‌های یکطرفه و دو طرفه به طور مشابه تعریف می‌شوند.
در شکل ۶، تعداد ۵ نمونه ممکن از توابع عضویت $\tilde{T}$، مربوط به آزمون یکطرفه راست فرض صفر $H_{\circ} : \mu \leq \mu_{\circ}$ در برابر فرض مقابل $H_a : \mu > \mu_{\circ}$ نشان داده است. با در نظرگرفتن ناحیه بحرانی $\{\tilde{T} \quad | \quad \tilde{T} > \tilde{T}_{\gamma}\}$ و بر اساس روابط (۱۰) و (۱۱) احتمال رد فرض صفر در حالت‌های a تا e به صورت زیر به دست می‌آیند

$$a) P_{\circ} = \circ$$
$$b) P_{\circ} = \frac{1}{\Delta}[\int_{\tilde{T}_{\gamma}}^{\tilde{T}_{\circ}^{U}} R(t)p(t)dt]$$
$$c) P_{\circ} = \frac{1}{\Delta}[\int_{\tilde{T}_{\gamma}}^{\tilde{T}_{1}^{U}} p(t)dt + \int_{\tilde{T}_{1}^{U}}^{\tilde{T}_{\circ}^{U}} R(t)p(t)dt]$$
$$d) P_{\circ} \frac{1}{\Delta}[\int_{\tilde{T}_{\gamma}}^{\tilde{T}_{1}^{L}} L(t)p(t)dt + \int_{\tilde{T}_{1}^{L}}^{\tilde{T}_{1}^{U}} p(t)dt + \int_{\tilde{T}_{1}^{U}}^{\tilde{T}_{\circ}^{U}} R(t)p(t)dt]$$
$$e) P_{\circ} = 1$$

چون بر اساس معادله (۴) رابطه‌ی بین آماره آزمون و داده‌های فازی غیر خطی است، بنابراین به دست آوردن $R(t)$ و $L(t)$ به طور دقیق بسیار دشوار است. اما می‌توان از روش‌های عددی استفاده نمود و مقادیر تقریبی برای $\int_{\gamma}^{\tilde{T}_{\circ}^{U}} R(t)p(t)dt, \int_{\tilde{T}_{\gamma}}^{\tilde{T}_{1}^{L}} L(t)p(t)dt, \int_{\tilde{T}_{1}^{U}}^{\tilde{T}_{\circ}} R(t)p(t)dt$ به دست آورد. آنگاه مطابق شیوه‌ی بیان شده در حالت σ معلوم عمل نموده و احتمال رد فرض صفر را محاسبه می‌کنیم.



شکل ۶:

مراجع

[1] Buckley, J.J., Eslami, E. (2002), An Introduction to Fuzzy Logic and Fuzzy Sets, Springer.

[2] Casals, R., Gil, M.A., Gil, P. (1986), On the use of Zadeh' probabilistic definition for testing statistical hypotheses from fuzzy information, Fuzzy Sets and Systems, 20: 175-190.

[3] Casals, M.R., Gil, M.A., Gil, P. (1986), The fuzzy decision problem: an approach to the problem of testing statistical hypothesis with fuzzy information, Euro. J. Oper. Res., 27: 371-382.

[4] Casals, R., Gil, M.A. (1989), A note on the operativeness of Neyman-Pearson tests with fuzzy information, Fuzzy Sets and Systems, 30: 215-220.

[5] Casella, G., Berger, R.L. (2002), Statistical Inference, Sec. Ed., Buxbury.

[6] Grzegorzewski, P. (2001), Fuzzy tests- defuzzification and randomization, Fuzzy Sets and Systems, 118: 437- 446.

[7] Leu, C.H., Chen, C.C. (2004), Randomized tests of mean for normal population with fuzzy data, Preprint.

[8] Montenegro, M., Casals, M.R., Lubiano, M.A., Gil, M.A. (2001), Two-sample hypothesis tests of means of a fuzzy random variable, Information Sciences, 133: 89- 100.

[9] Romer, C., Kandel, A. (1995), Statistical tests for fuzzy data, Fuzzy Sets and Systems, 72: 1- 26.

[10] Son, J.C., Song, I., Kim, H.Y. (1992), A fuzzy decision problem based on the generalized Neyman- Pearson criterion, Fuzzy Sets and Systems, 47: 65- 75.

[11] Taheri, S.M., Behboodian, J. (2004), On Bayesian approach to fuzzy hypothesis testing with fuzzy data, Ital. J. Pure and Appl. Math., 22.

[12] Watanabe, N., Imaizumi, T. (1993), A fuzzy statistical test of fuzzy hypotheses, Fuzzy Sets and Systems, 53: 167- 178.

[13] Zadeh, L.A. (1968), Probability measures of fuzzy events, J. Math. Anal. Appl., 23: 421- 427.

[14] Zimmerman, H.J. (1996), Fuzzy Set Theory and Its Applications, 3rd Edition, Kluwer, Boston.

تعیین جایگاه اقتصادی استان‌های کشور ـ ۱۳۷۹[1] از نظر ارزش افزوده و درصد نیروی شاغل در بخش‌های اقتصادی به روش تحلیل عاملی

وحید طیفوری

سازمان مدیریت و برنامه‌ریزی استان گیلان معاونت آمار و اطلاعات

چکیده: دو شاخص مهم در ارزیابی و مقایسه نحوه توزیع سرمایه و سایر عوامل اقتصادی در بین استان‌های کشور، ارزش افزوده ایجاد شده و درصد نیروی شاغل در هر یک از بخش‌های اقتصادی می‌باشد. بررسی هر یک از استان‌ها از بعد این دو شاخص بطور مجزا و توام ضمن فراهم نمودن امکان تحلیل‌های اقتصادی، ما را در شناخت امکانات، پتانسیل‌ها، نقاط قوت و ضعف استانها، یاری نموده و ابزار مناسبی برای سیاستگذاری‌های هدفمند در نحوه توزیع سرمایه و امکانات، بمنظور کاهش نابرابری‌های اقتصادی می‌باشد.

در این مقاله، استان‌های کشور از نظر ارزش افزوده ایجاد شده و درصد نیروی شاغل در ۱۵ بخش اقتصادی[2] در سال ۱۳۷۹ بعنوان متغیرهای مورد مطالعه بروش تحلیل عاملی[3] در دو بخش مورد بررسی قرار گرفته است. طبق نتایج بدست آمده در تحلیل ارزش افزوده بر اساس آزمون‌های انجام شده سه عامل با تبیین بیش از ۹۱ درصد واریانس کل پیشنهاد گردید که پس از محاسبات لازم، رتبه‌بندی استانها برای هر یک از سه عامل، مشخص گردیده است. در بررسی درصد نیروی شاغل در هریک از بخش‌های اقتصادی نیز با استفاده از همین روش پس از حذف برخی متغیرها با برازش مدل مناسب در ۱۱ بخش (متغیر) باقیمانده، مدل سه عاملی با تبیین حدود ۷۹ درصد واریانس پیشنهاد گردید و پس از محاسبه ضرایب عامل، استانها بر حسب امتیاز بدست آمده رتبه‌بندی شدند و در پایان هر بخش تحلیل کلی از این بررسی بعمل آمده است.

واژه‌های کلیدی: تحلیل عاملی، ارزش افزوده، درصد نیروی شاغل

۱ بخش اول

۱.۱ مقدمه

یکی از دستاوردهای بسیار مهم مرکز آمار ایران که با همکاری گسترده کارشناسان سازمان مدیریت و برنامه‌ریزی کلیه استانها طی سالهای اخیر به تحقق پیوسته، تهیه و تدوین حسابهای منطقه‌ای

۱) آخرین نشریه منتشر شده حساب تولید استانها مربوط به سال ۱۳۷۹ می‌باشد.

2) ISIC Rev.3 بر اساس 3) Factor Analysis

(حساب تولید استان‌ها) بوده است. این طرح با در نظر گرفتن تعاریف و مفاهیم، طبقه‌بندیها و جداول استاندارد کلیه فعالیت‌های اقتصادی را در هر استان مورد پوشش قرار داده و در این چارچوب اطلاعات مورد نیاز را جمع‌آوری و حساب اقتصادی استانها با استفاده از روش تولید محاسبه می‌گردد. مهمترین شاخصهای استنتاج شده از این طرح محاسبه محصول ناخالص داخلی[۴] و ارزش افزوده هر یک از فعالیت‌های اقتصادی در سطح استان می‌باشد. این حسابها بعنوان ابزار لازم در مقوله برنامه‌ریزی منطقه‌ای، یکی از زیر ساخت‌ها محسوب شده و مبنای موثقی در ارزیابی عملکرد نحوه توزیع سرمایه و امکانات در کشور و حرکتی برای امکان تدوین برنامه‌های لازم بمنظور رفع کاستیهای توسعه اقتصادی در کشور می‌باشد.

بمنظور سهولت در تحلیل سازوکارهای اقتصادی توسط برنامه‌ریزان تعیین جایگاه اقتصادی استانها از بعد دو شاخص ارزش افزوده و درصد نیروی شاغل در هر یک از بخش‌های اقتصادی بصورت مجزا و توام، می‌تواند در شناخت امکانات و پتانسیل‌ها، نقاط قوت و ضعف استانها، مورد ارزیابی قرار گیرد.

بنابر اهمیت موضوع در گزارش حاضر استان‌های کشور در سال ۱۳۷۹ از دو بعد ارزش افزوده و درصد نیروی کار در هر یک از فعالیتهای اقتصادی [۵] بروش تحلیل عاملی، در دو بخش بصورت مجزا رتبه‌بندی و مورد بررسی قرارگرفته و در بخش آخر تحلیلی بر رابطه این دو شاخص اقتصادی با یکدیگر بعمل آمده است. با توجه به اهمیت شاخصهای یاد شده و کاربرد متنوع آنها در تحلیل‌های اقتصادی، گزارش حاضر از جمله اهداف زیر را می‌تواند پوشش دهد.

- فراهم نمودن زمینه لازم در خصوص مطالعه منابع اقتصادی یا شناسایی و تعیین محورهای توسعه استانها.
- شناخت مهمترین فعالیت (فعالیتها) در هر استان.
- تعیین نقش و جایگاه هر استان در ساختار اقتصادی کشور.
- مبنای تحقیقات تفصیلی در نحوه توزیع سرمایه و امکانات در جهت حذف فاصله‌های عمیق اقتصادی و همچنین تمرکز یا جابجایی سرمایه و نیروی کار در بخشهای مستعد و یا پر بازده.

۲.۱ روشهای رتبه‌بندی

الف ـ روش وزن‌دهی

یکی از روشهایی که به رتبه‌بندی فعالیتها بر اساس برخی صفات که به صورت شاخصهای کمی یا کیفی بیان می‌شوند روش وزنی است. در این روش هر فعالیت به تناسب مقداری از صفت که به خود اختصاص می‌دهد امتیاز می‌گیرد. و در نهایت، جمع امتیازها، وضعیت فعالیت را نسبت به سایر فعالیتها مشخص می‌کند.

ویژگیهای این روش: سهولت و قابلیت کاربرد در تحقیقات مختلف، امکان استفاده از شاخصهای

4) Gross Domestic Products Region (GDPR) 5) International Standard Industrial Classification Of All Economic Active (ISIC Rev3)

کیفی، عدم لزوم همسویی شاخصها با یکدیگر و محدودیت‌های آن تورش نسبت به شاخصهایی که بایکدیگر همبستگی دارند، حساسیت زیاد این روش نسبت به چگونگی وزن‌دهی به شاخصها می‌باشد.

ب ــ روش تاکسونومی عددی

تاکسونومی عددی روشی است برای طبقه‌بندی و رتبه‌بندی موضوعات، به گونه‌ای که بین عناصر هر طبقه بیشترین تشابه وجود داشته باشد و در عین حال بیشترین اختلاف را با عناصر طبقات دیگر داشته باشد. به این طریق می‌توان مجموعه عناصر همگن را در یک طبقه قرار داد و سپس بر اساس امتیاز کسب شده رتبه‌بندی را انجام داد. از مزایای این روش می‌توان به سهولت در انجام محاسبات و امکان محاسبه شاخصهای توسعه یافتگی اشاره نمود اما محدودیتها این روش لزوم تعیین شاخصها با توجه به انتخاب و نظر کارشناسان هر رشته مطالعاتی و همچنین لزوم همسویی شاخصها می‌باشد.

ج ــ روش میزان انحراف از بهینه

این روش برای رتبه‌بندی جوامع از نظر شاخصهای کمی مورد استفاده قرار می‌گیرد. مبنای این روش بر استفاده از اعداد و ارقام خام شاخصها و استاندارد نمودن آنها می‌باشد. از ویژگیهای این روش، سهولت در محاسبات و سادگی در درک آن توسط کاربران، حذف مقیاس شاخصها بدلیل استاندارد نمودن آنها، تعیین میزان انحراف شاخص از مقدار بهینه بصورت یک نسبت و از محدویتهای این روش می‌توان به لزوم همسویی شاخصها، حساسیت نسبت به شاخصهای انتخابی، تورش نسبت به همبستگی بین شاخصها، اهمیت یکسان کلیه شاخصها نسبت به یکدیگر اشاره نمود.

د ــ روش تحلیل عاملی

کلی‌ترین شکل الگوی عاملی امکان اندازه‌گیری متغیرهایی که نمی‌توانند اندازه‌گیری شوند را فراهم می‌نماید. الگوی عاملی مجموعه‌ای از معادلات خطی هستند که به ویژه در علوم رفتاری، تصمیم در خصوص سوددهی شرکتها، ممیزی در استخدام‌ها، اقتصاد و ... کاربرد دارد. این روش عموماً برای خلاصه نمودن اطلاعات مورد استفاده قرار می‌گیرد. بدین ترتیب با تلخیص و تلفیق متغیرهای مختلف اثر آنها را در قالب چند عامل که بیشترین اختلاف را با یکدیگر دارند ارایه می‌شود. از این روش به منظور طبقه‌بندی، رتبه‌بندی و تجزیه و تحلیل در علوم مختلف استفاده می‌شود. از مزایای این روش می‌توان به استفاده از روشها و تکنیک‌های منطقی و پیشرفته برای برآورد مقادیر عاملها (نظیر روش کمترین توانهای دوم موزون و یا روش رگرسیونی)، امکان استاندارد کردن متغیرها، استفاده از تکنیک دوران به منظور سهولت در تفسیر عوامل اشاره نمود. محدویت‌های این روش عبارت است از نیاز به تصمیم گیری‌های زیاد در مطالعه تحلیل عاملی (نظیر تعیین تعداد عاملها، تشخیص دوران و ...) حجم زیاد محاسبات، حساسیت نسبت به چگونگی انتخاب و کیفیت اطلاعات و همچنین عدم تضمین در وجود عوامل معنی‌دار در تمامی تحلیل‌های آماری داده‌ها.

۳.۱ تعریف متغیرها

متغیرهای در نظر گرفته شده در بخش دوم عبارتند از ارزش افزوده ایجاد شده در هر یک از فعالیت‌های اقتصادی سال ۱۳۷۹، به تفکیک استان و در قالب ماتریسی تعریف شده است، که عنصر a_{ij} این ماتریس بیانگر ارزش افزوده فعالیت اقتصادی jام $(j = ۱, \ldots, ۱۵)$ در استان iام $(i = ۱, \ldots, ۲۸)$ می‌باشد. متغیرهای بخش سوم عبارت است از برآورد درصد نیروی شاغل در هریک از فعالیت‌های اقتصادی به تفکیک استان در سال ۱۳۷۹، که در قالب یک ماتریس تعریف شده است که عنصر b_{ij} این ماتریس بیانگر برآورد درصد نیروی شاغل در فعالیت اقتصادی jام $(j = ۱, \ldots, ۱۵)$ در استان iام $(i = ۱, \ldots, ۲۸)$ می‌باشد.

۴.۱ مراحل کلی انجام کار بروش تحلیل عاملی

الف ــ تعیین ماتریس ضرایب همبستگی بین متغیرها و تفسیر آن:
ماتریس همبستگی و آزمونهای منتج شده از آن ما را در تایید و یا رد استفاده از روش تحلیل عاملی یاری می‌کند. مهمترین شاخص مورد استفاده، شاخص KMO[۶] می‌باشد که برای مقایسه مقادیر ضرایب همبستگی ساده و جزیی بروی کلیه متغیرها مورد استفاده قرار می‌گیرد و مقادیر کمتر از ۰/۵ آن عدم کارایی این روش را بیان می‌کند و رابطه آن عبارت است از:

$$KMO = \frac{\sum\sum_{i \neq j} {r_{ij}}^۲}{\sum\sum_{i \neq j} {r_{ij}}^۲ + \sum\sum_{i \neq j} {d_{ij}}^۲}$$

که درآن
r_{ij} : ضریب همبستگی ساده
d_{ij} : ضریب همبستگی جزیی

ب ــ تعیین تعداد عاملها، محاسبه عاملها و بررسی میزان برازش مدل:
در صورت تایید استفاده از روش تحلیل عاملی، یکسری عوامل اولیه به یکی از روشهای موجود نظیر: مولفه‌های اصلی[۷]، بیشینه درستنمایی[۸]، کمترین مربعات ناموزون[۹] و . . . تعیین می‌شود. اغلب انتخاب تعداد عاملها مبتنی بر چند ترکیب : ۱ــ میزان تبیین واریانس هر عامل ۲ــ دانش مورد بحث ۳ــ عقلایی بودن نتایج، است. (البته لازم بذکر است که هر چه میزان تبیین واریانس در عوامل انتخابی بیشتر باشد، کارایی نتایج بهتر خواهد بود).
در این گزارش برای تعیین تعداد عاملها از روش مولفه‌های اصلی استفاده شده و با تکیه بر موارد مطرح شده، تعداد عاملهای نهایی تعیین شده است.

6) Kaiser-Meyer-Olkin Measure of Sampling Adequacy 7) Principal Component
8) Maximum Likelihood 9) Unweighted Least Square

ج ــ دوران:
هنگامیکه یک یا چند متغیر خاص توسط بیش از یک عامل تبیین شود آنگاه تفسیر عوامل مشکل خواهد بود لذا در شرایط مذکور و بمنظور سهولت در تفسیر، با انجام عمل دوران به یکی از روشهای متعامد یا مایل امکان تفسیر بهتر عوامل فراهم می‌گردد.

د ــ تعیین مقدار متناظر هر عامل برای متغیرهای مورد بررسی و تهیه جداول نهایی:
در این مرحله مشخص می‌شود که هر یک از عاملهای تعیین شده بر چه متغیر (متغیرهایی) تاکید دارند. پس از انطباق عوامل و با محاسبه ضرایب عاملها و بکارگیری متغیرهای استاندارد شده می‌توانیم بروش زیر امتیاز نهایی هر استان را بر اساس هر یک از عوامل تعیین نماییم.

$$A_{jk} = \sum_{i=۱}^{q} w_{ji} z_{ik}$$

$j = ۱, ۲, \ldots, m$
$k = ۱, ۲, \ldots, n$
m: تعداد عاملها
n: تعداد استانها
q: تعداد متغیرها
A_{jk}: امتیاز نهایی عامل j ام برای استان kام
w_{ji}: مقدار عامل jام برای فعالیت (بخش) iام
z_{ik}: مقدار استاندارد شده فعالیت (بخش) iام برای استان kام

۵.۱ تعاریف و مفاهیم

ارزش افزوده[۱۰]
عبارت است از تفاوت بین ارزش کالاها و خدمات تولید شده و ارزش کالاها وخدمات بکار رفته در جریان تولید و یا به بیان دیگر ارزش افزوده حاصل ارزش ستانده منهای ارزش داده است.

ستانده[۱۱]
مجموع ارزش کالاها و خدماتی است که در یک واحد تولیدی، تولید شده و برای استفاده در خارج آن واحد در دسترس قرار می‌گیرد.

داده[۱۲] **(مصارف واسطه)**[۱۳]
عبارت است از ارزش کالاها و خدماتی است که به صورت داده یک فرآیند تولید، در آن فرآیند، مصرف می‌شوند. به استثنای مصرف داراییهای ثابت (استهلاک).

شاغل
کلیه کسانی (ده ساله و بیشتر) که در هفت روز گذشته قبل از مراجعه مأمور آمارگیر حداقل دو روز به کاری اشتغال داشته‌اند و یا شاغل بوده‌اند و به عللی (از قبیل مرخصی و یا بیماری و

10) Value Added 11) Output 12) Input 13) Intermediate Comsumption

. . .) کار نکرده‌اند، شاغل محسوب می‌شوند.

درصد شاغلین

عبارت است از نسبت تعداد شاغلان ده ساله و بیشتر بـــه جمعیت فعال ده سالـــه و بیشتر (شاغل و بیکار) ضربدر ۱۰۰.

طبقه‌بندی فعالیتهای اقتصادی

در این طبقه‌بندی فعالیتهای اقتصادی بر اساس تشابه ساختار تولید و بعضی روابط موجود آنها، از اهمیت استراتژیک برخوردار می‌باشند. که جنبه عمده وجوه تشابه مورد نـــظر عبارت است از:

ـــ مشخصه کالاها و خدمات تولید شده.

ـــ موارد استفاده از کالا و خدمات.

ـــ داده‌های بکار رفته در تولید، فرآیند تولید و تکنولوژی مورد استفاده در آن.

ساختار کلی طبقه‌بندی استاندارد فعالیتهای اقتصادی بر اساس ISIC(rev3) در سطح رده‌های الفبایی بشرح زیر می‌باشد:

الف ـــ کشاورزی، شکار و جنگلداری

ب ـــ ماهیگیری

پ ـــ معدن

ت ـــ صنعت

ث ـــ تامین برق، گاز و آب

ج ـــ ساختمان

چ ـــ عمده فروشی، خرده فروشی و تعمیر وسایل نقلیه موتوری و کالاهای شخصی و خانگی

ح ـــ هتل و رستوران

خ ـــ حمل و نقل، انبارداری و ارتباطات

د ـــ واسطه‌گریهای مالی

ذ ـــ مستغلات، کرایه و خدمات کسب و کار

ر ـــ اداره امور عمومی، دفاعی و تامین اجتماعی

ز ـــ آموزش

ژ ـــ بهداشت و درمان و مددکاری اجتماعی

س ـــ سایر فعالیتهای خدمات عمومی، اجتماعی و شخصی

۲ بخش دوم

تعیین و بررسی تحلیلی جایگاه استانها از نظر ارزش افزوده فعالیتهای اقتصادی «بروش تحلیل عاملی»

۱.۲ تحلیل همبستگی

نتایج حاصل شده از ماتریس همبستگی ارزش افزوده بخش‌های مختلف اقتصادی به عنوان متغیرهای مورد بررسی نشان می‌دهد که اکثر متغیرهای دارای همبستگی مثبت و بالایی بوده و فقط سه متغیر کشاورزی و . . . ، ماهیگیری و معدن دارای کمترین میـــزان همبستگی با سایـــر متغیرها می‌باشـــد. با بررسی ماتریس سطح معنی‌داری (ماتریس مذکور به واقع بیانگر میزان احتمال ارتکاب اشتباه نوع اول می‌باشد) آزمون معنی‌دار بودن فرض $H_0 : r_{ij} = 0$ برای کلیه متغیرها رد شده و لذا همبستگی بین متغیرها مورد تایید قرار گرفت.

در گام بعدی با بررسی اعتبار روش تحلیل عاملی، با استفاده از آزمون‌های موجود به شرح زیر داریم.

با توجه به اینکه دترمینان ماتریس همبستگی برابر $۷/۴۳۲ \times ۱۰^{-۲۳}$ بوده و چون به اندازه کافی کوچک است پس روش تحلیل عاملی در این مرحله تایید می‌شود.

در بررسی ماتریس باقیمانده‌ها نیز چون بیش از ۹۲ درصد مقادیر این ماتریس کمتر از ۵ درصد می‌باشد و بدین ترتیب برازش مدل تحلیل عاملی پذیرفته می‌شود.

همچنین باتوجه به عناصر روی قطر ماتریس تصویر معکوس همبستگی که در واقع بیانگر مقدار MSA_iها برای هر یک از متغیرها می‌باشد و چون بزرگ بودن این عناصر دلیل بر امکان حضور متغیر مورد بررسی در مدل می‌باشد، لذا در نتایج بدست آمده از این ماتریس مشاهده می‌شود که مقدار MSA برای متغیرهای معدن و ماهیگیری نسبتاً پایین بوده اما بدلیلی اهمیت این دو بخش در برخی از استانها و همچنین با توجه به پایین بودن میزان باقیمانده‌ها برای متغیرهای مذکور از حذف آنها خودداری گردید (جدول ۲). مقدار بدست آمده برای شاخص KMO برابر ۰/۸۱۴ بوده که با توجه به بزرگ بودن مقدار آن، استفاده از روش تحلیل عاملی در این بررسی تایید می‌شود.

۲.۲ تعیین تعداد و روش محاسبه عاملها و میزان برازش مدل

برای تعیین تعداد عاملها از روش مولفه‌های اصلی استفاده شده است و لذا بر اساس مقادیر ویژه بدست آمده برای ۱۵ متغیر، چون مقادیر ویژه سه عامل نخست بزرگتر از یک بوده و بیشتر از ۹۱ واریانس کل را تبیین می‌کند (جدول ۳)، که در نتیجه روش تحلیل عاملی با سه عامل مورد پذیرش واقع می‌شود.

در ادامه بمنظور مشخص نمودن اینکه هر عامل بر چه متغیرهایی تاکید دارد اقدام نموده که در محاسبات اولیه بدلیل عدم تفکیک‌پذیری عوامل بر حسب متغیرهای مطرح شده انجام عمل دوران ضروری به نظر رسید. با توجه به نتیجه بدست آمده و برای سهولت در تفسیر عمل دوران متعامد بروش واریماکس[۱۴] انجام شد که نتیجه آن در جدول ۴ مشاهده می‌شود. بر اساس نتایج بدست آمده، تعبیر هر یک از سه عامل به شرح زیر می‌باشد:

ویژگی عامل اول: این عامل عمدتاً بر فعالیتهای غیر از کشاورزی، شکار و جنگلداری و معدن و ماهیگیری تکیه دارد.

ویژگی عامل دوم: این عامل صرفاً بر اهمیت بخش کشاورزی، شکار و جنگلداری تاکید می‌کند.

ویژگی عامل سوم: این عامل با توجه به ضرایب مثبت بزرگ برای فعالیت‌های معدن و ماهیگیری، بر اهمیت فعالیتهای مورد نظر، دلالت دارد.

۳.۲ جداول نهایی

پس از تعیین ویژگی عاملها، به روش رگرسیونی جدول ضرائب نهایی عوامل برای هر یک از متغیر ها در جدول ۵ محاسبه شده است. مقادیر این جدول از رابطه زیر که در آن از متغیرهای استاندارد شده استفاده شده، بدست آمده است.

$$A_{jk} = \sum_{i=۱}^{۱۵} w_{ji} z_{ik} \qquad j = ۱, ۲, \ldots, ۲۸ \ , \ \ k = ۱, ۲, ۳$$

بعنوان مثال در محاسبه امتیاز نهایی استان تهران تحت عامل اول داریم:

$$\begin{aligned} F_{۱۱} &= \sum_{i=۱}^{۱۵} w_{۱i} z_{i۱} = w_{۱۱} z_{۱۱} + w_{۱۲} z_{۲۱} + \cdots + w_{۱,۱۵} z_{۱,۱۵} \\ &= ۰/۱۱۷(\frac{۳۶۹۲ - ۲۶۰۸/۵}{۲۰۰۹/۲۸}) + \cdots + ۰/۱۰۷(\frac{۲۶۶۵ - ۲۳۱/۰۸}{۴۹۳/۰۴}) \\ &= ۴/۹۱۷ \end{aligned}$$

z_iها مقادیر استاندارد شده هر یک از متغیرها بر اساس میانگین و انحراف معیار متناظر با همان متغیر می‌باشند.

۴.۲ نتیجه‌گیری

بمنظور سهولت در تحلیل پس از تعیین امتیاز هر استان، در جدول ۶ و ۷ و ۸ بترتیب نزولی مرتب شده است. بنابراین ملاحظه می‌شود که تحت عامل اول استان تهران به فاصله نسبتاً

14) Varimax with Kaiser Normalization

زیادی نسبت به سایر استان‌های کشور از نظر اهمیت و افزوده‌های فعالیتهای غیر ازکشاورزی و ... ، معدن و ماهیگیری در مقام نخست قرار دارد و تمرکز بخشهایی با ارزش افزوده بالا نظیر: صنعت عمده‌فروشی و خرده‌فروشی و ... و امور عمومی در این استان وجود داشته و منجر به شکاف اقتصادی عمیق نسبت به سایر استانها شده است. از طرفی بمنظور طبقه‌بندی فرضی استان‌های تقریباً همگن تحت این عامل می‌توان استانها را به چهارگروه تقسیم نمود:

ـ گروه اول: تهران
ـ گروه دوم: اصفهان، خراسان، خوزستان
ـ گروه سوم: آذربایجان غربی، هرمزگان، مازندران، گیلان، فارس
ـ گروه چهارم: سایراستانها

در بررسی استانها تحت عامل دوم دیده می‌شود که در استان‌های کرمان، خوزستان، فارس و اصفهان اهمیت بخش کشاورزی، شکار و جنگلداری نسبت به سایر استانها بیشتر بوده و در مقابل استان‌های هرمزگان، بوشهر و سیستان و بلوچستان کمترین رتبه را داشته‌اند. تحت این عامل بر اساس یک دسته‌بندی فرضی می‌توان استان‌های تقریباً همگن را به چهارگروه بشکل زیر دسته‌بندی نمود:

ـ گروه اول: کرمان، خراسان، خوزستان، فارس، اصفهان
ـ گروه دوم: آذربایجان غربی، مازندران، قزوین، همدان، آذربایجان شرقی
ـ گروه سوم: سایر استانها
ـ گروه چهارم: سیستان و بلوچستان، بوشهر، هرمزگان

در بررسی عامل سوم که تاکید بر رونق فعالیت های معدن و ماهیگیری دارد ملاحظه می‌شود که استان‌های خوزستان و هرمزگان بترتیب دارای بالاترین رتبه بوده و در این خصوص استان‌های کرمان و اردبیل در این بخشها پایین‌ترین میزان اهمیت را داشته‌اند. تحت عامل سوم استان‌های همگن را به سه گروه همگن بشکل زیر دسته‌بندی می‌شوند:

ـ گروه اول: خوزستان، هرمزگان
ـ گروه دوم: بوشهر، کهگیلویه و بویراحمد، گیلان، مازندران، سیستان و بلوچستان، گلستان، اصفهان
ـ گروه سوم: سایر استانها

جدول ۱: ارزش افزوده فعالیتهای اقتصادی کشور به تفکیک استان ــ ۱۳۷۹† (میلیارد ریال)

استان	کشاورزی و ...	ماهیگیری	معدن	صنعت	تامین آب و برق و ...	ساختمان	عمده‌فروشی و ...
آذربایجان شرقی	۳۱۷۸	۴۸	۴۹	۱۸۳۱	۱۴۱	۷۳۷	۱۸۶۹
آذربایجان غربی	۳۹۴۲	۲	۶۰	۵۶۹۱	۷۲۵	۷۱۵	۵۵۵۳
اردبیل	۱۹۱۶	۳	۹	۶۷۲	۶۱	۲۵۵	۱۳۷۰
اصفهان	۳۶۰۴	۷	۱۹۳	۱۲۸۳۷	۱۹۷۳	۲۸۶۸	۶۳۵۲
ایلام	۴۰۱	۲	۱۶۲۶	۱۴۶	۲۲	۱۳۳	۳۶۴
بوشهر	۵۶۵	۲۸۴	۳۲۴	۸۵۲	۳۳۲	۲۷۴	۹۱۶
تهران	۳۶۹۲	۵	۳۰۷	۲۹۰۷۶	۲۴۸۳	۸۱۷۴	۳۰۳۰۹
چهارمحال وبختیاری	۱۰۱۶	۸	۳۱	۴۲۲	۴۳	۲۷۴	۶۲۰
خراسان	۶۷۷۳	۳	۲۶۵	۵۷۹۳	۹۷۲	۲۶۹۲	۷۷۹۶
خوزستان	۴۲۰۸	۲۱۲	۵۴۴۸۶	۱۱۱۴۷	۱۶۵۴	۱۴۱۴	۴۶۵۱
زنجان	۱۴۲۸	۳	۹۷	۱۲۹۶	۱۱۶	۲۷۰	۷۳۲
سمنان	۱۱۳۵	۱	۱۰۰	۷۵۵	۸۱	۳۵۴	۳۸۹
سیستان و بلوچستان	۱۰۲۹	۱۸۰	۲۷	۳۹۲	۱۱۰	۳۲۷	۱۰۵۰
فارس	۶۲۳۹	۱۳	۳۹۷	۳۲۷۵	۷۶۵	۱۵۵۴	۳۳۱۸
قزوین	۲۴۲۵	۲	۲۳	۳۳۹۳	۷۷۹	۵۰۴	۱۳۶۷
قم	۷۰۰	۱	۲۹	۱۴۰۶	۷۶	۴۶۶	۱۳۴۵
کردستان	۱۱۰۲	۶	۳۶	۴۶۶	۵۷	۳۰۱	۱۴۱۹
کرمان	۸۱۰۲	۲	۶۴۰	۳۳۰۳	۲۸۵	۸۱۲	۱۹۸۱
کرمانشاه	۱۵۳۰	۲	۴۰	۷۸۰	۲۲۳	۶۱۶	۱۲۵۷
کهگیلویه و بویراحمد	۵۷۳	۴	۳۰۰۲۳	۱۴۸	۳۷	۱۹۶	۴۱۱
گلستان	۲۷۴۲	۱۷۳	۳۰	۵۶۱	۷۷	۴۶۶	۲۰۳۴
گیلان	۲۶۴۰	۲۴۹	۷۵	۲۳۰۲	۵۶۸	۱۲۲۸	۳۳۸۶
لرستان	۲۱۸۱	۹	۱۷۵	۱۰۴۳	۵۷	۲۵۹	۱۵۳۱
مازندران	۵۲۷۰	۲۱۷	۱۱۸	۲۲۱۲	۶۰۸	۱۱۹۵	۴۲۷۱
مرکزی	۱۵۰۷	۱	۷۸	۵۸۱۲	۳۸۵	۷۵۹	۲۳۲۴
هرمزگان	۱۳۶۱	۴۹۳	۸۸	۱۴۰۸	۵۷۳	۲۳۹	۱۳۰۰
همدان	۲۵۹۰	۳	۳۲	۱۰۷۱	۴۷۲	۳۷۱	۱۵۴۹
یزد	۱۱۶۷	۱	۴۲۱	۱۳۸۷	۱۷۷	۵۲۲	۸۸۸

† نتایج حساب تولید استان‌های کشور ــ ۱۳۷۹

استان	هتل و رستوران	حمل و نقل و ...	واسطه‌گری مالی و ...	کسب وکار	امور عمومی و ...	آموزش	بهداشت و ...	سایر خدمات
آذربایجان شرقی	۲۰۷	۹۶۶	۱۵۳	۱۷۲۹	۱۵۸۷	۷۲۱	۵۰۶	۱۶۰
آذربایجان غربی	۳۶۸	۲۰۶۹	۲۷۷	۳۲۰۲	۱۴۱۴	۱۰۱۹	۷۰۶	۲۶۸
اردبیل	۸۶	۴۸۸	۶۱	۶۰۶	۳۹۴	۳۷۲	۲۳۶	۷۶
اصفهان	۵۰۸	۲۵۶۵	۶۵۶	۴۸۶۲	۲۰۷۶	۱۴۶۸	۱۱۲۳	۳۹۸
ایلام	۲۵	۱۸۹	۳۶	۱۹۶	۳۰۹	۲۲۴	۱۳۰	۲۶
بوشهر	۵۶	۷۰۱	۹۳	۸۳۲	۶۴۹	۳۰۹	۲۱۰	۴۳
تهران	۱۹۵۰	۱۴۲۱۷	۸۲۵۶	۳۴۵۹۶	۱۱۶۰۱	۵۰۰۹	۴۸۶۷	۲۶۶۵
چهارمحال وبختیاری	۲۳	۲۸۷	۶۴	۴۱۵	۲۴۳	۳۰۹	۱۷۲	۴۲
خراسان	۶۴۷	۳۲۹۸	۵۳۱	۴۸۴۱	۲۵۵۸	۲۱۱۴	۱۴۶۵	۵۴۷
خوزستان	۲۸۱	۲۸۹۳	۳۴۸	۲۹۸۳	۲۱۸۳	۱۱۷۸	۹۹۱	۱۹۴
زنجان	۶۱	۲۵۱	۵۶	۵۱۶	۳۷۹	۳۰۸	۱۶۰	۶۷
سمنان	۲۵	۴۳۵	۷۷	۴۲۳	۲۲۹	۲۵۶	۱۶۲	۵۰
سیستان و بلوچستان	۳۶	۸۲۳	۱۰۶	۹۵۲	۸۷۶	۵۷۰	۳۰۸	۸۳
فارس	۳۰۴	۲۵۴۵	۴۹۳	۴۱۸۲	۱۷۱۶	۱۴۱۳	۹۲۷	۳۰۶
قزوین	۷۸	۴۲۵	۸۷	۹۳۲	۴۰۵	۳۲۷	۲۰۸	۴۳
قم	۹۰	۳۱۰	۶۷	۱۰۳۴	۳۴۰	۲۴۵	۳۰۳	۲۵۸
کردستان	۶۱	۵۶۱	۶۹	۶۹۹	۷۰۲	۳۸۸	۲۹۸	۷۶
کرمان	۱۲۱	۱۲۲۹	۱۸۵	۱۶۷۷	۱۰۴۲	۸۸۵	۴۷۶	۹۵
کرمانشاه	۱۲۶	۵۳۶	۱۰۲	۱۲۵۹	۱۳۷۰	۵۶۸	۳۶۱	۹۲
کهگیلویه و بویر احمد	۱۹	۱۳۶	۳۱	۲۳۳	۲۳۶	۲۷۷	۱۵۱	۳۳
گلستان	۵۲	۴۶۸	۱۰۵	۸۲۹	۶۴۳	۴۷۳	۲۹۱	۶۹
گیلان	۲۴۶	۱۱۸۰	۲۰۰	۱۷۶۰	۸۶۴	۸۱۱	۵۹۳	۱۴۴
لرستان	۵۶	۷۷۷	۷۷	۹۱۴	۷۳۳	۵۱۲	۲۸۳	۶۵
مازندران	۳۵۸	۲۱۲۷	۳۱۴	۲۵۰۲	۹۳۱	۱۰۵۱	۸۷۳	۲۹۹
مرکزی	۵۳	۱۰۸۱	۹۸	۱۰۵۲	۴۸۷	۳۹۲	۳۵۵	۱۰۵
هرمزگان	۷۲	۲۹۵۵	۱۲۳	۹۱۵	۶۲۵	۳۲۴	۲۷۰	۵۱
همدان	۱۰۵	۶۳۹	۹۷	۹۴۹	۶۳۵	۵۳۹	۳۵۳	۱۱۱
یزد	۵۲	۵۵۳	۱۴۲	۷۰۱	۳۹۶	۳۷۲	۳۴۶	۱۰۶

ارزش افزوده فعالیتهای اقتصادی کشور به تفکیک استان ــ ۱۳۷۹ (میلیارد ریال) (دنباله جدول ۱)

۳ بخش سوم

تعیین و بررسی تحلیلی جایگاه استانها از نظر درصد نیروی کار در فعالیتهای اقتصادی «بروش تحلیل عاملی»

۱.۳ تحلیل همبستگی

در این بخش نیز نظیر آنچه قبلاً گفته شده با استفاده از اطلاعات برآورد درصد شاغلان در هر یک از فعالیتهای اقتصادی استانها که در جدول ۹ آمده است، ماتریس ضرایب همبستگی بین متغیرها محاسبه گردید. در بررسی اولیه این ماتریس مشاهده شد که بین اکثر متغیرها همبستگی لازم وجود داشته و در تعداد معدودی از متغیرها به دلیل پایین بودن میزان همبستگی و با توجه به نتایج آزمونهای مربوطه نسبت به حذف آنها اقدام گردید[۱۵] و بررسی نهایی روی یازده متغیر باقیمانده صورت گرفت.

بمنظور ارزیابی کارایی استفاده از روش تحلیل عاملی در این بررسی از ملاکهای مطرح شده

۱۵) توضیح اینکه چهار متغیر حذف شده عبارتند از :
ساختمان ــ عمده‌فروشی، خرده‌فروشی و . . . ، هتل و رستوران و سایر فعالیتهای خدمات عمومی و اجتماعی و . . .

جدول ۲: محاسبه مقادیر MSA منتج شده از ماتریس همبستگی ارزش افزوده به تفکیک فعالیت

فعالیت	مقادیر MSA
کشاورزی و ...	۰٫۵۷۴
ماهیگیری	۶٫۵۶۸E-2
معدن	۹٫۸۸۴E-2
صنعت	۰٫۷۴۲
تامین آب وبرق و ...	۰٫۸۲۲
ساختمان	۰٫۸۰۰
عمده فروشی و ...	۰٫۸۰۰
هتل و رستوران	۰٫۸۷۲
حمل و نقل و ...	۰٫۸۹۰
واسطه‌گری مالی و ...	۰٫۸۲۵
کسب و کار و ...	۰٫۸۴۸
امور عمومی و ...	۰٫۸۶۷
آموزش ...	۰٫۷۹۸
بهداشت و ...	۰٫۹۰۷
سایر خدمات	۰٫۸۸۰

جدول ۳: مقادیر ویژه و درصد واریانس و درصد تجمعی واریانس تبیین شده توسط هر یک از عاملها

عاملهای اولیه	مقادیر ویژه	درصد واریانس	درصد واریانس تجمعی
۱	۱۱٫۴۶۰	۷۶٫۳۹۷	۷۶٫۳۹۶
۲	۱٫۲۸۳	۸٫۵۵۴	۸۴٫۹۵۱
۳	۱٫۰۱۰	۶٫۷۳۶	۹۱٫۶۸۷
۴	۰٫۸۳۹	۵٫۵۹۳	۹۷٫۲۸۰
۵	۰٫۲۷۸	۱٫۸۵۵	۹۹٫۱۳۵
۶	۵٫۳۶۰E-2	۰٫۳۵۷	۹۹٫۴۹۲
۷	۲٫۰۲۷E-2	۰٫۱۳۵	۹۹٫۶۲۷
۸	۱٫۹۱۸E-2	۰٫۱۲۸	۹۹٫۷۵۵
.	.	.	.
.	.	.	.
.	.	.	.
۱۵	۳٫۲۸۶E-۴	۲٫۱۹۱E-۳	۱۰۰٫۰۰۰

جدول ۴: ماتریس ضرایب عامل دوران یافته بروش واریماکس به تفکیک فعالیت (در تحلیل ارزش افزوده)

فعالیت	عامل ۱	عامل ۲	عامل ۳
کشاورزی و ...	۰٫۲۱۸	۰٫۸۶۸	-۵٫۰۶۰E-۳
ماهیگیری	-۲٫۱۵۳E-۲	-۰٫۳۵۱	۰٫۷۴۵
معدن	-۲٫۶۶۶E-۲	۰٫۳۰۴	۰٫۷۵۴
صنعت	۰٫۹۲۳	۰٫۲۴۵	۰٫۱۱۵
تامین آب و برق و ...	۰٫۷۸۰	۰٫۳۸۰	۰٫۳۰۵
ساختمان	۰٫۹۶۸	۰٫۲۰۷	-۳٫۲۸۳E-۲
عمده فروشی و ...	۰٫۹۸۷	۰٫۱۳۰	-۲٫۷۹۱E-۲
هتل ورستوران	۰٫۹۷۲	۰٫۱۹۲	-۴٫۲۲۳E-۲
حمل ونقل و ...	۰٫۹۸۱	۹٫۵۴۵E-۲	۷٫۹۰۹E-۲
واسطه‌گری مالی و ...	۰٫۹۸۳	-۳٫۱۸۹E-۲	-۷٫۳۹۹E-۲
کسب و کار	۰٫۹۹۱	۴٫۹۴۱E-۲	-۶٫۳۹۳E-۲
امور عمومی و ...	۰٫۹۸۴	۰٫۱۰۷	-۱٫۸۸۲E-۲
آموزش	۰٫۹۴۸	۰٫۲۸۷	-۲٫۵۵۷E-۲
بهداشت و ...	۰٫۹۸۱	۰٫۱۸۰	-۹٫۰۰۵E-۲
سایر خدمات	۰٫۹۸۸	۵٫۸۵۲E-۲	-۹٫۴۹۴E-۲

جدول ۵: ماتریس ضرایب نهایی عوامل به تفکیک فعالیت بروش رگرسیونی (در تحلیل ارزش افزوده)

فعالیت	عامل ۱	عامل ۲	عامل ۳
کشاورزی و ...	-۰/۱۱۷	۰/۷۸۰	-۰/۰۵۳
ماهیگیری	۰/۰۶۳	-۰/۳۷۰	۰/۶۱۵
معدن	-۰/۰۴۶	۰/۲۴۵	۰/۵۸۴
صنعت	۰/۰۷۱	۰/۰۷۱	۰/۰۸۷
تامین آب و برق و ...	۰/۰۳۳	۰/۲۱۰	۰/۲۲۹
ساختمان	۰/۰۸۱	۰/۰۳۶	-۰/۰۲۸
عمده‌فروشی و ...	۰/۰۹۶	-۰/۰۳۹	-۰/۰۲۰
هتل و رستوران	۰/۰۸۴	۰/۰۲۲	-۰/۰۳۵
حمل و نقل و ...	۰/۱۰۲	-۰/۰۷۷	۰/۰۶۷
واسطه‌گری مالی و ...	۰/۱۲۲	-۰/۱۸۸	-۰/۰۴۷
کسب و کار	۰/۱۰۹	-۰/۱۱۴	-۰/۰۴۴
امور عمومی و ...	۰/۱۰۰	-۰/۰۶۱	-۰/۰۱۱
آموزش	۰/۰۶۵	۰/۱۱۵	-۰/۰۲۸
بهداشت و ...	۰/۰۸۷	۰/۰۰۸	-۰/۰۰۸
سایر خدمات	۰/۱۰۷	-۰/۱۰۳	-۰/۰۶۹

استفاده و نتایج آن بشرح زیر می‌باشد.

دترمینان ماتریس همبستگی با مقدار $۴/۸۷۹ \times ۱۰^{-۵}$ بوده و بعلت کوچک بودن مقدار بدست آمده، استفاده از روش تحلیل عاملی تایید شد.

بررسی مقدار MSAها باستثنای بخش صنعت که خیلی قوی نیست بر حضور سایر متغیرها تاکید دارد، اما بدلیل اهمیت بخش صنعت، حضور آن در مدل الزامی به نظر رسید.

مقدار شاخص KMO بدست آمده نیز تقریباً معادل ۰/۷ می‌باشد که مقدار این شاخص نیز استفاده از روش تحلیل عاملی را تایید می‌کند.

۲.۳ تعیین تعداد و روش محاسبه عاملها و میزان برازش مدل

برای تعیین تعداد عاملها در این بررسی با توجه به مقادیر ویژه بدست آمده در جدول ۱۲ و میزان واریانس تبیین شده توسط هر یک از عاملهای اولیه مشاهده شد که عاملهای اول تا سوم بدلیل آنکه دارای بیشترین مقادیر ویژه و با تبیین حدود ۸۰ درصد واریانس کل، بعنوان عاملهای منتخب پذیرفته شدند. در ادامه برای سهولت در تفسیر عاملها، بااستفاده از دوران متعامد ماتریس ضرایب عاملها برای هریک از سه عامل محاسبه و نتایج آن در جدول ۱۳ آمده است.

در تعبیر و تفسیر هر یک از عاملها با استفاده از مقادیر جدول ۱۳ بشرح زیر داریم:

جدول ۶: رتبه‌بندی ارزش افزوده استانها بر حسب عامل اول ــ ۱۳۷۹ (تاکید بر اهمیت بخشهای غیر از کشاورزی، معدن و ماهیگیری)

رتبه	استان	امتیاز کسب شده
۱	تهران	۴٫۹۲
۲	اصفهان	۰٫۵۹
۳	خراسان	۰٫۴۶
۴	خوزستان	۰٫۱۳
۵	آذربایجان غربی	۰٫۰۵
۶	هرمزگان	۰٫۰۳
۷	مازندران	۰٫۰۳
۸	گیلان	۰٫۰۲
۹	فارس	۰٫۰۱
۱۰	بوشهر	-۰٫۱۵۲
۱۱	آذربایجان شرقی	-۰٫۱۹۲
۱۲	سیستان و بلوچستان	-۰٫۱۹۴
۱۳	مرکزی	-۰٫۱۹۶
۱۴	کرمانشاه	-۰٫۲۴۱
۱۵	قم	-۰٫۲۶۳
۱۶	یزد	-۰٫۳۰۷
۱۷	گلستان	-۰٫۳۰۸
۱۸	کردستان	-۰٫۳۱۹
۱۹	همدان	-۰٫۳۳۲
۲۰	قزوین	-۰٫۳۴۶
۲۱	لرستان	-۰٫۳۵۷
۲۲	اردبیل	-۰٫۳۸۹
۲۳	سمنان	-۰٫۳۹۱
۲۴	زنجان	-۰٫۳۹۲
۲۵	چهارمحال و بختیاری	-۰٫۴۰۰
۲۶	ایلام	-۰٫۴۱۰
۲۷	کرمان	-۰٫۵۱۴
۲۸	کهگیلویه و بویر احمد	-۰٫۵۲۶

جدول ۷: رتبه‌بندی ارزش افزوده استانها بر حسب عامل دوم ــ ۱۳۷۹ (تاکید بر اهمیت بخش کشاورزی)

رتبه	استان	امتیاز کسب شده
۱	کرمان	۲/۳۱
۲	خراسان	۱/۹۴
۳	خوزستان	۱/۷۷
۴	فارس	۱/۶۱
۵	اصفهان	۱/۰۹
۶	آذربایجان غربی	۰/۷۵
۷	مازندران	۰/۵۸
۸	قزوین	۰/۲۷
۹	همدان	۰/۱۹
۱۰	آذربایجان شرقی	۰/۱۵
۱۱	کهگیلویه و بویراحمد	-۰/۰۹
۱۲	لرستان	-۰/۱۳
۱۳	اردبیل	-۰/۲۱
۱۴	مرکزی	-۰/۲۲
۱۵	کرمانشاه	-۰/۳۱
۱۶	زنجان	-۰/۳۷
۱۷	گلستان	-۰/۴۰
۱۸	یزد	-۰/۴۵
۱۹	گیلان	-۰/۴۸
۲۰	سمنان	-۰/۵۰
۲۱	تهران	-۰/۵۳
۲۲	کردستان	-۰/۵۵
۲۳	چهارمحال و بختیاری	-۰/۵۷
۲۴	قم	-۰/۷۱
۲۵	ایلام	-۰/۷۷
۲۶	سیستان و بلوچستان	-۱/۰۸
۲۷	بوشهر	-۱/۵۰
۲۸	هرمزگان	-۱/۸۰

جدول ۸: رتبه‌بندی ارزش افزوده استانها بر حسب عامل سوم ــ ۱۳۷۹ (تاکید بر اهمیت بخشهای معدن و ماهیگیری)

رتبه	استان	امتیاز کسب شده
۱	خوزستان	۳٫۸۱
۲	هرمزگان	۲٫۱۳
۳	بوشهر	۰٫۹۶
۴	کهگیلویه و بویر احمد	۰٫۹۵
۵	گیلان	۰٫۷۶
۶	مازندران	۰٫۵۲
۷	سیستان و بلوچستان	۰٫۳۲
۸	گلستان	۰٫۲۲
۹	اصفهان	۰٫۰۷
۱۰	قزوین	-۰٫۳۲
۱۱	مرکزی	-۰٫۴۰
۱۲	آذربایجان شرقی	-۰٫۴۱
۱۳	آذربایجان غربی	-۰٫۴۳
۱۴	فارس	-۰٫۴۶
۱۵	تهران	-۰٫۴۷
۱۶	همدان	-۰٫۴۸
۱۷	ایلام	-۰٫۴۹
۱۸	یزد	-۰٫۵۲
۱۹	خراسان	-۰٫۵۴
۲۰	چهارمحال و بختیاری	-۰٫۵۵۱
۲۱	زنجان	-۰٫۵۵۴
۲۲	کرمانشاه	-۰٫۵۶۲
۲۳	لرستان	-۰٫۵۶۴
۲۴	سمنان	-۰٫۵۶۶
۲۵	کردستان	-۰٫۵۷۱
۲۶	قم	-۰٫۵۹۸
۲۷	اردبیل	-۰٫۵۹۹
۲۸	کرمان	-۰٫۶۴۹

ویژگی عامل اول: این عامل عمدتاً بر فعالیتهای ماهیگیری، حمل و نقل، تامین آب و برق و امور عمومی تاکید دارد و دارای ضرایب منفی نسبتاً بزرگ برای فعالیت صنعت می‌باشد. به بیان دیگر، استان‌های که تحت این عامل از رتبه بالایی برخوردارند سهم اشتغال بخش صنعت در آنها ضعیف است.
ویژگی عامل دوم: این عامل بر اهمیت فعالیتهای بهداشت و مددکاری اجتماعی و . . . ، آموزش و معدن تاکید می‌کند.
ویژگی عامل سوم: فعالیتهایی که توسط این عامل بر اهمیت آنها تاکید شده عبارتند از کسب و کار، واسطه‌گریهای مالی و صنعت می‌باشد و بعکس با ضریب منفی نسبت به فعالیت کشاورزی کمترین میزان اهمیت را در مورد فعالیت کشاورزی نشان می‌دهد.

۳.۳ جداول نهایی

پس از تعیین تعداد و ویژگیهای هر عامل امتیاز نهایی هر یک از استانها را بر اساس فرمول زیر محاسبه می‌نماییم که پس از مرتب نمودن امتیازات مذکور به صورت نزولی در جدول ۱۵ و ۱۶ و ۱۷ نشان داده شده است.

$$B_{jk} = \sum_{i=۱}^{۱۱} w_{ji} z_{ik} \qquad j = ۱, ۲, \ldots, ۲۸ \qquad k = ۱, ۲, ۳$$

۴.۳ نتیجه‌گیری

در بررسی جدول ۱۵ مشاهده می‌شود که تحت عامل اول استان‌های هرمزگان، بوشهر، و خوزستان در ارتباط با اهمیت و تمرکز نیروی کار در رشته فعالیتهای ماهیگیری، حمل و نقل، تامین آب و برق و . . . ، امور عمومی بترتیب بالاترین رتبه را کسب نموده و در مقابل استان‌های یزد، زنجان در این خصوص در پایین‌ترین رتبه قرار داشته‌اند.

برای دسته‌بندی استان‌های همگن تحت عامل اول می‌توان آنها را در سه گروه فرضی بشکل زیر تقسیم نمود:

- گروه اول: هرمزگان، بوشهر، خوزستان
- گروه دوم: سیستان و بلوچستان، کرمانشاه، فارس، آذربایجان
- گروه سوم: سایر استانها

پس از مرتب نمودن استانها در ارتباط با عامل دوم ملاحظه می‌شود که استان‌های کهگیلویه و بویر احمد، سمنان، کرمان، ایلام و خوزستان در بخشهای بهداشت و مددکاری و . . . ، آموزش و معدن دارای بالاترین رتبه بوده‌اند و آذربایجان شرقی و غربی و قزوین در این خصوص کمترین رتبه را داشته‌اند.

تحت این عامل استان‌های تقریباً همگن عبارتند از:
ـ گروه اول: کهگیلویه و بویر احمد، سمنان، کرمان، ایلام
ـ گروه دوم: خوزستان، بوشهر، تهران، یزد، فارس، کرمانشاه، لرستان، مازندران
ـ گروه سوم: سایر استانها
ـ گروه چهارم: مرکزی، آذربایجان غربی و شرقی، قزوین

در تحلیل استانها بر اساس عامل سوم که بر اهمیت درصد نیروی کار در بخشهای صنعت، کسب و کار و واسطه‌گریهای مالی تاکید دارد نشان می‌دهد که استان‌های تهران، یزد، اصفهان و قم بترتیب بالاترین رتبه را در این خصوص دارا بوده و در مقابل استان‌های ایلام، کهگیلویه و بویر احمد و اردبیل پایین‌ترین رتبه را داشته‌اند.

طبقه‌بندی استان‌های تقریباً همگن تحت عامل سوم عبارت است از:
ـ گروه اول: تهران، یزد، اصفهان، قم
ـ گروه دوم: بوشهر، سمنان، مرکزی، آذربایجان شرقی، خوزستان، هرمزگان، چهار محال و بختیاری
ـ گروه سوم: قزوین، فارس، کرمان، مازندران، لرستان، زنجان، گیلان، همدان
ـ گروه چهارم: سایر استانها

۵.۳ نتیجه‌گیری کلی

ـ بر اساس رتبه‌های کسب شده استانها در خصوص دو شاخص ارزش افزوده و درصد نیروی کار در فعالیتهای اقتصادی، بارزترین نکته وضعیت استانها در دو بخش صنعت و خدمات بوده بطوریکه ملاحظه می شود استان تهران در ارزیابی این دو بخش فاصله زیادی نسبت به سایر استانها داشته است. به این ترتیب با تجمع و تمرکز سهم زیادی از رشته فعالیتهای عمده صنعتی و خدماتی عملاً منجر به جذب سرمایه و نیروی کار و در نهایت ایجاد شکاف اقتصادی با دیگر استانها شده است. بعنوان بارزترین مثال در بیان این موضوع نگاهی به سرانه تولید ناخالص داخلی نشان می‌دهد که در سال ۱۳۷۹ مقدار این شاخص در کشور معادل ۱۰٫۱۴ میلیون ریال بوده و در استان تهران برابر ۱۳٫۸۶ میلیون ریال می‌باشد و برای مقایسه بهتر چنانچه ارزش افزوده فعالیت استخراج نفت و گاز طبیعی که جنبه ملی دارد را از سر جمع تولید ناخالص داخلی کشور و استان تهران کسر نماییم، آنگاه سرانه مورد نظر در کشور با ۱٫۷۴ واحد کاهش به ۸٫۴۰ میلیون ریال نزول می‌کند در حالیکه این سرانه در استان تهران فقط با کسر ۰٫۰۲ واحد به ۱۳٫۸۴ میلیون ریال می‌رسد. بهر حال با توجه به اینکه به تجربه دیده شده بخش صنعت بعنوان یکی از ضرورتهای توسعه اقتصادی پایدار که ناشی از وجود رابطه خطی مثبت سهم این بخش، از GDP (GDPR) با تولید ناخالص داخلی سرانه می‌باشد، همواره مطرح بوده است. بنابراین هر چه سهم این بخش افزایش یابد، دسترسی به توسعه پایدار در منطقه تسهیل شده و لذا به منظور کاهش بخشی از عدم تعادل‌ها که این امر لزوم توجه جدی به تقویت بخش صنعت بخصوص در

جدول ۹: برآورد درصد شاغلان† هر یک از فعالیتهای اقتصادی‡ به تفکیک استان ــ ۱۳۷۹

استان	کشاورزی	ماهیگیری	معدن	صنعت	تامین آب و برق و ...	ساختمان	عمده فروشی و ...
آذربایجان شرقی	۲۵٫۳۸	۰٫۰۰	۰٫۱۴	۲۸٫۵۸	۰٫۶۵	۹٫۹۷	۱۲٫۱۲
آذربایجان غربی	۳۷٫۷۲	۰٫۰۳	۰٫۱۷	۱۲٫۶۷	۰٫۶۸	۱۰٫۷۰	۱۲٫۱۸
اردبیل	۳۷٫۶۸	۰٫۰۱	۰٫۱۸	۱۰٫۸۰	۰٫۶۷	۱۳٫۹۱	۱۲٫۰۴
اصفهان	۱۳٫۵۲	۰٫۰۱	۰٫۶۵	۳۲٫۲۴	۱٫۱۳	۱۰٫۳۶	۱۳٫۲۶
ایلام	۳۶٫۳۵	۰٫۰۰	۰٫۱۵	۵٫۷۱	۰٫۸۷	۱۱٫۶۱	۹٫۳۶
بوشهر	۱۳٫۳۵	۵٫۷۴	۲٫۰۳	۷٫۸۰	۱٫۶۲	۱۰٫۲۸	۱۱٫۲۷
تهران	۳٫۶۸	۰٫۰۱	۰٫۸۳	۲۳٫۹۴	۰٫۹۸	۹٫۴۰	۱۹٫۳۷
چهارمحال بختیاری	۲۳٫۴۷	۰٫۰۷	۰٫۲۸	۲۱٫۶۱	۰٫۶۵	۱۸٫۷۵	۸٫۵۶
خراسان	۲۹٫۸۰	۰٫۰۰	۰٫۲۸	۱۹٫۷۸	۰٫۷۴	۱۰٫۵۳	۱۲٫۰۴
خوزستان	۲۰٫۱۱	۱٫۴۸	۳٫۴۱	۱۴٫۲۲	۲٫۸۶	۱۰٫۸۸	۱۲٫۷۸
زنجان	۳۶٫۸۲	۰٫۰۰	۰٫۷۶	۲۲٫۱۰	۰٫۵۹	۹٫۰۲	۸٫۴۳
سمنان	۲۱٫۷۷	۰٫۰۰	۲٫۸۰	۱۵٫۷۵	۱٫۱۴	۹٫۶۲	۹٫۷۹
سیستان و بلوچستان	۲۷٫۳۵	۱٫۹۸	۰٫۳۴	۸٫۰۰	۰٫۹۷	۱۵٫۲۹	۱۱٫۶۸
فارس	۲۴٫۹۷	۰٫۰۱	۰٫۵۷	۱۲٫۷۵	۱٫۰۶	۱۳٫۰۰	۱۳٫۹۶
قزوین	۳۰٫۵۹	۰٫۰۳	۰٫۱۷	۲۵٫۴۵	۱٫۶۸	۶٫۲۶	۱۰٫۷۷
قم	۸٫۳۲	۰٫۰۰	۰٫۵۴	۲۸٫۰۰	۰٫۷۰	۱۲٫۸۶	۱۶٫۲۹
کردستان	۳۲٫۷۲	۰٫۱۰	۰٫۲۵	۱۳٫۲۵	۰٫۶۱	۱۳٫۳۸	۱۳٫۴۰
کرمان	۳۰٫۱۲	۰٫۰۰	۲٫۲۵	۱۰٫۵۹	۱٫۰۰	۱۱٫۲۵	۱۰٫۷۹
کرمانشاه	۲۹٫۳۴	۰٫۰۲	۰٫۳۶	۷٫۷۲	۱٫۰۲	۱۲٫۲۶	۱۳٫۳۴
کهگیلویه و بویر احمد	۳۰٫۰۴	۰٫۰۱	۲٫۵۶	۴٫۷۳	۱٫۰۸	۱۸٫۶۰	۸٫۵۰
گلستان	۳۹٫۹۵	۰٫۷۷	۰٫۵۴	۱۶٫۹۱	۰٫۷۳	۷٫۹۹	۱۰٫۲۳
گیلان	۳۷٫۳۶	۰٫۷۱	۰٫۲۱	۱۳٫۶۶	۰٫۸۲	۵٫۱۲	۱۳٫۷۳
لرستان	۲۹٫۲۲	۰٫۰۱	۰٫۲۸	۸٫۴۶	۰٫۶۶	۱۸٫۶۳	۱۱٫۶۹
مازندران	۳۶٫۷۷	۰٫۲۰	۰٫۴۸	۱۱٫۲۸	۰٫۸۲	۹٫۲۹	۱۲٫۷۱
مرکزی	۲۵٫۹۳	۰٫۰۰	۰٫۷۲	۲۷٫۰۵	۰٫۹۰	۱۱٫۰۲	۱۰٫۳۳
هرمزگان	۱۷٫۹۱	۷٫۰۰	۰٫۴۹	۶٫۰۷	۲٫۶۸	۱۲٫۴۹	۱۲٫۱۶
همدان	۳۲٫۱۳	۰٫۰۰	۰٫۱۹	۱۳٫۹۲	۱٫۰۷	۱۳٫۷۴	۱۳٫۳۷
یزد	۱۳٫۲۹	۰٫۰۲	۱٫۴۱	۳۰٫۳۸	۰٫۸۲	۱۱٫۱۹	۱۱٫۵۹

† برآورد بر اساس نتایج سرشماری نفوس و مسکن ۱۳۷۵ و نتایج طرح آمارگیری از ویژگیهای اشتغال و بیکاری کشور ــ ۱۳۷۹

‡ توضیح اینکه برخی ارقام مندرج در ستون فعالیت ماهیگیری بعلت گرد شدن بصورت صفر نمایش داده شده است.

جدول ۱۰: برآورد درصد شاغلان هر یک از فعالیتهای اقتصادی به تفکیک استان ـ ۱۳۷۹ (دنباله)

استان	هتل و رستوران	حمل و نقل و ...	واسطه‌گری مالی	کسب وکار	امور عمومی و ...	آموزش	بهداشت و ...	سایر خدمات
آذربایجان شرقی	۰٫۷۵	۶٫۲۴	۰٫۶۴	۰٫۵۹	۶٫۳۱	۵٫۱۳	۱٫۴۱	۱٫۴۳
آذربایجان غربی	۰٫۵۵	۵٫۸۴	۰٫۵۹	۰٫۶۰	۹٫۰۲	۵٫۳۶	۱٫۵۱	۱٫۷۰
اردبیل	۱٫۰۷	۵٫۷۳	۰٫۶۶	۰٫۴۸	۵٫۹۳	۶٫۳۲	۱٫۵۰	۲٫۰۸
اصفهان	۰٫۶۱	۶٫۷۳	۰٫۹۶	۰٫۹۴	۸٫۴۸	۷٫۰۲	۱٫۹۷	۱٫۵۰
ایلام	۰٫۴۷	۵٫۶۳	۰٫۹۴	۰٫۳۹	۱۳٫۸۳	۹٫۶۱	۲٫۲۸	۱٫۶۸
بوشهر	۰٫۳۳	۱۳٫۵۰	۱٫۴۴	۱٫۷۴	۱۷٫۶۳	۸٫۴۸	۲٫۴۱	۱٫۷۱
تهران	۰٫۸۰	۸٫۸۷	۲٫۲۰	۲٫۵۴	۱۲٫۹۲	۶٫۸۸	۲٫۷۱	۳٫۵۱
چهارمحال و بختیاری	۰٫۳۵	۶٫۰۴	۰٫۶۸	۰٫۵۹	۶٫۸۳	۸٫۶۶	۱٫۹۲	۰٫۷۷
خراسان	۰٫۵۸	۵٫۷۳	۰٫۸۹	۰٫۸۳	۶٫۸۶	۶٫۸۹	۱٫۶۵	۲٫۵۶
خوزستان	۰٫۶۱	۷٫۶۷	۱٫۰۱	۱٫۰۸	۱۱٫۶۰	۷٫۱۷	۲٫۵۴	۱٫۸۶
زنجان	۰٫۵۱	۴٫۶۲	۰٫۶۰	۰٫۴۵	۶٫۱۴	۶٫۵۴	۱٫۳۱	۱٫۶۴
سمنان	۰٫۴۶	۷٫۱۷	۱٫۷۸	۰٫۴۶	۱۴٫۴۵	۹٫۳۳	۲٫۶۵	۲٫۲۱
سیستان و بلوچستان	۰٫۴۲	۸٫۲۱	۰٫۷۸	۰٫۴۲	۱۲٫۲۳	۷٫۳۶	۱٫۹۳	۲٫۰۷
فارس	۰٫۵۷	۹٫۱۸	۰٫۸۳	۰٫۶۶	۸٫۹۷	۸٫۵۶	۲٫۱۷	۱٫۹۴
قزوین	۰٫۵۹	۷٫۶۹	۰٫۷۰	۰٫۵۶	۵٫۰۹	۷٫۶۱	۱٫۳۴	۱٫۴۹
قم	۰٫۹۷	۷٫۲۴	۰٫۷۸	۱٫۰۶	۷٫۹۶	۸٫۰۰	۱٫۸۰	۳٫۷۷
کردستان	۰٫۶۵	۵٫۷۶	۰٫۵۷	۰٫۵۴	۸٫۴۴	۶٫۴۶	۱٫۷۰	۱٫۶۶
کرمان	۰٫۴۱	۶٫۵۶	۱٫۱۲	۰٫۶۷	۱۰٫۲۲	۹٫۸۱	۲٫۴۸	۱٫۷۹
کرمانشاه	۰٫۴۳	۷٫۹۹	۰٫۷۳	۰٫۷۴	۱۳٫۴۱	۷٫۲۱	۲٫۲۶	۱٫۶۸
کهگیلویه و بویراحمد	۰٫۳۵	۴٫۸۲	۰٫۹۳	۰٫۵۳	۱۰٫۷۹	۱۲٫۱۳	۳٫۰۹	۱٫۰۹
گلستان	۰٫۴۴	۵٫۰۴	۰٫۸۳	۰٫۳۹	۴٫۷۱	۷٫۵۱	۱٫۶۴	۲٫۳۲
گیلان	۱٫۱۳	۶٫۵۹	۰٫۸۸	۰٫۷۹	۶٫۵۸	۶٫۹۴	۲٫۲۶	۲٫۳۳
لرستان	۰٫۴۴	۶٫۵۲	۰٫۸۶	۰٫۸۱	۱۰٫۲۳	۷٫۷۳	۱٫۹۴	۱٫۳۷
مازندران	۰٫۶۳	۶٫۷۰	۰٫۸۸	۰٫۹۸	۶٫۵۸	۷٫۳۳	۲٫۳۲	۲٫۲۸
مرکزی	۰٫۴۱	۶٫۲۹	۰٫۹۷	۰٫۵۱	۵٫۷۳	۶٫۰۵	۱٫۵۵	۱٫۷۷
هرمزگان	۰٫۶۹	۱۲٫۱۳	۱٫۰۵	۱٫۲۹	۱۲٫۸۸	۶٫۲۲	۲٫۷۰	۳٫۰۳
همدان	۰٫۵۱	۶٫۶۲	۰٫۸۳	۰٫۴۰	۷٫۵۰	۶٫۲۴	۱٫۷۲	۱٫۳۸
یزد	۰٫۵۰	۷٫۵۹	۱٫۲۶	۰٫۸۱	۷٫۷۷	۸٫۲۳	۲٫۳۹	۱٫۹۶

جدول ۱۱: محاسبه مقادیر MSA منتج شده از ماتریس همبستگی درصد شاغلین به تفکیک فعالیت

فعالیت	مقادیر MSA
کشاورزی و ...	۰٫۵۵۷
ماهیگیری	۰٫۸۰۹
معدن	۰٫۷۱۴
صنعت	۰٫۳۴۵
تامین آب و برق و ...	۰٫۶۹۹
حمل ونقل و ...	۰٫۷۵۸
واسطه‌گری مالی	۰٫۶۶۴
کسب وکار	۰٫۷۵۴
امور عمومی و ...	۰٫۶۹۰
بهداشت و ...	۰٫۷۲۵
آموزش	۰٫۶۱۰

جدول ۱۲: مقادیر ویژه و درصد واریانس و درصد تجمعی واریانس تبیین شده توسط هر یک از عاملها

عاملهای اولیه	مقادیر ویژه	درصد واریانس	درصد واریانس تجمعی
۱	۴٫۹۰۵	۴۴٫۵۹۲	۴۴٫۵۹۲
۲	۲٫۰۴۰	۱۸٫۵۴۹	۶۳٫۱۴۱
۳	۱٫۷۳۶	۱۵٫۷۸۴	۷۸٫۹۲۵
۴	۰٫۸۹۴	۷٫۷۱۴	۶۳۹٫۸۶
۵	۰٫۴۵۳	۴٫۱۱۹	۹۰٫۷۵۸
.	.	.	.
.	.	.	.
.	.	.	.
۱۱	۴٫۰۱۶E-۲	۰٫۳۶۵	۱۰۰٫۰۰۰

جدول ۱۳: ماتریس ضرایب عامل دوران یافته بروش واریماکس به تفکیک فعالیت برای درصد شاغلان

فعالیت	عامل ۱	عامل ۲	عامل ۳
کشاورزی و ...	-۰٫۲۳۱	-۰٫۱۵۸	-۰٫۸۹۰
ماهیگیری	۰٫۹۲۹	-۱٫۲۲۱E-۲	۶٫۶۵۴E-۲
معدن	۰٫۱۲۰	۰٫۷۷۰	۰٫۲۱۰
صنعت	-۰٫۵۱۹	-۰٫۴۲۱	۰٫۶۴۷
تامین آب و برق و ...	۰٫۷۳۰	۰٫۱۸۸	۰٫۱۴۹
حمل و نقل و ...	۰٫۸۴۷	۵٫۰۷۴E-۲	۰٫۴۰۱
واسطه‌گری مالی	۰٫۱۶۵	۰٫۵۱۵	۰٫۷۰۸
کسب وکار	۰٫۴۳۳	۹٫۲۲۰E-۲	۰٫۷۵۴
امور عمومی و ...	۰٫۶۱۳	۰٫۵۸۶	۰٫۱۶۳
بهداشت و ...	۰٫۳۵۷	۰٫۸۲۴	۰٫۲۲۷
آموزش	-۰٫۱۳۱	۰٫۸۷۶	-۰٫۱۰۷

جدول ۱۴: ماتریس ضرایب نهایی عوامل بروش رگرسیونی به تفکیک فعالیت برای درصد شاغلان

فعالیت	عامل ۱	عامل ۲	عامل ۳
کشاورزی و ...	۰٫۰۴۰	۰٫۰۱۷	-۰٫۳۶۳
ماهیگیری	۰٫۳۷۰	-۰٫۱۳۴	۰٫۰۸۷
معدن	-۰٫۰۸۱	۰٫۲۹۱	۰٫۰۳۵
صنعت	-۰٫۲۳۶	-۰٫۱۴۲	۰٫۳۸۱
تامین آب و برق و ...	۰٫۲۵۲	-۰٫۰۲۹	-۰٫۰۳۶
حمل و نقل و ...	۰٫۲۸۶	-۰٫۱۱۶	۰٫۰۷۱
واسطه‌گریهای مالی	-۰٫۰۹۲	۰٫۱۵۱	۰٫۲۶۸
کسب و کار	۰٫۰۶۸	-۰٫۰۶۴	۰٫۲۸۰
امور عمومی و ...	۰٫۱۴۹	۰٫۱۵۱	-۰٫۰۳۸
بهداشت و ...	۰٫۰۰۵	۰٫۲۸۰	۰٫۰۱۰
آموزش	-۰٫۱۵۷	۰٫۳۸۷	-۰٫۰۸۲

استانهایی که پتانسیل بیشتری در این خصوص دارند را طلب می‌نماید.

ـ مقایسه برخی رتبه‌های کسب شده در استانها بیانگر عدم تطابق الگوی ارزش افزوده فعالیتها در ارتباط با درصد نیروی کار در فعالیتهای متناظر می‌باشد. به بیان دیگر بر خلاف آنکه انتظار می‌رفت یک ارتباط نسبتاً قوی بین دو شاخص مذکور وجود داشته باشد (باستثنای بخشهایی که بر تکنولوژی‌های نوین که از نیروی انسانی کمتری استفاده می‌کنند)، ملاحظه نمی‌شود.

ـ برای بررسی این موضوع، طی بررسی تحلیل آماری همبستگی و آزمون‌های مربوطه بین این دو شاخص نتایجی بشرح زیر بدست آمد:

در بخش ماهیگیری مقدار همبستگی بدست آمده حدود ۰٫۷ بوده و نتایج آزمون همبستگی، وجود ارتباط بین دو متغیر ارزش افزوده و درصد نیروی شاغلین را در این بخش تایید می‌کند.

یکی از نکات جالب بدست آمده آن بود که رابطه همبستگی بین دو شاخص مورد نظر در بخشهای بهداشت و درمان و ساختمان منفی و بترتیب حدود ۰٫۱۵− و ۰٫۳− بوده و در بخش آموزش حدود ۰٫۱ می‌باشد که البته نتایج آزمون در سطح ۵ درصد همبستگی دو شاخص را رد می‌نماید.

این بررسی در مورد سایر بخشها نیز بیانگر یک ارتباط ضعیف ولی مثبت بوده، که مقدار همبستگی در آنها مجموعاً بین ۰٫۱ تا ۰٫۶ در نوسان بوده ودر سطح ۵ درصد تایید می‌شوند.

موارد مذکور در واقع به نوعی مؤید پایین بودن میزان بهره‌وری درکشور و عدم بکارگیری مفید و بهینه از ظرفیتهای موجود درکلیه استانها بوده و لزوم بررسی همه جانبه آن از ضروری‌ترین

جدول ۱۵: رتبه‌بندی استانها بر حسب عامل اول برای درصد نیروی شاغل در فعالیتهای اقتصادی ــ ۱۳۷۹ (تاکید بر اهمیت بخشهای ماهیگیری، حمل و نقل، تامین آب و . . . ، امور عمومی و . . .)

رتبه	استان	امتیاز کسب شده
۱	هرمزگان	۳٫۴۹
۲	بوشهر	۲٫۶۷
۳	خوزستان	۱٫۰۶
۴	سیستان و بلوچستان	۰٫۸۴
۵	کرمانشاه	۰٫۵۵
۶	فارس	۰٫۱۷
۷	آذربایجان غربی	۰٫۰۰
۸	همدان	-۰٫۰۲۸
۹	گیلان	-۰٫۰۲۸
۱۰	تهران	-۰٫۰۴۵
۱۱	لرستان	-۰٫۰۵۹
۱۲	مازندران	-۰٫۰۹۳
۱۳	ایلام	-۰٫۱۰۱
۱۴	قزوین	-۰٫۱۲۵
۱۵	کردستان	-۰٫۲۱۶
۱۶	اردبیل	-۰٫۲۵۰
۱۷	کرمان	-۰٫۴۴۶
۱۸	سمنان	-۰٫۴۷۸
۱۹	خراسان	-۰٫۵۴۱
۲۰	آذربایجان شرقی	-۰٫۵۷۸
۲۱	گلستان	-۰٫۵۸۷
۲۲	اصفهان	-۰٫۶۱۱
۲۳	مرکزی	-۰٫۶۷۹
۲۴	قم	-۰٫۶۹۳
۲۵	کهگیلویه و بویر احمد	-۰٫۷۰۵
۲۶	چهارمحال و بختیاری	-۰٫۷۶۶
۲۷	زنجان	-۰٫۸۳۰
۲۸	یزد	-۰٫۹۲۰

جدول ۱۶: رتبه‌بندی استانها بر حسب عامل دوم برای درصد نیروی شاغل در فعالیتهای اقتصادی ــ ۱۳۷۹ (تاکید بر اهمیت بخشهای بهداشت و مددکاری و . . . ، آموزش و معدن)

رتبه	استان	امتیاز کسب شده
۱	کهگیلویه و بویر احمد	۲٫۸۵
۲	سمنان	۲٫۱۲
۳	کرمان	۱٫۶۲
۴	ایلام	۱٫۰۷
۵	خوزستان	۰٫۹۴
۶	بوشهر	۰٫۶۱
۷	تهران	۰٫۴۵
۸	یزد	۰٫۴۴
۹	فارس	۰٫۲۰
۱۰	کرمانشاه	۰٫۱۵
۱۱	لرستان	۰٫۰۷
۱۲	مازندران	۰٫۰۰
۱۳	چهارمحال و بختیاری	-۰٫۱۲
۱۴	سیستان	-۰٫۱۷
۱۵	گیلان	-۰٫۲۷
۱۶	گلستان	-۰٫۳۹
۱۷	قم	-۰٫۴۴
۱۸	اصفهان	-۰٫۵۱
۱۹	خراسان	-۰٫۶۱
۲۰	کردستان	-۰٫۶۱
۲۱	هرمزگان	-۰٫۶۵
۲۲	همدان	-۰٫۶۷
۲۳	اردبیل	-۰٫۸۰
۲۴	زنجان	-۰٫۸۱
۲۵	مرکزی	-۰٫۹۰
۲۶	آذربایجان غربی	-۰٫۹۹
۲۷	قزوین	-۱٫۰۳
۲۸	آذربایجان شرقی	-۱٫۵۳

جدول ۱۷: رتبه‌بندی استانها بر حسب عامل سوم برای درصد نیروی شاغل در فعالیتهای اقتصادی ــ ۱۳۷۹ (تاکید بر اهمیت بخشهای کسب و کار، واسطه‌گریهای مالی و صنعت)

رتبه	استان	امتیاز کسب شده
۱	تهران	۳٫۳۲
۲	یزد	۱٫۵۵
۳	اصفهان	۱٫۳۹
۴	قم	۱٫۳۲
۵	بوشهر	۰٫۸۹
۶	سمنان	۰٫۵۴
۷	مرکزی	۰٫۵۰
۸	آذربایجان شرقی	۰٫۴۳
۹	خوزستان	۰٫۳۵
۱۰	خراسان	۰٫۰۷
۱۱	هرمزگان	۰٫۰۳
۱۲	چهارمحال وبختیاری	۰٫۰۱
۱۳	قزوین	-۰٫۰۲
۱۴	فارس	-۰٫۲۱
۱۵	کرمان	-۰٫۳۹
۱۶	مازندران	-۰٫۴۸
۱۷	لرستان	-۰٫۵۱
۱۸	زنجان	-۰٫۵۲
۱۹	گیلان	-۰٫۵۲
۲۰	همدان	-۰٫۵۵
۲۱	کرمانشاه	-۰٫۶۵
۲۲	کردستان	-۰٫۷۴
۲۳	سیستان و بلوچستان	-۰٫۸۱
۲۴	گلستان	-۰٫۸۱
۲۵	آذربایجان غربی	-۰٫۸۵
۲۶	اردبیل	-۰٫۹۹
۲۷	کهگیلویه و بویر احمد	-۱٫۰۶
۲۸	ایلام	-۱٫۲۸

گامهای مطاعاتی مطالعات می‌باشد. در انتها بر اساس نتایج این تحقیق لزوم انجام مطالعات کلان اقتصادی در زمینه‌های: امکان برقراری توزیع عادلانه ثروت درکشور، استفاده از سازکارهای نوین در بکارگیری بهینه پتانسیل‌های اقتصادی موجود در کلیه استانها، اجرایی نمودن الگوهای افزایش بهره‌وری و کاهش هزینه‌ها و ... ضروری به نظر می‌رسد.

مراجع

[۱] استنباط آماری چند متغیره، نارایان سی. جری، ترجمه ابوالقاسم بزرگ نیا ـ موسسه چاپ و انتشارات آستان قدس رضوی. ۱۳۶۶.

[۲] آشنایی با روشهای چند متغیره، بی .اف . جی. مانلی، مترجم دکتر محمد مقدم، مهندس سید ابوالقاسم محمدی شوطی، مهندس مصطفی آقایی برزه، انتشارات پیشتاز علم، ۱۳۷۳.

[۳] حسابهای منطقه‌ای، اصول نظری و کاربرد (ویرایش دوم)، دفتر حسابهای اقتصادی، مرکز آمار ایران، ۱۳۷۹.

[۴] تحلیل آمار چند متغیره کاربردی، ریچارد. آ. جانسون، دین. دبلیو. دیچرن، ترجمه دکتر حسینعلی نیرومند، مشهد، دانشگاه فردوسی (مشهد)، ۱۳۷۸.

[۵] حسابهای ملی ایران. حسابهای منطقه‌ای، حساب تولید استان‌های کشور ۱۳۷۹، دفتر حسابهای اقتصادی، مرکز آمار ایران، ۱۳۸۱.

[۶] آمارگیری از ویژگیهای اشتغال و بیکاری خانوار ۱۳۷۹، مرکز آمار ایران، ۱۳۷۹.

[۷] سالنامه آماری کشور ـ ۱۳۷۹، مرکز آمار ایران، اسوه ۱۳۸۰.

برآورد فاصله‌ای برای خانواده توزیع‌های نمایی طبیعی

محسن عارفی، غلامرضا محتشمی برزادران

گروه آمار دانشگاه بیرجند

چکیده: در این مقاله یک برآورد فاصله‌ای برای میانگین، در توزیع‌های نمایی طبیعی که دارای تابع واریانس درجه دوم بر حسب میانگین می‌باشد، مورد توجه قرار گرفته است. این خانواده توزیع‌های نمایی طبیعی شامل توزیع‌های دوجمله‌ای، دوجمله‌ای منفی، پواسن، نرمال، گاما و سکانت هذلولوی تعمیم یافته می‌باشد. در ضمن توزیع گوسین معکوس که جزء خانواده توزیع‌های نمایی طبیعی با تابع واریانس درجه سوم بر حسب میانگین می‌باشد، نیز مورد بررسی قرار گرفته است. برای این توریعها، فاصله استاندارد والد و چهار فاصله اطمینان دیگر و مشخصه‌های مرتبط با آنها مورد ارزیابی قرار گرفته است. نتایج و محاسبات نشان می‌دهد که در میان این فواصل، فاصله نسبت درستنمایی کارایی بهتری را نشان می‌دهد.

۱ مقدمه

هدف از انجام این مقاله، بدست آوردن یک برآورد فاصله‌ای برای میانگین، در خانواده توزیع‌های نمایی طبیعی با تابع واریانس مورد نظر می‌باشد. بدین منظور در بخش دوم، خانواده توزیع‌های نمایی طبیعی و تابع واریانس درجه دوم و تابع واریانس درجه سوم را معرفی کرده‌ایم و ضرایب این تابع واریانسها را نیز مورد محاسبه قرار داده‌ایم.

در این مقاله ما بیشتر روی خانواده توزیع‌های نمایی طبیعی گسسته متمرکز می‌شویم و برای آنها انواع فواصل اطمینان را بدست می‌آوریم. همچنین در بین خانواده توزیع‌های نمایی طبیعی پیوسته، توزیع گوسین معکوس را که دارای تابع واریانس درجه سوم نسبت به میانگین می‌باشد، را بررسی می‌نمائیم.

در بخش سوم، انواع فواصل اطمینان را همراه با طول مورد انتظار وخاصیت اریبی را برای حالت گسسته با واریانس درجه دوم مورد ارزیابی قرار می‌دهیم. بخش چهارم شامل توزیع گوسین معکوس با تابع واریانس درجه سوم و فواصل مربوط به این توزیع، همراه با طول مورد انتظار می‌باشد.

۲ خانواده توزیع‌های نمایی طبیعی[1]

در این مقاله ما به دنبال بررسی یک برآورد فاصله‌ای برای میانگین، در خانواده توزیع‌های نمایی (NEF) با واریانس درجه دوم نسبت به میانگین و برای یک توزیع خاص با واریانس درجه سوم نسبت به میانگین می‌باشیم. خانواده توزیع‌های نمایی طبیعی با واریانس درجه دوم (NEF-QVF) شامل ۶ توزیع مهم «دوجمله‌ای، دوجمله‌ای منفی، پواسون، نرمال، گاما و توزیع سکانت هذلولوی تعمیم یافته (NEF-GHS)» می‌باشد (برای بررسی بیشتر درمورد توزیع سکانت هذلولوی تعمیم یافته به نوشته‌های موریس (۱۹۸۲) و برون (۱۹۸۶) مراجعه نمایید). همچنین خانواده توزیع‌های نمایی طبیعی با واریانس درجه سوم (NEF-CVF) علاوه بر ۶ توزیع فوق شامل ۶ توزیع دیگر می‌باشد، که ما در اینجا تنها توزیع گوسین معکوس را مورد بررسی قرار داده‌ایم (برای بررسی این توزیعها به نوشته‌های لتاک و مورا (۱۹۹۰) مراجعه نمائید).

یک خانواده توزیع‌های نمایی طبیعی را به صورت

$$f(x/\xi) = \exp[\xi x - \psi(\xi)]h(x); \qquad (۱)$$

نمایش می‌دهند، که به ξ پارامتر طبیعی می‌گویند. برای این خانواده توزیعها

$$\mu = \psi'(\xi) \quad , \quad \sigma^۲ = \psi''(\xi).$$

به ترتیب میانگین و واریانس می‌باشند. این واریانس تنها به میانگین μ بستگی دارد، که به صورت زیر نشان داده می‌شود:

$$\sigma^۲ \equiv var(\mu) = a_۰ + a_۱\mu + a_۲\mu^۲ + a_۳\mu^۳; \qquad (۲)$$

که برای تابع واریانس درجه دوم $a_۳ = ۰$ و $\Delta = a_۱^۲ - ۴a_۰a_۲$ می‌باشد. در نتیجه برای توزیع‌های مختلف این خانواده داریم:

- **توزیع برنولی،** $Bin(۱, p)$:
برای این توزیع نسبتهای $\xi = \log\frac{p}{q}$، $\psi(\xi) = \log\frac{۱}{q} = \log(۱ + e^{\xi})$ و $h(x) = ۱$
با میانگین $\mu = p$ و تابع واریانس درجه دوم $v(\mu) = pq = \mu - \mu^۲$ بدست می‌آید که ضرایب تابع واریانس به صورت $a_۰ = ۰$، $a_۱ = ۱$ و $a_۲ = -۱$ می‌باشد.
- **توزیع دوجمله‌ای منفی،** $NBin(۱, p)$ **(تعداد موفقیتها قبل از اولین شکست:** x**):**
در این توزیع نسبتهای زیر را داریم:

$$\xi = \log(p) \quad , \quad \psi(\xi) = -\log(۱ - p) = -iog(۱ - e^{\xi}) \quad , \quad h(x) = ۱;$$

1) Natural Exponential Family (NEF)

همچنین میانگین به صورت $\mu = \frac{p}{q}$ و تابع واریانس بر حسب این میانگین به صورت $v(\mu) = \frac{p}{q^۲} = \mu + \mu^۲$ با ضرایب $a_۰ = ۰$، $a_۱ = ۱$ و $a_۲ = ۱$ می‌باشد.

- **توزیع پواسن،** $Poi(\lambda)$:
نسبتهای فوق در این توزیع به صورت زیر می‌باشد:

$$\xi = \log \lambda \quad , \quad \psi(\xi) = \lambda = e^{\xi} \quad , \quad h(x) = \frac{۱}{x!};$$

که تابع واریانس به صورت $v(\mu) = \lambda = \mu$ با ضرایب $a_۰ = ۰$، $a_۱ = ۱$ و $a_۲ = ۰$ می‌باشد.

- **توزیع نرمال** $N(\mu, \sigma^۲)$ **($\sigma^۲$ معلوم):**
در این توزیع نسبتهای مورد نظر به صورت زیر می‌باشد:

$$\xi = \frac{\mu}{\sigma^۲} \quad , \quad \psi(\xi) = \frac{\mu^۲}{۲\sigma^۲} = \frac{\sigma^۲}{۲}\xi \quad , \quad h(x) = \frac{۱}{\sqrt{۲\pi\sigma^۲}} \exp\{-\frac{x^۲}{۲\sigma^۲}\};$$

در این حالت تابع واریانس به صورت $v(\mu) = \sigma^۲$ با ضرایب $a_۰ = \sigma^۲$ و $a_۱ = a_۲ = ۰$ می‌باشد.

- **توزیع گاما،** $Gamma(r, \lambda)$ **(r معلوم):**
روابط مورد نظر برای این توزیع به صورت

$$\xi = -\frac{۱}{\lambda} \quad , \quad \psi(\xi) = r \log \lambda = -r \log(-\xi) \quad , \quad h(x) = \frac{x^{r-۱}}{\Gamma(r)};$$

با میانگین و تابع واریانس و با ضرایب $a_۰ = a_۱ = ۰$ و $a_۲ = \frac{۱}{r}$ می‌باشد.

- **توزیع سکانت هذلولوی تعمیم یافته،** $GHS(r, \lambda)$ **(r معلوم):**
برای بررسی تابع چگالی این توزیع می‌توانید به نوشته‌های موریس (۱۹۸۲) و برون (۱۹۸۶) مراجعه نمائید. با توجه به تابع چگالی آن، نسبتهای زیر بدست می‌آید:

$$\xi = \tan^{-۱}(\lambda), \quad \psi(\xi) = \tfrac{۲}{۲} \log(۱ + \lambda^۲) = -r \log(\cos \xi), \quad h(x) = f_{r,۰}(x);$$

برای این توزیع میانگین و تابع واریانس به صورت زیر بدست می‌آید:

$$\mu = \psi'(\xi) = r \tan(\xi) = r\lambda \quad , \quad \nu(\mu) = \frac{r}{\cos^۲(\xi)} = r + \frac{\mu^۲}{r};$$

که ضرایب این تابع واریانس به صورت $a_۰ = r$ ، $a_۱ = ۰$ و $a_۲ = \frac{۱}{r}$ می‌باشد.

- **توزیع گوسین معکوس،** $IG(m,\lambda)$ **(**λ **معلوم):**
در این توزیع روابط زیر را می‌توان از فرم نمایی آن نتیجه گرفت:

$$\xi = \frac{-\lambda}{2m^2}\ ,\ \psi(\xi) = \frac{-\lambda}{m} = -(-2\lambda\xi)^{1/2}\ ,\ h(x) = \sqrt{\frac{\lambda}{2\pi x^3}}\exp\{\frac{-\lambda}{2x}\};$$

که میانگین و تابع واریانس درجه سوم آن به صورت زیر محاسبه می‌گردد:

$$\mu = \psi'(\xi) = m\ \ ,\ \ \nu(\mu) = \psi''(\xi) = \frac{m^3}{\lambda} = \frac{\mu^3}{\lambda}$$

در نتیجه ضرایب تابع واریانس برابر $a_0 = a_1 = a_2 = 0$ و $a_3 = \frac{1}{\lambda}$ می‌باشد.

اکنون ما روی توزیع‌های نمایی گسسته با واریانس درجه دوم متمرکز می‌شویم و از میان توزیع‌های پیوسته، توزیع گوسین معکوس را که دارای تابع واریانس درجه سوم می‌باشد، نیز مورد بررسی قرار می‌گیرد.

۳ خانواده توزیع‌های نمایی طبیعی گسسته با واریانس درجه دوم

این خانواده توزیعها شامل سه توزیع دوجمله‌ای، دوجمله‌ای منفی، پواسن می‌باشد. برای این توزیعها ما انواع فواصل اطمینان و بعضی روابط خاص را بررسی می‌کنیم.

۱.۳ بسط اریبی

فاصله استاندارد والد به صورت $\hat{\mu} \pm z_{\frac{\alpha}{2}} n^{-\frac{1}{2}}\sqrt{a_0 + a_1\hat{\mu} + a_2\hat{\mu}^2}$ می‌باشد، که بر اساس کمیت محوری زیر پایگذاری می‌شود:

$$W_n = \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sqrt{a_0 + a_1\hat{\mu} + a_2\hat{\mu}^2}} \xrightarrow{L} N(0,1)$$

که در آن $\hat{\mu} = \bar{X}$، یک برآورد MLE از μ می‌باشد. در واقع به طور کلی W_n دارای توزیع مجانبی نرمال استاندارد می‌باشد. از آنجایی که $Z = \frac{\sqrt{n}(\bar{\mu}-\mu)}{\sqrt{a_0+a_1\mu+a_2\mu^2}}$ دارای توزیع نرمال استاندارد می‌باشد، می‌توانیم میزان اریبی را محاسبه نمائیم. از یک عبارت جبری ساده رابطه بین Z و W_n در توزیعهایی با تابع واریانس درجه دوم به صورت زیر در می‌آید:

$$W_n = Z(1 + (a_1 + 2a_2\mu)\sigma^{-1}n^{-\frac{1}{2}}Z + a_2 n^{-1}Z^2)^{-\frac{1}{2}}$$

این فرآیند را برای سه توزیع گسسته، می‌توان در نوشته‌های برون، کی وداس گیوپتا (۱۹۹۹b)به صورت زیر ملاحظه نمود، که این روابط براساس سری مکلورن برای $f(Z) = W_n$ بدست می‌آید.

- برنولی، $Bin(۱, p)$:

$$E(W_n) = \frac{p - \frac{۱}{۲}}{\sqrt{npq}}(۱ + \frac{۷}{۲n} + \frac{۹(p - \frac{۱}{۲})^۲}{۲npq}) + O(n^{-۲}) \qquad (۳)$$

- دوجمله‌ای منفی، $NBin(۱, p)$:

$$E(W_n) = -\frac{۱ + p}{۲\sqrt{np}}(۱ + \frac{۱}{n} + \frac{۹q^۲}{۸np}) + O(n^{-۲}) \qquad (۴)$$

- پواسن، $Poi(\lambda)$:

$$E(W_n) = -\frac{۱}{۲\sqrt{n\lambda}}(۱ + \frac{۹}{۸n\lambda}) + O(n^{-۲}) \qquad (۵)$$

با توجه به رابطه (۳) معلوم می‌گردد که در حالت دوجمله‌ای، W_n دارای اریبی منفی برای $p < \frac{۱}{۲}$ و دارای اریبی مثبت برای $p > \frac{۱}{۲}$ می‌باشد و در نتیجه فاصله والد حول $\frac{۱}{۲}$ نوسان می‌کنند. از معادله (۴) و (۵) این طور به نظر می‌رسد که W_n همواره دارای اریبی منفی می‌باشد.

۲.۳ انواع فاصله‌های اطمینان

فرض کنید $\hat{\mu} = \bar{X}$ برآورد MLE از μ باشد. با توجه به قضیه حد مرکزی[۲] و قضیه اسلوتسکی[۳] روابط زیر حاصل می‌شود:

$$W_n = \frac{\sqrt{n}(\hat{\mu} - \mu)}{\hat{\sigma}} = \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sqrt{a_۰ + a_۱\hat{\mu} + a_۲\hat{\mu}^۲}} \xrightarrow{L} N(۰, ۱); \qquad (۶)$$

$$Z = \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma}\frac{\sqrt{n}(\hat{\mu} - \mu)}{\sqrt{a_۰ + a_۱\mu + a_۲\mu^۲}} \xrightarrow{L} N(۰, ۱); \qquad (۷)$$

بر اساس دو تقریب (۶) و (۷) ما برخی از فاصله‌های اطمینان را به صورت زیر پایگذاری می‌کنیم:
فاصله # ۱: فاصله استاندارد والد بر اساس رابطه (۶) به صورت زیر بدست می‌آید:

$$CI_S = \hat{\mu} \pm k\hat{\sigma}n^{-\frac{۱}{۲}} = \hat{\mu} \pm k(a_۰ + a_۱\hat{\mu} + a_۲\hat{\mu}^۲)^{\frac{۱}{۲}}n^{-\frac{۱}{۲}}; \qquad (۸)$$

2) Central Limit Theorem 3) Slutsky's Theorem

فاصله # ۲: فاصله اسکور بر اساس تقریب (۷) پایگذاری می‌شود. این فاصله به وسیله وارونی از آزمون اسکور با دمهای برابر رائو تحت فرض $H_0 : \mu = \mu_0$ بدست می‌آید. بنابراین فرض H_0 پذیرفته می‌شود، اگر μ در فاصله مورد نظر باشد. با حل یک معادله درجه دوم براساس تقریب (۷) فاصله اسکور بدست می‌آید:

$$CI_R = \tilde{\mu} \pm \frac{kn^{\frac{1}{2}}}{n - k^2 a_2} = (a_0 + a_1\hat{\mu} + a_2\hat{\mu}^2 + \frac{k^2}{4n}\Delta)^{\frac{1}{2}}; \qquad (۹)$$

که $\tilde{\mu} = \frac{n\hat{\mu}+k^2 a_1/2}{n-k^2 a_2}$ می‌باشد.

فاصله # ۳: فاصله اگرستی-کویل (AC) شکلی شبیه فاصله استاندارد والد دارد، با این تفاوت که به جای $\hat{\mu}$ مقدار $\tilde{\mu}$ و به جای n مقدار تصحیح شده $\tilde{n} = n - k^2 a_2$ را قرار می‌دهیم:

$$CI_{AC} = \tilde{\mu} \pm k(\nu(\tilde{\mu}))^{\frac{1}{2}}\tilde{n}^{-\frac{1}{2}} = \tilde{\mu} \pm k(a_0 + a_1\tilde{\mu} + a_2\tilde{\mu}^2)^{\frac{1}{2}}\tilde{n}^{-\frac{1}{2}}; \qquad (۱۰)$$

فاصله # ۴: فاصله نسبت درستنمایی به وسیله وارونی از آزمون نسبت درستنمایی با پذیرش فرض صفر $H_0 : \mu = \mu_0$، اگر $-2\log\Lambda_n \leq \chi^2_{\alpha,1}$ باشد، بدست می‌آید. که در آن

$$\Lambda_n = \frac{L(\mu_0)}{\sup L_\mu(\mu)};$$

نسبت درستنمایی و L تابع درستنمایی بر اساس n مشاهده مستقل و همتوزیع می‌باشد. برای مطالعه بیشتر در این زمینه می‌توانید به نوشته‌های رائو (۱۹۷۳) و سرفلینگ (۱۹۸۰) مراجعه نمائید.

فاصله # ۵: فاصله جفریز:

فرض کنید $b(.) = (\psi')^{-1}(.)$ باشد، به وضوح آشکار است که $b(\mu) = \xi$ می‌باشد. فرض کنیم:

$$I(\mu) = n(a_0 + a_1\mu + a_2\mu^2)^{-1}$$

تابع پیشین جفریز را متناسب با $I^{1/2}(\mu)$ می‌گیریم، بنابراین تابع پسین به صورت زیر بدست می‌آید:

$$f(\mu|x) \propto \exp\{xb(\mu) - n\psi(b(\mu)) - \frac{1}{2}\log(a_0 + a_1\mu + a_2\mu^2)\}$$

فاصله جفریز با دمهای برابر برای μ از تابع پسین جفریز، به صورت زیر بدست می‌آید:

$$CI_J = \left[J_{\frac{\alpha}{2}}, J_{1-\frac{\alpha}{2}}\right]$$

که $J_{\frac{\alpha}{2}}$ و $J_{1-\frac{\alpha}{2}}$ چندکهای $\frac{\alpha}{2}$ و $1-\frac{\alpha}{2}$ تحت توزیع پسین جفریز و بر مبنای n مشاهده می‌باشد (برای مطالعه بیشتر به نوشته‌های کس و واسرمن (۱۹۹۶) مراجعه نمائید). اکنون فاصله جفریز را برای سه توزیع گسسته محاسبه می‌نمائیم:

دوجمله‌ای: در این حالت $\psi(\xi) = \log(1+e^{\xi})$ و $b(\mu) = \log \frac{\mu}{1-\mu}$ می‌باشد. بنابراین تابع پیشین جفریز دارای توزیع $\beta(\frac{1}{2}, \frac{1}{2})$ و تابع پسین دارای توزیع $\beta(X+\frac{1}{2}, n-X+\frac{1}{2})$ با $X = \sum_{i=1}^{n} X_i$ می‌باشد. بنابراین فاصله جفریز با دمهای برابر برای در سطح $100(1-\alpha)$ درصد به صورت زیر بدست می‌آید:

$$CI_J = [p_L, p_U] = [B_{\{\frac{\alpha}{2}, X+\frac{1}{2}, n-X+\frac{1}{2}\}}, B_{\{1-\frac{\alpha}{2}, X+\frac{1}{2}, n-X+\frac{1}{2}\}}]$$

دوجمله‌ای منفی: در این حالت $\psi(\xi) = -\log(1-e^{\xi})$ و $b(\mu) = \log \frac{\mu}{1+\mu}$ می‌باشد. تابع پیشین برای μ متناسب با $\mu^{-\frac{1}{2}}(1+\mu)^{-\frac{1}{2}}$ می‌باشد، که یک تابع بتا پریم نامیده می‌شود (برای اطلاعات بیشتر به جانسون (۱۹۹۵) مراجعه نمائید). برای حل این مشکل ابتدا یک فاصله اطمینان برای p بدست می‌آوریم. فاصله پیشین برای p متناسب با $p^{-\frac{1}{2}}(1-p)^{-1}$ و تابع پسین دارای توزیع $Beta(X+1/2, n)$ می‌باشد. پس فاصله اطمینان برای p در سطح $100(1-\alpha)$ درصد برابر است با:

$$CI_J^p = [p_L, p_U] = [B_{\{\alpha/2, X+1/2, n\}}, B_{\{1-\alpha/2, X+1/2, n\}}]$$

از آنجایی که $\mu = p/(1-p)$ می‌باشد، پس فاصله جفریز برای μ عبارتست از:

$$CI_J = [p_L/(1-p_L), p_U/(1-p_U)]$$

پواسن: در این حالت $\psi(\xi) = e^{\xi}$ و $b(\mu) = \log \mu$ می‌باشد. تابع پیشین متناسب با $\lambda^{-\frac{1}{2}}$ و تابع پسین دارای توزیع $Gamma(X+1/2, 1/n)$ می‌باشد. بنابراین فاصله اطمینان جفریز با دمهای برابر برای λ در سطح $100(1-\alpha)$ درصد برابر با

$$CI_J = [\lambda_L, \lambda_U] = [G_{\{\frac{\alpha}{2}, X+\frac{1}{2}, \frac{1}{n}\}}, G_{\{1-\frac{\alpha}{2}, X+\frac{1}{2}, \frac{1}{n}\}}]$$

خواهد بود.

۳.۳ طول فواصل اطمینان مورد انتظار

در میان فاصله اطمینانهای بدست آمده، کمترین طول برای یک فاصله اطمینان، همواره موضوع مورد اهمیتی بوده است. بنابراین در میان کلیه فواصل اطمینان، امید طولهای آنها را محاسبه و با یکدیگر مورد مقایسه قرار می‌دهیم.

قضیه: فرض کنید CI یک نمایش کلی برای هر یک از پنج فاصله اطمینان CI_s، CI_R، CI_{AC}، CI_{LR} و CI_J باشد، امید طولها برابر است با:

$$L(n,\mu) = E(L_{CI}) = ۲k(\mu+a_۲\mu^۲)^{۱/۲}n^{-۱/۲}(۱-\frac{\delta(k,\mu)}{۷۲n(\mu+a_۲\mu^۲)})+O(n^{-۲})$$

که $\delta(k,\mu)$ به صورت زیر محاسبه می‌شود:

$$\begin{aligned}
\delta(k,\mu) &= ۹\Delta \quad for \quad CI_s;\\
&= ۹(۱-k^۲)\Delta - ۷۲k^۲a_۲(\mu+a_۲\mu^۲) \quad for \quad CI_R;\\
&= ۹(۱-۲k^۲)\Delta - ۱۰۸k^۲a_۲(\mu+a_۲\mu^۲) \quad for \quad CI_{AC};\\
&= (۹-۲k^۲)\Delta - ۲۶k^۲a_۲(\mu+a_۲\mu^۲) \quad for \quad CI_{LR};\\
&= (۵-۲k^۲)\Delta - ۲(۱۳k^۲+۱۷)a_۲(\mu+a_۲\mu^۲) \quad for \quad CI_J;
\end{aligned}$$

نتیجه (۱): حالت پواسن را در نظر بگیرید. طول مورد انتظار برای انواع فاصله اطمینانها عبارتست از:

$$\begin{aligned}
E(L_s) &= ۲k\lambda^{۱/۲}n^{-۱/۲}(۱-\frac{۹}{۷۲n\lambda})+O(n^{-۲});\\
E(L_{LR}) &= ۲k\lambda^{۱/۲}n^{-۱/۲}(۱+\frac{۹(k^۲-۱)-۷k^۲}{۷۲n\lambda})+O(n^{-۲});\\
E(L_J) &= ۲k\lambda^{۱/۲}n^{-۱/۲}(۱+\frac{۹(k^۲-۱)+۴-۷k^۲}{۷۲n\lambda})+O(n^{-۲});\\
E(L_R) &= ۲k\lambda^{۱/۲}n^{-۱/۲}(۱+\frac{۹(k^۲-۱)}{۷۲n\lambda})+O(n^{-۲});\\
E(L_{AC}) &= ۲k\lambda^{۱/۲}n^{-۱/۲}(۱+\frac{۹(۲k^۲-۱)}{۷۲n\lambda})+O(n^{-۲});
\end{aligned}$$

با توجه به فرآیندهای بالا معلوم می‌شود که برای هر

$$k > \frac{۲}{\sqrt{۷}} = ۰/۷۶$$

فواصل CI_{AC}و$CI_R, CI_J, CI_{LR}, CI_S$ به ترتیب دارای کوچکترین تا بزرگترین طول می‌باشند. بنابراین فاصله نسبت درستنمایی دارای کوچکترین طول می‌باشد.

نتیجه (۲): برآورد فاصله‌ای برای $\mu = p/q$ را در حالت دوجمله‌ای در نظر بگیرید. طول مورد انتظار برای فواصل مختلف به صورت زیر می‌باشد:

$$\begin{aligned}
E(L_s) &= 2kp^{1/2}q^{-1}n^{-1/2}(1-\frac{9q^2}{72np})+O(n^{-2});\\
E(L_{LR}) &= 2kp^{1/2}q^{-1}n^{-1/2}(1-\frac{9q^2-2k^2(1+11p+p^2)}{72np})+O(n^{-2});\\
E(L_J) &= 2kp^{1/2}q^{-1}n^{-1/2}\\
&\quad (1-\frac{9q^2-2k^2(1+11p+p^2)-2(2+13p+2p^2)}{72np})+O(n^{-2});\\
E(L_R) &= 2kp^{1/2}q^{-1}n^{-1/2}(1-\frac{9q^2-9k^2(1+6p+p^2)}{72np})+O(n^{-2});\\
E(L_{AC}) &= 2kp^{1/2}q^{-1}n^{-1/2}(1-\frac{9q^2-18k^2(1+4p+p^2)}{72np})+O(n^{-2});
\end{aligned}$$

با توجه به بسطهای بالا نتیجه می‌شود که برای هر $p > 0$ و برای هر $k > \sqrt{17/23} = 0/86$ فواصل $CI_{AC}, CI_R, CI_J, CI_{LR}, CI_S$ به ترتیب دارای کوچکترین تا بزرگترین طول می‌باشند. بنابراین فاصله نسبت درستنمایی دارای کوچکترین طول مورد انتظار می‌باشد.

بر خلاف دو حالت پواسن و دوجمله‌ای منفی، یک ترتیب یکنواختی برای طول مورد انتظار در حالت دوجمله‌ای وجود ندارد. ولی اگر نسبت به p از طولهای مورد انتظار انتگرال بگیریم، یک ترتیب واضحی به صورت زیر وجود دارد.

نتیجه (۳): حالت دوجمله‌ای را در نظر بگیرید. انتگرالی از طولهای مورد انتظار برای فواصل مختلف به صورت زیر می‌باشد:

$$\begin{aligned}
\int_0^1 E(L_J)dp &= \frac{k\pi}{4}n^{-1/2}-(\frac{37}{36}+\frac{5k^2}{36})\frac{k\pi}{4}n^{-3/2}+O(n^{-2});\\
\int_0^1 E(L_{LR})dp &= \frac{k\pi}{4}n^{-1/2}-(1+\frac{5k^2}{36})\frac{k\pi}{4}n^{-3/2}+O(n^{-2});\\
\int_0^1 E(L_R)dp &= \frac{k\pi}{4}n^{-1/2}-\frac{k\pi}{4}n^{-3/2}+O(n^{-2});\\
\int_0^1 E(L_s)dp &= \frac{k\pi}{4}n^{-1/2}-\frac{k\pi}{4}n^{-3/2}+O(n^{-2});\\
\int_0^1 E(L_{AC})dp &= \frac{k\pi}{4}n^{-1/2}-(\frac{k^2}{2}-1)\frac{k\pi}{4}n^{-3/2}+O(n^{-2});
\end{aligned}$$

بنابراین فواصل CI_J، CI_{LR}، CI_R، CI_S و CI_{AC} به ترتیب از کوچکترین به بزرگترین طول می‌باشد. به طور تقریبی فواصل CI_Jو CI_{LR} با هم برابر می‌باشند، پس فاصله نسبت درستنمایی طول بهتری دارد.

۴ فاصله اطمینان برای تابع توزیعی با واریانس درجه سوم

خانواده توزیع‌های نمایی طبیعی با واریانس درجه سوم (NEF-CVF) شامل ۶ توزیع علاوه بر توزیعهای دیگر می‌باشد که ما در میان این ۶ توزیع تنها بعضی خواص توزیع گوسین معکوس را مورد بررسی قرار می‌دهیم (برای بررسی این توزیعها به نوشته‌های لتاک و مورا (۱۹۹۰) مراجعه نمائید).

برای توزیع گوسین معکوس $IG(m,\lambda)$ که λ معلوم می‌باشد، میانگین برابر $\mu = m$ و واریانس تحت میانگین به صورت زیر محاسبه می‌شود:

$$var(\mu) = \frac{m^3}{\lambda} = \frac{\mu^3}{\lambda};$$

در نتیجه ضرایب تابع واریانس درجه سوم برابر $a_0 = a_1 = a_2 = 0$ و $a_3 = \frac{1}{\lambda}$ می‌باشد.

با توجه به اینکه $W_n = \frac{\sqrt{n}(\hat{\mu}-\mu)}{\hat{\sigma}^2}$ می‌باشد، براساس یک عبارت جبری ساده و با توجه به تابع واریانس درجه سوم W_n به صورت زیر تبدیل می‌شود:

$$W_n = Z(1 + (a_1 + 2a_2\mu + 3a_3\mu^2)\sigma^{-1}n^{-1/2}Z + (a_2 + 3a_3\mu)n^{-1}Z^2 + a_3n^{-3/2}Z^3\sigma)^{-1/2}$$

با توجه به ضرایب تابع واریانس درجه سوم در توزیع گوسین معکوس داریم:

$$W_n = Z(1 + 3a_3\mu^2\sigma^{-1}n^{-1/2}Z + 3a_3n^{-1}Z^2\mu + a_3n^{-3/2}Z^3\sigma)^{-1/2}$$

از سری مکلورن برای $f(Z) = W_n$ می‌توانیم میزان اریبی را محاسبه نمود:

$$E(W_n) = -\frac{3}{2}n^{-1/2}\mu^{1/2}\lambda^{-1/2}(1 + \frac{35\mu}{8n\lambda}) + O(n^{-2});$$

پس اریبی در توزیع گوسین معکوس همواره منفی می‌باشد.

۵ انواع فاصله اطمینانهای برای توزیع گوسین معکوس

برای این توزیع ما سه فاصله اطمینان را به صورت زیر بدست می‌آوریم:
□**فاصله استاندارد والد:** با توجه به تقریب اسلوتسکی این فاصله به صورت زیر بدست می‌آید:

$$CI_S = \hat{\mu} \pm kn^{\frac{-۱}{۲}}\hat{\sigma} = \hat{\mu} \pm kn^{\frac{-۱}{۲}}\hat{\mu}^{\frac{۳}{۲}}\lambda^{\frac{-۱}{۲}}$$

که $\hat{\mu} = \bar{X}$ برآورد MLE از μ می‌باشد.
□ **فاصله نسبت درستنمایی:** این فاصله با توجه به نسبت درستنمایی Λ_n و با پذیرش فرض $H_۰ : \mu = \mu_۰$ و بر اساس رابطه $-۲\log\Lambda_n \leq k^۲$ به صورت زیر بدست می‌آید:

$$\begin{aligned}\Lambda_n &= \frac{L(\mu_۰)}{\sup_\mu L(\mu)} = \frac{\exp\{\frac{-\lambda}{۲\mu^۲}\sum_{i=۱}^{n} X_i + \frac{n\lambda}{\mu}\}\prod_{i=۱}^{n} h(X_i)}{\exp\{\frac{-\lambda}{۲\bar{\mu}^۲}\sum_{i=۱}^{n} X_i + \frac{n\lambda}{\bar{\mu}}\}\prod_{i=۱}^{n} h(X_i)} \\ &= \exp\{\frac{-n\lambda}{۲\mu^۲}\bar{X} + \frac{n\lambda}{\mu} - \frac{n\lambda}{\bar{X}}\};\end{aligned}$$

در نتیجه می‌توان نوشت:

$$-۲\log\Lambda_n = -۲\{\frac{-n\lambda}{۲\mu^۲}\bar{X} + \frac{n\lambda}{\mu} - \frac{n\lambda}{\bar{X}}\} \leq k^۲$$

با حل یک معادله درجه دوم فاصله مورد نظر برای $\frac{۴n\lambda}{۲۷\mu} \leq k^۲ \leq \frac{n\lambda}{\mu}$ به صورت زیر بدست می‌آید:

$$CI_{LR} = \frac{n\hat{\mu}}{n - k^۲a_۳\hat{\mu}} \pm \frac{k\hat{\mu}\sqrt{na_۳\hat{\mu}}}{n - k^۲a_۳\hat{\mu}}$$

□ **فاصله اسکور:** این فاصله اطمینان بر اساس قضیه حد مرکزی بر اساس آزمون اسکور با دمهای برابر و با پذیرش فرض صفر $H_۰ : \mu = \mu_۰$ محاسبه می‌شود. با توجه به قضیه حد مرکزی می‌توان نوشت:

$$Z = \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} = \frac{\sqrt{n\lambda}(\hat{\mu} - \mu)}{\sqrt{\mu^۳}} \leq k$$

برای بدست آوردن این فاصله باید یک معادله درجه سوم را محاسبه نمود. ابتدا تغییر متغیر $y = \sqrt{\mu}$ را در رابطه بالا اعمال می‌کنیم:

$$\frac{\sqrt{n\lambda}(\hat{\mu} - y^۲)}{y^۳} \leq k \Rightarrow \sqrt{n\lambda}(\hat{\mu} - y^۲) \leq ky^۳ \Rightarrow ky^۳ + \sqrt{n\lambda}y^۲ - \sqrt{n\lambda}\hat{\mu} \geq ۰;$$

با تغییر متغییر $y = \frac{۱}{x}$ معادله درجه سوم زیر حاصل می‌شود:

$$x^۳ - \frac{۱}{\hat{\mu}} - \frac{k}{\sqrt{n\lambda}\hat{\mu}} \leq ۰$$

با توجه به مقدار $\frac{۴n\lambda}{۲۷\hat{\mu}} \leq k^۲ \leq \frac{n\lambda}{\hat{\mu}}$ داریم:

$$\Phi = \frac{-۴}{\hat{\mu}^۳} + \frac{۲۷k^۲}{n\lambda\hat{\mu}^۲} > ۰$$

از دستور کاردان برای معادلات درجه سوم رابطه زیر حاصل می‌گردد:

$$x < (\frac{k}{\sqrt{n\lambda}\hat{\mu}} + \sqrt{\frac{\Phi}{۱۰۸}})^{۱/۳} + (\frac{k}{\sqrt{n\lambda}\hat{\mu}} - \sqrt{\frac{\Phi}{۱۰۸}})^{۱/۳}$$

اکنون اگر تغییر متغییرهای فوق را اعمال نمائیم، یک فاصله اطمینان یکطرفه به صورت زیر بدست می‌آید:

$$\big[(\frac{k}{\sqrt{n\lambda}\hat{\mu}} + \sqrt{\frac{\Phi}{۱۰۸}})^{۱/۳} + (\frac{k}{\sqrt{n\lambda}\hat{\mu}} - \sqrt{\frac{\Phi}{۱۰۸}})^{۱/۳}\big]^{-۲} \leq \mu$$

۱.۵ طول فاصله اطمینان مورد انتظار برای توزیع گوسین معکوس

- **فاصله استاندارد والد:** طول این فاصله اطمینان عبارتست از:

$$L_s = ۲kn^{-۱/۲}\lambda^{-۱/۲}\hat{\mu}^{۳/۲} = ۲kn^{-۱/۲}\lambda^{-۱/۲}(Z\frac{\sigma}{\sqrt{n}} + \mu)^{۳/۲}$$

با استفاده از سری مکلورن برای رابطه بالا، طول مورد انتظار به صورت زیر محاسبه می‌شود:

$$E(L_s) = ۲kn^{-۱/۲}\lambda^{-۱/۲}\mu^{۳/۲}(۱ + \frac{۳\mu}{۸n\lambda}) + O(n^{-۲})$$

- **فاصله اطمینان نسبت درستنمایی:** طول این فاصله اطمینان عبارتست از:

$$L_{LR} = ۲kn^{-۱/۲}\lambda^{-۱/۲}\hat{\mu}^{۳/۲}(n - k^۲\lambda^{-۱}\hat{\mu})^{-۱}$$

در نتیجه بر اساس سری مکلورن، طول مورد انتظار برابر است با:

$$E(L_{LR}) = \frac{۲kn^{-۱/۲}a_۳^{۳/۲}\mu^{۳/۲}}{n - k^۲a_۳}(۱ + \frac{۳\mu}{۴} + \frac{۳k^۲a_۳\mu^۲}{n - k^۲a_۳\mu} + \frac{۲k^۴a_۳^۲\mu^۳}{(n - k^۲a_۳\mu)^۲}) + O(n^{-۲})$$

طول فاصله یکطرفه اسکور بدست آمده، بینهایت می‌باشد. با مقایسه طولهای مورد انتظار بالا به نظر می‌رسد که فاصله نسبت درستنمایی دارای کوچکترین طول می‌باشد.

۶ نتیجه‌گیری

نتایج و محاسبات بالا نشان می‌دهد که فاصله استاندارد والد خیلی ضعیف عمل می‌کند. بدست آوردن دو فاصله اطمینان نسبت درستنمایی و جفریز کار ساده‌ای نیست، ولی این دو فاصله اطمینان کارایی بهتری را نشان می‌دهند و همچنین دارای کوتاهترین طول نیز می‌باشند. فاصله‌های اطمینان اسکور و AC در مقایسه با دو فاصله قبلی برای توزیعهایی با تابع واریانس درجه دوم، راحتر بدست می‌آید، ولی برای توزیعهایی با تابع واریانس درجه سوم، نیاز به حل معادلات درجه سوم می‌باشد، که به سادگی قابل انجام نمی‌باشد.

مراجع

[1] Agresti,A. & Coull,B.A. (1998). Approximate is better than exact for interval estimation of binomial proportions. The American Statistician 52, 119-126.

[2] Brown,L.D. (1986). Fundamental of statistical exponential families with applications in statistical decision theory. Lecture Notes-Monograph Series, Institute of Mathematical Statistics, Hayward, California.

[3] Brown, L.D., Cai, T. & DasGupta, A. (1999b). Confidence interval for a binomial proportion and Edgeworth expansions. Technical Report, Department of Statistical, Purdue University. Submitted.

[4] Brown, L.D., Cai, T. & DasGupta, A. (2003). Interval estimation in exponential families. Technical Report,University of Pennsylvania and Purdue University. Submitted.

[5] Johnson, N.L., Kotz, s. & Balakrishnan, N. (1995).Continuous univareiate distributions. second edition, Vol. 2. Wiley, New york.

[6] Kass, R.E. & Wasserman, L. (1996). The selection of prior distribution by formal rules. J. Amer. Stat. Assoc., 91, 1343-70.

[7] Letac, G. & Mora, M. (1990). Natural real exponential families with cubic variance functions.

[8] Morris, C.N. (1982). Natural exponential families with quadratic variance functions. Ann. Statist. 10, 65-80.

[9] Morris, C.N. (1983). Natural exponential families with quadratic variance functions: statistical theory. Ann. Statist. 11, 515-529.

[10] Rao, C.R. (1973). Linear statistical inference and its applications. Wiley, New york.

[11] Serfling, R.J. (1980). Approximation theorems of mathematical statistics. Wiley, New York.

روش درستنمایی وزنی در برآوردیابی برای ناحیه کوچک

ملیحه عباس نژاد مشهدی

گروه آمار، دانشکده علوم ریاضی دانشگاه فردوسی مشهد

چکیده: برآوردیابی برای ناحیه کوچک[1] در سالهای اخیر مورد توجه زیادی قرارگرفته است، به دلیل این که تقاضا برای برآوردگرهای معتبر نواحی کوچک هم از جانب بخش‌های دولتی و هم از جانب بخش‌های خصوصی افزایش یافته است. از بررسیهای نمونه‌ای می‌توان برآوردهای مستقیم معتبری برای پارامترهای کل جامعه به دست آورد اما برآوردگرهای مستقیم برای هر ناحیه کوچک یعنی برآوردگرهایی که تنها بر مبنای داده‌های موجود در ناحیه به دست می‌آیند به دلیل کوچکی حجم نمونه در هر ناحیه به خطاهای استاندارد بزرگی منجر می‌شوند. بنابراین ضروری به نظر می‌رسد روشهایی برای این منظور جایگزین شوند که علاوه بر اطلاعات موجود در خود ناحیه، اطلاعات سایر نواحی (یا اطلاعات سرشماری) را نیز مورد استفاده قرار دهند. در این مقاله ابتدا به غیر مجاز بودن برآوردگر درستنمایی ماکزیمم میانگین نرمال در برآوردیابی همزمان[2] یعنی $\bar{X}$ اشاره‌ای کوتاه خواهیم داشت، روش درستنمایی وزنی[3] را در خانواده نمایی به طور کلی بیان کرده و سپس آن را در مسأله برآوردیابی برای ناحیه کوچک مورد بحث قرار می‌دهیم.

واژه‌های کلیدی: ناحیه کوچک، درستنمایی وزنی، برآوردگر جیمز استاین[4]

۱ مقدمه

ناحیه کوچک (یا ناحیه محلی) به یک ناحیه جغرافیایی کوچک مانند یک استان، شهر یا یک بخش سرشماری اطلاق می‌گردد. همچنین می‌توان یک زیرجامعه کوچک همانند یک گروه از مردم درون یک ناحیه جغرافیایی بزرگ را که از سن و جنس و نژاد معینی هستند به عنوان یک ناحیه کوچک در نظر گرفت.

مسأله برآوردیابی برای ناحیه کوچک مسأله‌ای است که در سالهای اخیر توجه زیادی به آن شده است. این امر به این علت است که تقاضا برای آماره‌های نواحی کوچک چه از جانب بخشهای دولتی و چه از جانب بخشهای خصوصی رشد زیادی داشته است. به عنوان مثال بخش دولتی برای برنامه‌ریزی کلان در سطح کشور نیاز به آماره‌هایی از قبیل نرخ بیکاری، متوسط درآمد خانوارها و . . . در استانها و شهرهای مختلف دارد. به عنوان مثال فرض کنید که بعضی

1) Small Area 2) Simulataneous Estimation 3) Weighted Likelihood 4) James-stein Estimator

از نواحی جغرافیایی یا زیرگروههایی از جامعه از قبیل بعضی استانها و یا قشر خاصی از افراد از نظر سطح درآمد پائینتر از یک میانگین قابل قبول و در حد انتظار باشند. واضح است که قبل از انجام هرکاری و هر نوع برنامه‌ریزی لازم است که چنین نواحی یا زیرگروههایی مشخص شوند.

داده‌های به دست آمده از بررسیهای نمونه‌ای را می‌توان برای استخراج برآوردهای مستقیم معتبر برای نواحی بزرگ یا قلمروهای بزرگ (نواحی با نمونه‌های بزرگ) استفاده نمود. اما برآوردهای مستقیم یعنی برآوردهایی که تنها بر مبنای داده‌هایی که از نمونه اخذ شده از آن ناحیه استخراج می‌شوند قرار دارند دارای خطای بسیار بزرگی خواهند بود که ناشی ازکوچک بودن بیش از اندازه حجم نمونه در آن نواحی است. دلیل این امر این است که در ابتدا دقت مورد نظر برای برآورد پارامترکل جامعه در نظرگرفته شده و حجم نمونه اصلی بر مبنای آن تعیین گردیده است. یعنی در هنگام برنامه‌ریزی و تعیین حجم نمونه فقط حجم نمونه کل مورد نیاز محاسبه شده است و طبعاً سهم هر ناحیه کوچک از کل حجم نمونه ناچیز است، در نتیجه برای نواحی کوچک این برآوردها دقت کافی را نخواهند داشت.

واضح است که از هر ناحیه نیز نمی‌توان نمونه‌ای بزرگ اختیارکرد. لذا ناچار باید به دنبال راهی بود که دقت برآوردها را تا آنجا که ممکن است بالا برد. برای این منظور می‌توان از داده‌های نواحی دیگر و اطلاعاتی که در آن نواحی براساس نمونه به دست آمده است در جهت بهبود بخشیدن به برآوردهای پارامترهای نواحی کوچک استفاده نمود. به این طریق حجم نمونه مؤثر افزایش یافته و دقت برآوردها افزایش می‌یابد. در مقاله حاضر روش درستنمایی وزنی را به این منظور مورد بحث قرار می‌دهیم.

۲ روش درستنمایی وزنی

روش درستنمایی بدون شک یکی از قدرتمندترین ابزارها در مسائل برآوردیابی و آزمون فرض می‌باشد؛ اما زمانی که در برآوردیابی همزمان با مسأله ترکیب اطلاعات روبرو هستیم روش درستنمایی کلاسیک کارآمد نخواهد بود. اهمیت این مسأله در مقاله استاین (۱۹۵۶) دیده شد. در واقع مقاله او نشان داد که در برآوردیابی همزمان میانگین‌های جوامع نرمال برآوردگر درستنمایی ماکزیمم یعنی میانگین نمونه غیر مجاز می‌باشد. او نشان داد که برای برآورد میانگین‌های جوامع نرمالی که از یک جهت شبیه یکدیگرند می‌توان برای به دست آوردن برآوردگری که دقت بیشتری از میانگین‌های نمونه دارد از سایر نمونه‌های مرتبط کمک گرفت.

مسأله ترکیب اطلاعات از منابع گوناگون خارج از روش درستنمایی قرارگرفته بود تا زمانی که تیبشیرانی و هاستی (۱۹۸۷) روش درستنمایی محلی[5] را در محدوده رگرسیون ناپارامتری ارائه نمودند. هو روش درستنمایی محلی را خارج از حوزه رگرسیون ناپارامتری مورد بررسی قرار داد که این روش درستنمایی وزنی مرتبط[6] (REWL) نامیده شد. علاوه بر این هو و زیدک (۲۰۰۰)

5) Local Likelihood 6) Relevance Weighted Likelihood

نشان دادند که چطور می‌توان برآوردگر معروف جیمز-استاین را به عنوان برآوردگر REWL استخراج نمود مشروط بر این که وزنها از نمونه برآورد شوند. بنابراین می‌توان روش درستنمایی را برای ترکیب اطلاعات نمونه‌های جوامع مختلف توسعه داد.

روش درستنمایی وزنی با روش درستنمایی کلاسیک متفاوت است. در روش کلاسیک فرض بر این است که مشاهدات از یک جامعه همسان خارج شده‌اند. در عوض درستنمایی وزنی در استنباط پارامتری زمانی پیش می‌آید که علاوه بر نمونه مورد نظر از جامعه مورد مطالعه نمونه‌های مرتبط اما مستقل از جوامع دیگر نیز در دسترس هستند که این همان مسأله مطلوب در برآوردیابی برای ناحیه کوچک می‌باشد. روش درستنمایی وزنی با اختصاص وزنهایی به این نمونه‌ها بر اساس میزان ارتباطشان اطلاعات آنها را ترکیب می‌کند.

در این مقاله به تولید برآوردگرهای همزمان میانگین‌های چندین توزیع می‌پردازیم که هر یک از این توزیعها متعلق به یک خانواده نمایی طبیعی با توابع واریانس درجه دوم (NEF-QVF)[7] هستند.

۳ روش درستنمایی وزنی در خانواده نمایی

فرض کنید m جامعه داریم. بردارهای نمونه‌های مستقل با $(y_{i1}, ..., y_{in_i})'$ $i = 1, ..., m$ نشان داده می‌شوند که $n_i \geq 0$. در عمل همیشه این امکان وجود دارد که از یک ناحیه محلی هیچ مشاهده‌ای گرفته نشود. می‌دانیم Y دارای توزیع نمایی با پارامتر θ است اگر

$$f(y, \theta) = exp[\theta y - \phi(\theta) + h(y)]$$

بنابراین اگر $\theta = (\theta_1, , , \theta_m)'$ و ω_{ij}^* وزن اختصاص یافته به نمونه jام در برآورد پارامتر iام باشد تابع درستنمایی به صورت زیر تعریف می‌شود:

$$\begin{aligned} WL(\theta) &= WL(\theta_1, ..., \theta_m) \\ &= \prod_{i=1}^{m} \prod_{j=1}^{m} \prod_{l=1}^{n_j} exp[(\theta_i y_{jl}) - \phi(\theta_i) + h(y_{jl})]^{\omega_{ij}^* n_j^{-1}} \end{aligned}$$

با فرض

$$\bar{y_j} = \frac{1}{n_j} \sum_{l=1}^{n_j} y_{jl} \qquad j = 1, ..., m$$

7) Natural Exponential Family with Quadratic Variance Function

داریم:

$$WL(\theta) = exp[\sum_{i=1}^{m}[\theta_i \sum_{i=1}^{m} \omega_{ij}^* \bar{y_j} - \phi(\theta_i) \sum_{j=1}^{m} \omega_{ij}^* + \sum_{j=1}^{m} \omega_{ij}^* n_j^{-1} \sum_{l=1}^{n_j} h(y_{jl})]] \quad (1)$$

بنابراین اگر ω_{ij}^* ها معلوم باشند

$$\frac{\partial LnWL(\theta)}{\partial \theta_i} = \sum_{j=1}^{m} \omega_{ij}^* \bar{y_j} - \phi'(\theta_i) \sum_{j=1}^{m} \omega_{ij}^* = 0$$

$$\frac{\partial^2 LnWL(\theta)}{\partial \theta_i^2} = -\phi''(\theta_i) \sum_{j=1}^{m} \omega_{ij}^* \leq 0$$

برآوردگرهای درستنمایی ماکزیمم وزنی (MWLE's) برای میانگینهای جامعه یعنی $\phi_i = \phi'(\theta_i)$ عبارتند از:

$$\hat{\phi}_{i_{WL}} = \frac{\sum_{j=1}^{m} \omega_{ij}^* \bar{y_j}}{\sum_{j=1}^{m} \omega_{ij}^*} = \sum_{j=1}^{m} \omega_{ij} \bar{y_j} \quad (2)$$

که

$$\omega_{ij} = \frac{\omega_{ij}^*}{\sum_{j=1}^{m} \omega_{ij}^*}$$

اما در عمل چون وزنها معلوم نیستند باید از روی داده‌ها برآورد شوند.

۱.۳ روش درستنمایی وزنی در مدل تعویض پذیر

در این حالت فرض کنید

$$\omega_{ii} = \omega \qquad i = 1, ..., m \qquad \omega_{ij} = \omega^* \qquad j \neq i = 1, ..., m$$

چنین فرضی این ایده را در بردارد که نمونه‌های سایر جوامع نسبت به اطلاعاتی که در مورد iامین پارامتر دارند تعویض پذیر[8] هستند. این فرض صریحاً به برآوردگر جیمز استاین در حالت نرمال زمانی‌که وزنها با استفاده از داده‌ها برآورد می‌شوند منجر می‌شود. تحت این فرض

$$\hat{\phi}_{i_{WL}} = \omega \bar{y_i} + \omega^* \sum_{j \neq i} \bar{y_j}$$

8) Exchangable

و چون مجموع وزنها برابر یک می‌باشد پس

$$\omega + (m-۱)\omega = ۱ \quad \Rightarrow \quad \omega^* = \frac{۱-\omega}{m-۱}$$

و

$$\hat{\phi}_{i_{WL}} = \omega\bar{y}_i + \frac{۱-\omega}{m-۱}\sum_{j\neq i}\bar{y}_j \qquad (۳)$$

برای پیدا کردن ω بهینه باید $g(\omega) = \sum_{i=۱}^{m} E(\hat{\phi}_i - \phi_i)^۲$ نسبت به ω می‌نیمم شود. با فرض

$$u_i = V(\bar{y}_i) = \frac{۱}{n_i}\phi''(\theta_i) \qquad \bar{u} = \frac{۱}{m}\sum_{i=۱}^{m} u_i$$

$$s_\phi^۲ = \frac{۱}{m-۱}\sum_{i=۱}^{m}(\phi_i - \bar{\phi})^۲$$

داریم:

$$g(\omega) = m[\omega^۲\bar{u} + (۱-\omega)^۲(m-۱)^{-۱}\bar{u} + m(۱-\omega)^۲(m-۱)^{-۱}s_\phi^۲] \qquad (۴)$$

$$\Rightarrow \quad \frac{\partial g(\omega)}{\partial\omega} = m[۲\omega\bar{u} - ۲(۱-\omega)(m-۱)^{-۱}\bar{u} - ۲m(۱-\omega)(m-۱)^{-۱}s_\phi^۲] = ۰$$

و لذا

$$\omega_{opt} = (\bar{u} + ms_\phi^۲)[m(\bar{u} + s_\phi^۲)]^{-۱} \qquad (۵)$$

با جایگذاری (۵) در رابطه (۳) داریم:

$$\hat{\phi}_{i_{WL}}^{opt} = (۱-\beta_{opt})\bar{y}_i + \beta_{opt}\bar{y} \qquad (۶)$$

که

$$\bar{y} = \frac{۱}{m}\sum_{i=۱}^{m}\bar{y}_i \qquad \beta_{opt} = \frac{\bar{u}}{\bar{u} + s_\phi^۲}$$

باز هم $\bar{u}$ و $s_\phi^۲$ نامعلومند و باید برآورد شوند.
فرض کنید:

$$s_y^۲ = \frac{۱}{m-۱}\sum_{i=۱}^{m}(y_i - \bar{y})^۲$$

آنگاه $E(s_y^۲) = \bar{u} + s_\phi^۲$. اما چون جوامع دارای توزیع نمایی با توابع واریانس درجه دوم هستند لذا

$$V(y_{jl}) = \nu_۰ + \nu_۱\phi_j + \nu_۲\phi_j^۲ = Q(\phi_j) \qquad l = ۱, ..., n_j \quad , \quad j = ۱, ..., m$$

$$E[Q(\bar{y}_j)] = \nu_۰ + \nu_۱\phi_j + \nu_۲[V(\bar{y}_j) + \phi_j^۲]$$

$$\Rightarrow E[Q(\bar{y}_j)] = Q(\phi_j) + \nu_۲ n_j^{-۱} Q(\phi_j) = (n_j + \nu_۲)u_j \qquad (۷)$$

پس

$$E[\frac{۱}{m}\sum_{j=۱}^{m}(n_j + \nu_۲)^{-۱}Q(\bar{y}_j)] = u$$

فرض کنید

$$\hat{\bar{u}} = \frac{۱}{m}\sum_{j=۱}^{m}(n_j + \nu_۲)^{-۱}Q(\bar{y}_j)$$

بنابراین β_{opt} را به وسیله $\hat{\beta}$ برآورد می‌کنیم که عبارتست از:

$$\hat{\beta} = min\{\frac{\hat{\bar{u}} + ۱/m}{s_y^۲ + ۱/m}, ۱\} \qquad (۸)$$

اکنون ϕ_i به صورت زیر برآورد می‌شود:

$$\hat{\phi}_{i_{WL}} = (۱ - \hat{\beta})\bar{y}_i + \hat{\beta}\bar{y} \qquad i = ۱, ..., m \qquad (۹)$$

تذکر: حالت خاصی را در نظر بگیرید که نمونه‌هایی با حجم مساوی n به طور مستقل از توزیعهای $N(\theta_i, \sigma^۲) \quad i = ۱, ..., m$ استخراج شده‌اند. و $\sigma^۲ > ۰$ معلوم است. در این صورت $\nu_۰ = \sigma^۲, \nu_۱ = \nu_۲ = ۰$ پس $\hat{\bar{u}} = \frac{\sigma^۲}{n}$.
بنابراین با جایگذاری در (۸) داریم:

$$\hat{\phi}_{i_{WL}} = (۱ - \frac{\sigma^۲/n}{s_y^۲})\bar{y}_i + \frac{\sigma^۲/n}{s_y^۲}\bar{y} \qquad i = ۱, ..., m$$

۲.۳ روش درستنمایی وزنی در مدل تعویض پذیر محلی

فرض کنید $\omega_{ij} = \omega_{i*} \quad j \neq i = ۱, ..., m \quad , \quad \omega_{ii} = \omega_i$ یعنی فرض تعویض پذیری سراسری بر داشته شده اما فرض تعویض پذیری درون هر ناحیه محلی باقی می‌ماند. آنگاه

$$\hat{\phi}_{i_{WL}} = \omega_i\bar{y}_i + \omega_{i*}\sum_{j\neq i}\bar{y}_j$$

و چون $\omega_i + (۱ - m)\omega_{i*} = ۱$ داریم:

$$\hat{\phi}_{i_{WL}} = \omega_i \bar{y}_i + (۱ - \omega_i)(۱ - m)^{-۱} \sum_{j \neq i} \bar{y}_j$$

برای به دست آوردن ω_iهای بهینه اکنون $g(\omega_i) = E(\hat{\phi}_{i_{WL}} - \phi_i)^۲$ را نسبت به ω_i به ازای هر $i = ۱, ..., m$ می‌نیمم می‌کنیم.

$$g(\omega_i) = \omega_i^۲ u_i + (۱-\omega_i)^۲ (m-۱)^{-۱}(m\bar{u}-u_i) + m^۲(m-۱)^{-۲}(۱-\omega_i)^۲(\phi_i-\bar{\phi})^۲$$

$$g'(\omega_i) = ۲\omega_i u_i - ۲(۱-\omega_i)(m-۱)^{-۱}(m\bar{u}-u_i) - ۲m^۲(m-۱)^{-۲}(۱-\omega_i)(\phi_i-\bar{\phi})^۲$$

$$g''(\omega_i) = ۲u_i + \frac{۲}{m-۱}(m\bar{u} - u_i) + \frac{۲m^۲}{(m-۱)^۲}(\phi_i - \bar{\phi})^۲ > ۰$$

پس

$$\omega_{i_{opt}} = \frac{m\bar{u} - u_i + m^۲(\phi_i - \bar{\phi})^۲}{(m-۱)^۲ u_i + m\bar{u} - u_i + m^۲(\phi_i - \bar{\phi})^۲}$$

بنابراین

$$\hat{\phi}^{opt}_{i_{WL}} = (۱ - \beta_{i_{opt}})\bar{y}_i + \beta_{i_{opt}}\bar{y} \qquad (۱۰)$$

که

$$\beta_{i_{opt}} = \frac{(m-۱)u_i}{(m-۲)u_i + \bar{u} + m(\phi_i - \bar{\phi})^۲}$$

اما در عمل β_iها معلوم نیستند و باید برآورد شوند. توجه کنید که

$$\begin{aligned} E(\bar{y}_i - \bar{y})^۲ &= V(\bar{y}_i - \bar{y}) + (\phi_i - \bar{\phi})^۲ \\ &= \frac{۱}{m}[(m-۲)u_i + \bar{u} + m(\phi_i - \bar{\phi})^۲] \end{aligned}$$

همچنین

$$E[Q(\bar{y}_i)] = \nu_۰ + \nu_۱\phi_i + \nu_۲(n_i + \nu_۲) = (n_i + \nu_۲)u_i$$

بنابراین $E[Q(\bar{y}_i)]/(n_i + \nu_۲) = u_i$. از این رو $\beta_{i_{opt}}$ به وسیله $\hat{\beta}_i$ برآورد می‌شود که

$$\hat{\beta}_i = max[\frac{m-۱}{m}\frac{Q(\bar{y}_i)(n_i + \nu_۲)^{-۱} + m^{-۱}}{(\bar{y}_i - \bar{y})^۲ + m^{-۱}}, ۱]$$

و لذا

$$\hat{\phi}_{i_{WL}} = (۱ - \hat{\beta}_i)\bar{y}_i + \hat{\beta}_i\bar{y}$$

۳.۳ روش درستنمایی وزنی بدون فرض تعویض پذیری

حالتی را در نظر بگیرید که هیچگونه فرض تعویض پذیری وجود ندارد. اگر $\hat{\phi}_i = \sum_{j=1}^{m} \omega_{ij}\bar{y}_j$ پس ω_{ij}های بهینه با می‌نیمم کردن $E(\sum_{j=1}^{m} \omega_{ij}\bar{y}_j - \phi_i)^2$ نسبت به ω_{ij} به دست می‌آیند با این شرط که $\sum_{j=1}^{m} \omega_{ij} = 1$. فرض کنید $\Omega_i = (\omega_{i1}, ..., \omega_{im})'$ و فرض کنید

$$g(\Omega_i) = E(\sum_{j=1}^{m} \omega_{ij}\bar{y}_j - \phi_i)^2 - 2\lambda_i(\sum_{j=1}^{m} \omega_{ij} - 1)$$

که λ_iها ضرایب لاگرانژ می‌باشند.

$$g(\Omega_i) = \sum_{j=1}^{m} \omega_{ij}^2 u_j + (\sum_{j=1}^{m} \omega_{ij}\phi_j - \phi_i)^2 - 2\lambda_i(\sum_{j=1}^{m} \omega_{ij} - 1)$$

بنابراین

$$\frac{\partial g}{\partial \omega_{ij}} = 2\omega_{ij}u_j + 2(\sum_{j=1}^{m} \omega_{ij}\phi_j - \phi_i)\phi_j - 2\lambda_i$$

$$\frac{\partial^2 g}{\partial \omega_{ij}^2} = 2u_j + 2\phi_j^2 > 0 \qquad \frac{\partial^2 g}{\partial \omega_{ij}\partial \omega_{kl}} = 0 \quad j \neq l$$

از این رو ماتریس همسیان برای g معین مثبت است و ω_{ij}های بهینه با حل معادلات $\frac{\partial g}{\partial \omega_{ij}} = 0 \quad i, j = 1, ..., m$ به دست می‌آیند. با استفاده از نماد ماتریسی این معادلات به صورت زیر نوشته می‌شوند:

$$M\Omega_i = \lambda_i 1_m + \phi_i \Phi$$

که

$$M = D + \Phi\Phi' \ , \quad D = Diag(u_1, ..., u_m) \quad , \quad \Phi = (\phi_1, ..., \phi_m)'$$

بنابراین Ω_i بهینه به صورت زیر به دست می‌آید.

$$\Omega_i^{opt} = M^{-1}(\lambda_i 1_m + \phi_i \Phi) \qquad (11)$$

چون $\sum_{j=1}^{m} \omega_{ij} = 1$ بنابراین

$$1 = 1_m'\Omega_i^{opt} = \lambda_i 1_m' M^{-1} 1_m + \phi_i 1_m' M^{-1}\Phi$$

$$\lambda_i = \frac{۱ - \phi_i ۱'_m M^{-۱} \Phi}{۱'_m M^{-۱} ۱_m} \qquad (۱۲)$$

اکنون با استفاده از (۱۰) و (۱۱) داریم:

$$\Omega_i^{opt} = \frac{۱ - \phi_i ۱'_m M^{-۱} \Phi}{۱'_m M^{-۱} ۱_m} (M^{-۱} ۱_m) + \phi_i M^{-۱} \Phi$$

و

$$M^{-۱} = D^{-۱} - \frac{D^{-۱} \Phi \Phi' D^{-۱}}{۱ + \Phi' D^{-۱} \Phi}$$

حال چون $E[(n_j + \nu_۲)^{-۱} Q(\bar{y}_j)] = u_j$ که $E \quad (\bar{y}_i) = \phi_j \quad , \quad$

$$Q(\bar{y}_j) = \nu_۰ + \nu_۱ \bar{y}_j + \nu_۲ \bar{y}_j^{۲}$$

داریم:

$$\hat{u}_j = (n_j + \nu_۲)^{-۱} Q(\bar{y}_j)$$

$$\hat{\Omega}_i^{opt} = \frac{۱ - \bar{y}_i ۱'_m \hat{M}^{-۱} Z}{۱'_m \hat{M}^{-۱} ۱_m} (\hat{M}^{-۱} ۱_m) + \bar{y}_i \hat{M}^{-۱} Z$$

که

$$Z = (\bar{y}_۱, ..., \bar{y}_m)' \quad , \quad \hat{M} = (\hat{D} + ZZ')^{-۱} \quad , \quad \hat{D} = Diag(\hat{u}_۱, ..., \hat{u}_m)$$

مراجع

[1] Ghosh, M., and Rao, J.N.K. (1994), Small area estimation: an appraisal (with discussion), Statistical Science, 9, 65-93.

[2] Hu, F. and Zidek, J.V. (2000), The relevance weighted likelihood with applications.In Empirical Bayes and Likelihood Inference. Lecture Notes in Statistics. Eds. S.E. Ahmed and N. Reid. Springer-Verlag, New York, pp 211-235.

[3] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In proceedings of the Third Berkeley Symposium in Mathematical Statistics and Probability, V1. University of California Press, pp 197-206.

[4] Tibshirani. R. and Hastie, T. (1987). Local likelihood estimation. J. Amer. Statist. Assoc., 82, 559-567.

رگرسیون چندک

فرهاد فتاحی[۱]، عباس گرامی[۲]

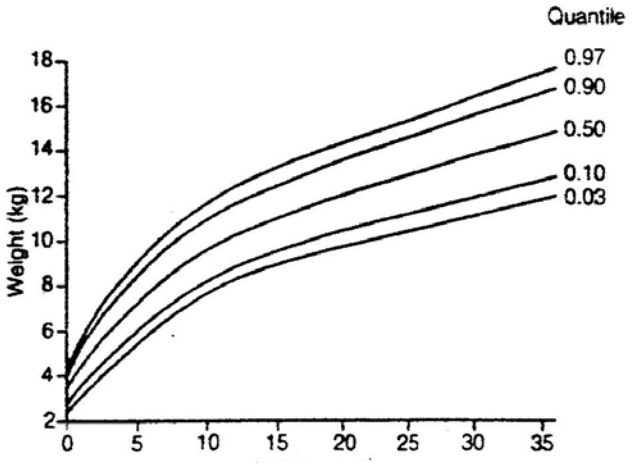
[۱] دانشگاه تربیت مدرس
[۲] پژوهشکده آمار

چکیده: رگرسیون چندک را می‌توان بعنوان تعمیم طبیعی برآورد حداقل مربعات کلاسیک مدلهای میانگین شرطی در برآورد مجموعه‌ای از مدلها، برای چندک شرطی مد نظر قرار داد. مورد خاص و عمده برآوردگر رگرسیون میانه می‌باشد که مجموع قدرمطلق خطاها را مینیمم می‌نماید. سایر توابع چندک شرطی با مینیمم کردن یک مجموع موزون نامتقارن از قدر مطلق خطاها برآورد می‌گردد. با کنار هم قرار گرفتن کلیهٔ توابع برآورد شده چندک شرطی می‌توان دید بسیار کامل‌تری در مورد اثر متغیرهای کمکی بر روی پارامترهای موقعیت، مقیاس و شکل توزیع متغیر پاسخ بدست آورد.

واژه‌های کلیدی: رگرسیون چندک، تابع چندک تجربی، چندک رگرسیونی، چندک نمونه‌ای، چندک شرطی

۱ مقدمه

در شکل (۱) چندکهای مختلف وزن نوزادان پسر در مقابل سن‌شان رسم شده است. مسئله مورد علاقه ممکن است فقط بر ارتباط متوسط این دو متغیر متمرکز نشده باشد، بلکه ممکن است بر ارتباط بین چندکهای آخری و سن مبتنی باشد. ممکن است چندکها بعنوان کرانهای مناسب مشکلات تغذیه‌ای بالقوه نوزادان را منعکس سازند بنابراین وقتی که فرایند افزایش وزن را در طول زمان مورد آزمایش قرار می‌دهیم بسیار سودمند واقع می‌گردد. واضح است که در این حالت و بسیاری حالات مشابه، ممکن است یک مدل رگرسیون استاندارد برای پیش‌بینی ارتباط بین کلیه چندکهای یک متغیر وابسته Y از یک سو و بردار متغیرهای مستقل X از سوی دیگر مناسب نباشد. به همین دلیل کانکر، باست و هاگ روش رگرسیون (MAD) یعنی میانگین قدر مطلق انحرافات را برای رگرسیون چندک تعمیم دادند که در آن تحت فرض خطی بودن برآوردگرها، روشی را برای مدل‌بندی چندکهای تابع توزیع شرطی $F(y|X)$ بیان کردند. در تشریح این روش ضرورت دارد چگونگی ارتباط بین چندکهای معمولی به عنوان جوابی برای معادلات برآورد مشخص شود که بواسطه آن یک تابع زیان نمونه‌ای خاص مینیمم می‌گردد. این امر ما را در بدست آوردن برآورد مدلهای رگرسیونی برای چندکهای شرطی قادر می‌سازد. با در دست داشتن نمونه‌ای nتایی از



شکل ۱: نمودار سن و وزن

مشاهدات یک متغیره $\{y_1, y_2, \ldots, y_n\}$، pامین چندک نمونه‌ای $\hat{\mu}_p$، $0 < p < 1$ باید در رابطه زیر صدق کند.

$$\sum_{i=1}^{n}[pI(y_i \geq \hat{\mu}_p) - (1-p)I(y_i < \hat{\mu}_p)] = 0 \qquad (1)$$

در رابطه بالا I تابع نشانگر بوده که در صورتی که شرط درون پرانتز برقرار باشد مقدار یک می‌گیرد در غیر اینصورت مقدار صفر را اختیار می‌کند. توجه کنید که $\hat{\mu}_p$ برآوردی از چندک جامعه متناظر خود یعنی μ_p می‌باشد که در رابطه زیر صدق می‌کند:

$$p[1 - F(\mu_p) - (1-p)F(\mu_p)] = 0$$

رابطه (۱) را می‌توان بصورت یک معادله برآوردکننده نیز نشان داد که به باقیمانده‌های مثبت $r_i = y_i - \hat{\mu}_p$ وزن p و به باقیمانده‌های منفی وزن $1-p$ داده می‌شود و مقدار مورد نظر توسط علامت باقیمانده اندازه‌گیری می‌شود بنابراین رابطه (۱) را می‌توان به فرم $\sum_{i=1}^{n}\psi_p(r_i) = 0$ نشان داد. که در آن

$$\psi_p(r) = \begin{cases} p\psi(r) & r > 0 \\ (1-p)\psi(r) & o.w \end{cases} \qquad (2)$$

گاهی اوقات $\psi(r)$ را بعنوان تابع تأثیر در برآورد چندکهای رگرسیونی با تابع زیان نامتقارن متناظر با آن که به فرم زیر است بکار می‌برند.

$$\rho_p(r) = \begin{cases} p\rho(r) & r > ۰ \\ (۱-p)\rho(r) & o.w \end{cases} \qquad (۳)$$

که در آن $\rho(r) = |r|$. رابطه برآوردکننده فوق و به تبع آن $\hat{\mu}_p$، را می‌توان با مینیمم‌کردن متوسط تابع زیان نمونه‌ای زیر بدست آورد.

$$n^{-۱} \sum_{i=۱}^{n} \rho_p(y_i - \mu_p) \qquad (۴)$$

برای اینکه روش بالا را به رگرسیون چندک تعمیم دهیم، روشی همانند رگرسیون کمترین مربعات خطی را در پیش می‌گیریم و μ_p موجود در رابطه (۴) را به $\mu_p(X) = \beta x' \beta_p$ تبدیل می‌کنیم، که در آن β_p یک بردار ضرائب رگرسیونی برای pامین چندک شرطی است. می‌توان برآورد β_p را با حل رابطه زیر بدست آورد.

$$\sum_{i=۱}^{n} \psi_p(y_i - \underline{x}_i' \hat{\beta}_p) \underline{x}_i = ۰ \qquad (۵)$$

در حالت خاص $p = ۰/۵$، سطح برازش داده شده معادل با میانه رگرسیون y نسبت به X است که رابطه زیر را مینیمم می‌کند (که عبارتست از متوسط قدرمطلق انحرافات باقیمانده‌ها).

$$n^{-۱} \sum_{i=۱}^{n} |y_i - X_i' \hat{\beta}_{۰/۵}| \qquad (۶)$$

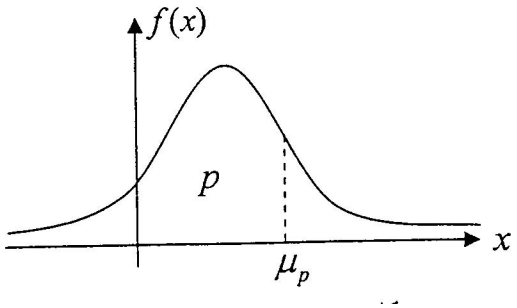
۲ معرفی چندک

فرض کنید متغیر تصادفی X، دارای تابع توزیع تجمعی $F(x)$ باشد پارامتر μ_p را چندک مرتبه p برای $F(x)$ یا برای X، می‌نامیم هرگاه نامساوی دو طرفه زیر را داشته باشیم.

$$p(X < \mu_p) \leq p \leq p(X \leq \mu_p) \qquad ۰ < p < ۱$$

معنی این نامساوی دو طرفه این است که مقدار احتمال در فاصله باز $(-\infty, \mu_p)$ حداکثر p و در فاصله نیم باز $[-\infty, \mu_p)$ حداقل p می‌باشد (شکل۲).

μ_p را که در آن $۰ < p < ۱$ است، چندک مرتبه p می‌نامند هرگاه تقریباً $۱۰۰p$ درصد داده‌ها کوچکتر از آن باشند.



شکل ۲: نمودار چندک p

چندکها کلی تر از میانه می‌باشند و در حقیقت $p_{0/5}$ همان میانه است به زبان هندسی اگر از نقطه μ_p خطی به موازات محور y ها رسم کنیم مساحت زیر منحنی فراوانی که در سمت چپ این خط قرار دارد برابر p می‌باشد. چندکهای معروف عبارتند از:

الف) چارکها که به ازای $p = 0/25, 0/5, 0/75$ بدست می‌آیند و آنها را به ترتیب با $\mu_{0/25}, \mu_{0/5}, \mu_{0/75}$ نشان می‌دهند.

ب) دهکها که به ازای $p = 0/1, \ldots, 0/9$ بدست می‌آیند و آنها را به ترتیب با $D_1, \ldots, D_9$ نشان می‌دهند.

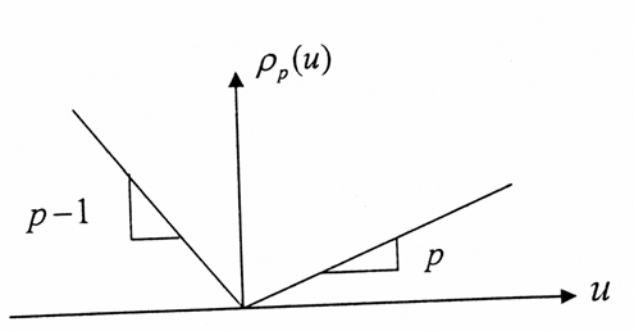
ج) صدکها که به ازای $p = 0/01, \ldots, 0/99$ بدست می‌آیند و آنها را به ترتیب با $p_1, \ldots, p_{99}$ نشان می‌دهند.

هر متغیر تصادفی را می‌توان توسط تابع توزیعش $F(x) = p(X \leq x)$ مشخص کرد و برای هر $0 < p < 1$ چندک pام متغیر X برابر $F^{-1}(p) = inf\{x : F(x) \geq p\}$ که در حالت $p = \frac{1}{2}$ میانه $F^{-1}(\frac{1}{2})$ را خواهیم داشت.

چندکها از یک مسئله بهینه‌سازی ساده ناشی می‌شوند به این صورت که اگر به ازای هر $p \in (0, 1)$ تابع زیان را بصورت $\rho_p(u) = u(p - I(u < 0))$ تعریف کنیم شکل این تابع زیان بصورت (شکل ۳) خواهد بود.

در اینجا هدف یافتن برآورد نقطه‌ای x یعنی $\hat{x}$ است، بطوریکه با انتخاب $\hat{x}$، امید زیان مینیمم شود بنابراین اگر امید زیان بصورت زیر تعریف شود:

$$M = E\rho_p(x - \hat{x}) = (p - 1)\int_{-\infty}^{\hat{x}} (x - \hat{x})dF(x) + p\int_{\hat{x}}^{\infty} (x - \hat{x})dF(x)$$



شکل ۳: شکل تابع زیان

با مشتق‌گیری نسبت به $\hat{x}$ داریم:

$$\frac{dM}{d\hat{x}} = (1-p)\int_{-\infty}^{\hat{x}} dF(x) - p\int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - p$$

چون تابع F یکنواست بنابراین هر عضو مجموعه $\{x : F(x) = p\}$ امید زیان را مینیمم می‌کند. در صورتیکه $\hat{x} = F^{-1}(p)$ یکتا باشد جواب مورد نظر بدست می‌آید. در غیر اینصورت فاصله‌ای از چندکهای pام را داریم و با توافق به این قرار داد که تابع چندک تجربی از طرف چپ پیوسته است. پس کوچکترین عضو این مجموعه را به عنوان جواب ممکن انتخاب می‌کنیم. پیداکردن بهترین برآوردگر نقطه‌ای برای زیان خطی نامتقارن، منجر به مفهوم چندکها می‌شود. در صورت متقارن بودن زیان، قدر مطلق آن، میانه را به خوبی مشخص خواهد کرد. برای مثال اگر مقادیر کوچکتر از برآورد، سه برابر مقادیر بزرگتر از برآورد باشد، $\hat{x}$ را به گونه‌ای انتخاب می‌کنیم که مقدار احتمال $P(X \leq x)$ سه برابر مقدار احتمال $P(X > x)$ باشد یعنی $\hat{x}$ را صدک ۷۵ام توزیع F انتخاب می‌کنیم. اکنون اگر تابع توزیع تجربی،

$$F_n(x) = n^{-1}\sum_{i=1}^{n} I(X_i \leq x)$$

جایگزین F شود می‌توان $\hat{x}$ را با مینیمم کردن امید زیان،

$$\int (x-\hat{x})dF_n(x) = n^{-1}\sum_{i=1}^{n} \rho_p(x_i - \hat{x})$$

بدست آورد در اینجا وقتی pn یک عدد صحیح باشد، بازه‌ای از جوابهایی را که به صورت $\{x : F_n(x) = p\}$ هستند، خواهیم داشت بنابراین جواب، دقیق و روشن نخواهد بود و خواهیم دید که این نتیجهٔ زیاد سودمندی هم نیست. اکنون حالت ساده‌ای از چندکهای نمونه‌ای مرسوم را با یک مقدار جزئیات بیشتر شرح می‌دهیم. می‌توان مسئله پیدا کردن چندک نمونه‌ای pام را بصورت:

$$\min_{\mu \in R} \sum_{i=1}^{n} \rho_p(y_i - \mu) \qquad (۷)$$

که در آن $\rho_p(u) = u(p - I(u < 0))$ است بیان کرد که ممکن است بصورت زیر بعنوان یک مسئله برنامه‌ریزی خطی با معرفی $2n$ متغیر غیر واقعی u_i وv_i، $\{u_i, v_i : i = 1, \ldots, n\}$، که u_i و v_i قسمتهای مثبت و منفی بردار باقیمانده‌ها را نشان می‌دهند فرموله شود.

$$\min_{(\mu,u,v) \in R \times R^{+2n}} \{p 1_n' u + (1 - p) 1_n' v | I_n p + u - v = y\} \qquad (۸)$$

$1_n'$ برداری n بعدی از یک‌ها می‌باشد. بوضوح می‌بینیم که در (۸) یک تابع خطی روی یک مجموعه محدودیت چند وجهی مینیمم می‌شود.این مجموعه محدودیت چند وجهی شامل اشتراک ابر صفحه‌های $(1 + 2n)$ بعدی تولید شده، با شرایط تساوی و مجموعه $R \times R^{2n}$ است. در تابع هدف (۷) در صورتی جواب بهینه خواهیم داشت که مشتقات راست و چپ $R(\mu) = \sum_{i-1}^{n} \rho_p(y_i - \mu)$ که در زیر بدست آمده‌اند هر دو نامنفی باشند.

$$\acute{R}(\mu, +1) = \lim_{h \to 0} \frac{(R(\mu + h) - R(\mu))}{n} = \sum_{i=1}^{n} I((y_i < \mu + h) - \mu) \qquad (۹)$$

$$\acute{R}(\mu, -1) = \lim_{h \to 0} \frac{(R(\mu - h) - R(\mu))}{n} = \sum_{i=1}^{n} (\mu - I((y_i < \mu - h)) \qquad (۱۰)$$

یعنی اگر عدد np در فاصله بسته $[N^-, N^+]$ قرار داشته باشد $N^{\pm} = \{y_i < \mu \pm 0\}$ در اینصورت $\hat{\mu}_p$ بین دو آماره ترتیبی مجاور هم قرار می‌گیرد و وقتی مقدار این آماره‌های ترتیبی مجاور برابر شود در اینصورت $\hat{\mu}_p$ یکتا است.

۳ چندکهای نمونه‌ای

اگر بخواهیم اطلاعات بیشتری را در مورد یک مجموعه از مشاهدات ارائه دهیم، ساده‌ترین کار، محاسبه چندکهای نمونه‌ای است. در اینجا هدف معرفی ابتدائی چندکهای نمونه‌ای است که همانا در مرتب شدن مشاهدات نقش بسزایی دارند. اگر $\{Y_t : t = 1, \ldots, n\}$ یک نمونه

تصادفی از متغیر تصادفی Y با تابع توزیع $F(\circ)$ باشد آنگاه برای $\circ < p < ۱$، pامین چندک نمونه‌ای، عبارتست از هر جوابی که از مینیمم کردن عبارت زیر بدست می‌آید.

$$\min_{\mu_p \in R}[\sum_{t \in \{t: y_t \geq \mu\}} p(y_t - \mu_p) + \sum_{t \in \{t: y_t \leq \mu\}} (p-۱)(y_t - \mu_p)] \qquad (۱۱)$$

در حالت خاص $p = \frac{۱}{۲}$، میانه نمونه، $\circ/۵\circ$امین چندک نمونه‌ای است. حال این تعریف را به مدلهای خطی تعمیم داده و چندکهای رگرسیونی را تعریف می‌کنیم.

۴ چندکهای رگرسیونی

رابطه ریاضی مدل رگرسیونی خطی:

$$y_i = \beta_۱ x_{i۱} + \beta_۲ x_{i۲} + \ldots + \beta_p xip + u_i \qquad x_{i۱} = ۱, i = ۱, ۲, \ldots, n$$

را در نظر می‌گیریم که u_iها دارای تابع توزیع $F(\circ)$ باشند. فرم ماتریسی رابطه عبارتست از:

$$y = X\beta + u$$

که $u = (u_۱, \ldots, u_n)'$ بردار خطا، $\beta = (\beta_۱, \ldots, \beta_p)'$ بردار پارامترهای مجهول، X یک ماتریس $n \times p$ معلوم رتبه کامل ستونی و $y = (y_۱, \ldots, y_n)'$ بردار مشاهدات متغیر وابسته است. اگر $\{X_t, t = ۱, \ldots, n\}$ نشان دهنده دنباله‌ای از بردارهااز سطرهای ماتریس X و $\{y_t, t = ۱, \ldots, n\}$ نشان دهنده اعداد تصادفی از فرایند رگرسیونی $u_t = y_t - x_t\beta$ با تابع توزیع $F(\circ)$ باشد آنگاه برای $\circ < p < ۱$، pامین چندک رگرسیونی بصورت،

$$\min_{\beta \in R^p}[\sum_{t \in \{t; y_t \geq x_t'\beta\}} p(y_t - x_t'\beta) + \sum_{t \in \{t; y_t < x_t'\beta\}} (p-۱)(y_t - x_t'\beta)] \qquad (۱۲)$$

بدست می‌آید. که آن را با نماد $\beta(p)$ نشان می‌دهیم. در اینجا ساختار عبارت (۱۲) را بر اساس تعریف چندکهای نمونه‌ای (۱۱) بیان کردیم. ولی در عمل برای ساده‌تر نشان دادن عبارت (۱۲) ابتدا تابع زیان $\rho_p(\circ)$ را بصورت $\rho_p(u) = u(p - I(u < \circ))$ تعریف می‌کنیم. سپس $\beta(p)$ را از مینیمم کردن تابع زیر روی β در فضای R^p برای مقادیر ثابت p، $\circ < p < ۱$ بدست می‌آوریم.

$$\min_{\beta \in R^p} \sum_{i=۱}^{n} \rho_p(y_i - x_i'\beta) \qquad (۱۳)$$

در عمل عبارت (۱۳) با عبارت (۱۴) یکسان است. با در نظرگرفتن $\hat{\beta}(p)$، برآورد $\beta(p)$، برآورد تابع چندک شرطی pام، $\hat{\Phi}_Y(p|x) = x'\hat{\beta}(p)$ بدست می‌آید. می‌توان مسئله رگرسیون چندک (۱۳) را بصورت یک مسئله برنامه‌ریزی خطی مشابه برنامه‌ریزی خطی بخش قبل به صورت زیر فرموله کرد.

$$\min_{(\beta,u,v)\in R^p\times R^{+2n}} \{pI_n'u + (1-p)I_n'v | X\beta + u - v = y\} \qquad (۱۴)$$

که X ماتریس طرح رگرسیون $n \times p$ است. در اینجا هم دوباره تابع خطی روی یک مجموعه محدودیت چند وجهی مینیمم می‌شود و بیشتر خصوصیات شناخته شده $\hat{\beta}(p)$ از خصوصیات شناخته شدهٔ جوابهای برنامه‌ریزی خطی پیروی می‌کند.

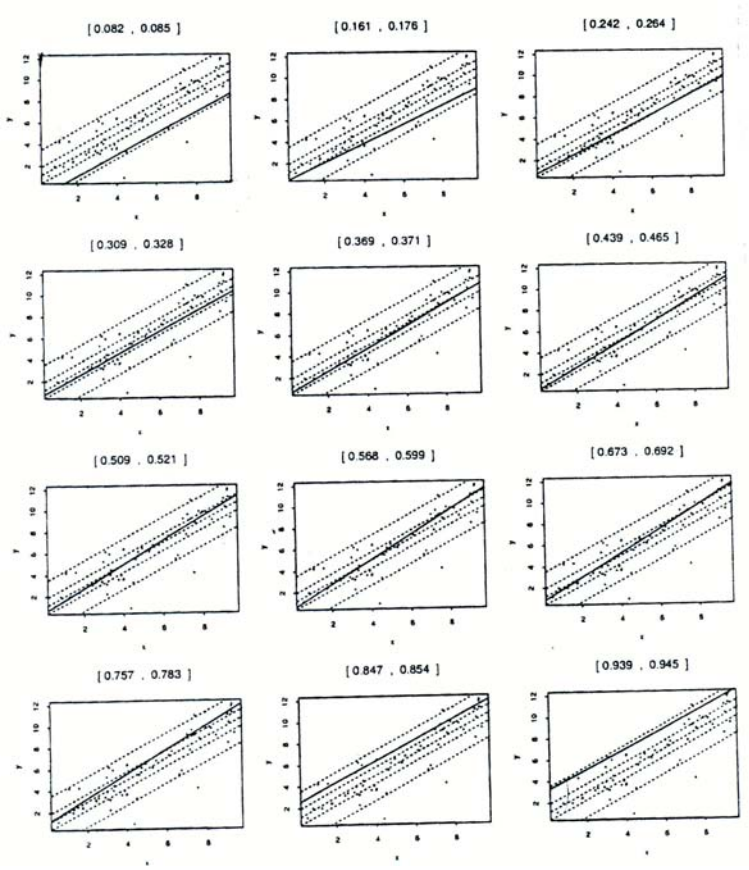
۵ مثال ۱

یک مدل رگرسیون بصورت، $y_i = \beta_0 + x_i'\beta_1 + u_i$ داریم که در آن خطاها هم توزیع ومستقل از هم هستند. بنابراین تابع چندک y_i بصورت $\Phi(p|x) = \beta_0 + x'\beta_1 + F_u^{-1}(p)$ خواهد بود که F_u تابع توزیع مشترک خطاها است. $\hat{\beta}(p)$ برآورد پارامترهای جامعه می‌باشد. در شکل (۴)، ۶۰ مشاهده تولید شده از توزیع نرمال با خطاهای مستقل و هم توزیع، بهمراه خطوط چندک رگرسیون برازش شده متناظر با $p \in \{۰/۰۵, ۰/۲۵, ۰/۵, ۰/۷۵, ۰/۹۵\}$ نشان داده شده است. خط ممتد در هر شکل، چندک شرطی برآورد شده را برای فاصله p (که در بالای هر شکل نشان داده شده است) شرح می‌دهد. می‌بینیم که با افزایش p، این خط ممتد با شیبی تقریباً برابر با شیب توابع چندک شرطی، در بین داده‌ها تغییر می‌کند و به سمت خطوط چندک شرطی بالاتر می‌رود. در بالای هر شکل بازه‌ای برای هر p نشان داده شده است که پاسخ متناظر آن بهینه است.

۶ اساس کار رگرسیون و رگرسیون چندک

تعبیر شهودی نحوه کار رگرسیون معمولی بیشتر بر اساس تعبیر هندسی طرح حداقل مربعات است. مینیمم کردن فاصله اقلیدسی $\|Y - \hat{Y}\|$ روی تمام yها، در فضای خطی تولید شده ستونهای X را می‌توان به این تشبیه کرد که توپی با مرکز y آنقدر باد می‌شود تا به زیر فضای تولید شده توسط X مماس شود. جایگزینی توپهای اقلیدسی با لوزیهای چند وجهی فاصله ρ_p ، بطوریکه، $d_p(y, \hat{y}) = \sum_{i=1}^n \rho_p(y_i - \hat{y}_i)$ بعضی مسائل جدید را مطرح می‌کند اما از بسیاری جهات بصورت قبل عمل می‌کند. در روش حداقل مربعات برای بدست آوردن معادلات نرمال عبارت زیر را مینیمم می‌کردیم:

$$\|y - \hat{y}(\beta)\|^2 = (y - X\beta)'(y - X\beta)$$



شکل ۴: برازش چندکهای رگرسیونی

با مشتق گیری نسبت بهβ:

$$\nabla \|y, \hat{y}(\beta)\|^{۲} = X^{'}(y - X\beta)) = ۰ \quad \longrightarrow \quad \hat{\beta} = (X^{'}X)^{-۱}X^{'}y$$

$\hat{\beta}$ بدست می‌آمد. در رگرسیون چندک هم با شیوه‌ای مشابه عمل می‌کنیم یعنی از تابع:

$$R(\beta) = d_p(y, \hat{y}(\beta)) = \sum_{i=۱}^{n} \rho_p(y_i - x_i\beta) \qquad (۱۵)$$

بر حسب β مشتق می‌گیریم و با توجه به اینکه ممکن است مشتقات، به جهت بستگی داشته باشند، مشتقات سوئی زیر را در نظر می‌گیریم:

$$\begin{aligned} \nabla R(\beta, w) &= \frac{d}{dt} R(\beta + tw)|_{t=۰} \\ &= \frac{d}{dt} \sum_{i=۱}^{n} u_i(\beta + tw)[p - I_{(u_i(\beta+tw)<۰)}] \\ &= -\sum_{i=۱}^{n} \psi^*(y_i - x_i^{'}\beta, -x_i^{'}w)x_i^{'}w \qquad (۱۶) \end{aligned}$$

که در آن:

$$\psi^*(u, v) = \begin{cases} p - I(u < ۰) & if \quad u \neq ۰ \\ p - I(v < ۰) & if \quad u = ۰ \end{cases} \qquad (۱۷)$$

اگر به ازای تمام $w \in R^p$ و با فرض $\|w\| = ۱$، داشته باشیم $\nabla R(\hat{\beta}, w) \geq ۰$ ، در این صورت $\hat{\beta}$ تابع $R(\beta)$ را مینیمم می‌کند.

اساس تعریف چندکهای رگرسیونی:

همانگونه که جستجو برای یافتن میانه، ما را به یک مشاهده میانی یکتا یا یک جفت از مشاهدات وسطی مجاور هدایت می‌کند که هر دو جواب، مجموع قدر مطلق باقی مانده‌ها را مینیمم می‌کنند. در رگرسیون چندک هم به جستجوی زیر مجموعه‌های p عضوی که جواب را مشخص می‌کنند، رهنمون می‌شویم. در اصطلاحات برنامه‌ریزی خطی، به این زیر مجموعه‌های p عضوی، جوابهای پایه گفته می‌شود. در بررسی این زیر مجموعه‌های p عضوی از مشاهدات چند نماد را معرفی می‌کنیم $B(p)$ را مجموعه تمام نقاطی در نظر می‌گیریم که عبارت $\sum_{i=}^{n} \rho_p(y_i - x_i\beta)$ را مینیمم می‌کند و $\beta(p)$ را یک عضو این مجموعه تعریف می‌کنیم. همچنین فرض می‌کنیم

$N = \{۱, ۲, \ldots, n\}$ و H نشاندهندهٔ مجموعه تمام زیر مجموعه‌های p عضوی از N ، باشد که اعضای آن را با h و متمم مجموعه h را با $h^- = N - h$ نشان می‌دهیم. ماتریس X و بردار y را بر این اساس تقسیم بندی می‌کنیم. $X(h)$ نشان دهندهٔ یک ماتریس $p \times p$ با سطرهای $\{x_i : i \in h\}$ است و حال آنکه $y(h)$ یک بردار p بعدی با مختصات $\{y_i : i \in h\}$ است. ip یک بردار p بعدی با عناصر یک است یعنی $i'p = (۱, ۱, \ldots, ۱)$ و در نهایت H^*، مجموعهٔ زیر مجموعه‌های p عضوی از N می‌باشد که به ازای هر $h \in H$،$X(h)$، رتبه کامل باشد. یعنی

$$H^* = \{h \in H | rank X(h) = p \ ; \quad |X(h)| \neq ۰\}$$

اکنون با بکارگیری این نمادها، می‌توان هر جواب پایه‌ای را که از نقاط $\{(x_i, y_i) : i \in h\}$ ، می‌گذرد با فرض ناویژه بودن، ماتریس $X(h)$ بصورت $b(h) = X^{-۱}(h)y(h)$ بیان کرد.

تعریف: در مشاهدات رگرسیونی (y, X)، اگر به ازای هر p مشاهده، یک برازش دقیق و منحصر بفرد داشته باشیم در موقعیت عمومی هستند یعنی برای هر $h \in H^* = \{h \in H \ ; \ |X(h)| \neq ۰\}$ داشته باشیم:

$$y_i - x_i b(h) = ۰ \qquad \forall i \in h$$

خواص چندکهای رگرسیونی:

قضیه ۱: اگر X دارای رتبه p باشد و برای بعضی از عناصر H^*، مجموعه $B(p)$، حداقل یک عضو به صورت $b(h) = X^{-۱}(h)y(h)$ داشته باشد، آنگاه $B(p)$، یک پوسته محدب می‌باشد و تمام اعضای آن دارای همین شکل هستند. این قضیه شکل چندکهای رگرسیونی را تحت یک قانون کلی بیان می‌کند.

قضیه ۲: اگر (y, X) در موقعیت عمومی باشند، آنگاه یک جواب منحصر بفرد برای مسئله چندک رگرسیونی بصورت $b(h) = X^{-۱}(h)y(h)$ وجود دارد اگر و فقط اگر به ازای یک مقدار $h \in H^* = \{h \in H \ ; \ |X(h) \neq ۰\}$ داشته باشیم:

$$(p - ۱)I_p \leq \mu_h \leq pI_p$$

که در آن: $\mu_h = -\sum_{i \in h} \psi_p (y_i - x_i' b(h) \quad , \quad - x' i X^{-۱}(h) v x_i' X^{-۱}(h))$

$$\psi_p = p - I(u < ۰) \qquad , \quad I'p = (۱, ۱, \ldots, ۱)_{۱ \times p}$$

علاوه بر این، $b(h)$ جواب یکتاست اگر و فقط اگر نامعادله‌ها اکید باشند در غیر اینصورت مجموعه جواب $B(p)$، پوسته محدبی از جوابهای متعددی به شکل $b(h)$ است.

توجه: تباهیدگی اولیه در مسئله رگرسیون چندک به این معنی است که مشاهدات رگرسیونی (y, X) در موقعیت عمومی نباشند یعنی بیشتر از p باقیماندهٔ صفر داشته باشیم.

قضیه ۳: فرض کنید $N^۰, N^-, N^+$ بترتیب برابر تعداد مولفه‌های مثبت، منفی و صفر بردار

باقیمانده $y - X'\hat{\beta}(p)$ ، باشند اگر $a \in R^p$ وجود داشته باشد بطوریکه $Xa = I_n$ باشد، پس برای هر $\hat{\beta}(p)$ که جواب معادله:

$$\min_{\beta \in R^p} \sum_{i=1}^{n} \rho_p(y_i - x_i'\beta)$$

می‌باشد داریم: $N^+ \leq n(1-p) \leq N^+ + N^\circ$ و $N^- \leq np \leq N^- + N^\circ$

نتیجه: اگر $N^\circ = p$ باشد، در این صورت هیچ تباهیدگی وجود ندارد بنابراین نسبت باقیمانده‌های منفی تقریباً برابر p و نسبت باقیمانده‌های مثبت تقریباً برابر $(1-p)$ است یعنی:

$$\frac{N^+}{n} \leq (1-p) \leq \frac{N^+ + p}{n} \quad , \quad \frac{N^-}{n} \leq p \leq \frac{N^- + p}{n}$$

قضیه ۴: فرض کنید $A_p \times p$، یک ماتریس ناویژه در $\gamma \in R^p$ و $a > 0$ باشد اگر $\beta(p) \in B(p)$ باشد آنگاه جوابهای $R(\beta) = \sum_{i=1}^{n} \rho_p(y_i - x_i'\beta)$ به ازای هر $p \in (0, 1)$ دارای خواص زیر هستند:

$$\begin{aligned}
&(i) \quad \hat{\beta}(p\,; ay, X) = a\hat{\beta}(p\,; y, X) \\
&(ii) \quad \hat{\beta}(p\,; -ay, X) = a\hat{\beta}(1-p\,; y, X) \\
&(iii) \quad \hat{\beta}(p\,; y + X\gamma, X) = \hat{\beta}(p\,; y, X) + \gamma \\
&(iv) \quad \hat{\beta}(p\,; y, XA) = A^{-1}\hat{\beta}(p\,; y, X)
\end{aligned}$$

خواص (i) و (ii) پایایی مقیاس و (iii) پایایی انتقال یا پایایی رگرسیونی و (iv) پایایی دوباره پارامتری کردن طرح نامیده می‌شود.

کانکر و باست برای مدل رگرسیون خطی $y_i = x_i'\beta + u_i \qquad i = 1, 2, \ldots, n$ جواب:

$$\hat{\beta}_n(p) = \min_{\beta \in R^p} \sum_{i=1}^{n} \rho_p(y_i - x_i'\beta)$$

را بعنوان برآورد چندک رگرسیونی pام معرفی کردند. این چندکهای رگرسیونی p بعدی یک دنباله از ابر صفحه‌ها را تعیین می‌کنند که توابع چندک شرطی متغیر پاسخ y را برآورد می‌کنند همانند مدل یک نمونه‌ای، زمانی که p بین $(0, 1)$ تغییر می‌کند، یک برنامه‌ریزی خطی پارامتری داریم که می‌تواند به سادگی حل شود.

۷ مثال ۲

انگل در سال ۱۸۵۷، رابطه بین هزینه غذا و درآمد خانوار را مورد تجزیه و تحلیل قرار داد. او یک رگرسیون ساده از داده‌ها را که مربوط به ۲۳۵ خانوار اروپایی از رده کارگران می‌باشد، در

نظر گرفت. در اینجا ساده بدین مفهوم است که بردار متغیر تصادفی، تابعی یک بعدی است و داده‌ها را می‌توان به وسیله نمودار پراکندگی زوجهای (x_i, y_i) نشان داد که y_i میزان هزینه غذا و x_i میزان درآمد خانوار را نشان می‌دهد. در شکل (۵)، نمودار ۸ خط رگرسیون برازش داده شده نشان داده شده است که ۷ خط مربوط به خطوط چندکهای رگرسیونی متناظر با $p \in \{۰/۰۵, ۰/۱, ۰/۲۵, ۰/۵, ۰/۷۵, ۰/۹, ۰/۹۵\}$ و یک خط هم مربوط به برآورد خط حداقل مربعات تابع میانگین شرطی می‌باشد که به صورت خط چین رسم شده است. برازش میانه در $p = ۰/۵$، توسط خط ممتد تیره‌تر نشان داده شده است. نمودار به روشنی معلوم می‌کند که هزینه غذا با افزایش درآمد خانوار افزایش پیدا می‌کند فضای خطوط رگرسیون چندک، همچنین معلوم می‌کند که توزیع شرطی هزینه غذا به طرف چپ متمایل شده است. فضای کم چندکهای بالاتر (مرتبه بالاتر) نشان می‌دهد که فشردگی داده‌ها زیاد است، یعنی در این فضا داده‌ها بیشترند و فضای زیاد بین چندکهای پایین‌تر (مرتبه پایین‌تر) نشان می‌دهد که در این فضا فشردگی داده‌ها کمتر است. یعنی در این قسمت داده‌ها کمتر هستند.

برازشهای میانگین و میانه شرطی در این مثال کاملاً متفاوتند و برازش خط حداقل مربعات یک برآورد نسبتاً ضعیفی از میانگین شرطی برای فقیرترین خانواده‌ها در نمونه می‌باشد چون از بالای اکثر مشاهدات با درآمد خیلی پایین می‌گذرد.

۸ بحث

وقتی که خطاها دارای توزیع نرمال هستند برآوردهای حداقل مربعات کاراترین برآوردهای نااریب برای ضرایب رگرسیون هستند اما وقتی توزیع خطاها دمهای طولانی داشته باشد این برآوردها نمی توانند خیلی کارا باشند تحت این شرایط ما انتظار داریم که در داده‌ها نقاط دور افتاده وجود دارد. در حالتی که داده‌ها دارای نقاط دور افتاده باشند برازش حداقل مربعات روش خوبی نیست. به همین دلیل است که روش‌های برازش دیگری ارائه گردیده است که نسبت به نقاط دور افتاده حساس نیستند.

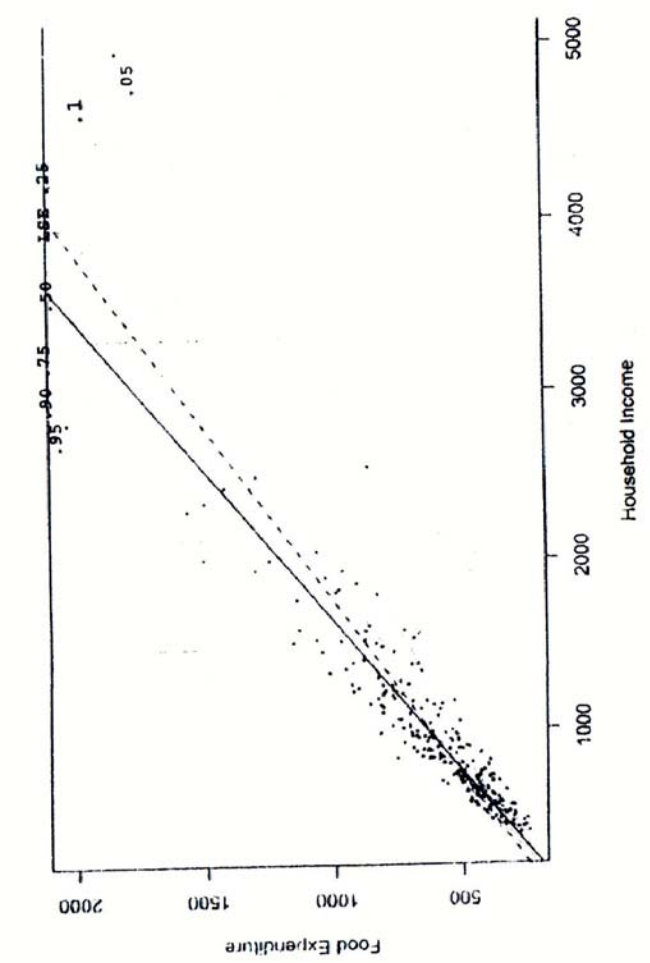
حساسیت کمترین مربعات به نقاط دور افتاده وابسته به دو عامل است.

(۱) اگر ما از معیار مربعات باقیمانده برای اندازه‌گیری استفاده کنیم هر باقیمانده بزرگ در مقایسه با سایر باقیمانده‌ها اندازه خیلی بزرگ خواهد داشت.

(۲) با استفاده از یک معیار مکانی مانند میانگین که نیرومند نیست هر مربع خطای بزرگ تأثیر خیلی قوی روی معیار، خواهد داشت در نتیجه در داده‌های انتهایی تأثیر نامناسبی بر برازش خواهد داشت.

دو راه ساده برای حل این مسئله وجود دارد.

(۱) ما می‌توانیم به جای معیار اندازه $e^۲$ از تابع دیگری از e یعنی $\rho(e)$ استفاده کنیم که اندازه باقیمانده‌ها را با کرانگین کمتری انعکاس دهد برای اینکه این تابع اندازه محسوس باشد تابع ρ باید: (الف) متقارن باشد. $\rho(e) = \rho(-e)$ (ب) مثبت باشد. $\rho(e) \geq ۰$ (ج) یکنوا باشد. یعنی



شکل ۵: برازش حداقل مربعات میانگین و چندکهای رگرسیونی داده‌های انگل

$\rho(|e_1| \geq \rho|e_2|) \quad if \quad |e_1| \geq |e_2|$ این نظریه ما را به سمت برآورد M (برآوردی که تابعی از $\rho(e)$ به کار می‌برد که ترکیبی از e^2 و $|e|$ می‌باشد.) می‌برد.

(۲) می‌توانیم به جای مجموع (یا بطور معادل میانگین) از معیار مکانی نیرومندتری از قبیل میانه (حالت خاص رگرسیون چندک) یا میانگین پیراسته $(trimmedmean)$ استفاده می‌کنیم. اساس روشهای رگرسیونی بر گرفته از این ایده شامل حداقل میانه مربعات و حداقل میانگین پیراسته مربعات می‌شود.

مهمترین انگیزه برای پیدایش رگرسیون چندک نیرومندی ذاتی آن در مقابل مشاهدات دورافتاده در متغیر پاسخ بوده است. در حالیکه رگرسیون معمولی به یک نقطه دور افتاده حساسیت دارد تأثیر مشاهدات دور افتاده بر $\hat{\beta}(p)$ محدود است در واقع دور شدن مشاهدات از چندک رگرسیونی برازش داده شده (مرتبهٔ yام) هیچ تأثیری روی برازش مدل ندارد (که در مثال ذکر شده نمایان است).

مراجع

[1] Georgea, F. Seber and Alanj, Lee (2003). Linear Regression Analysis Sonc. Inc. Allrights reserved Auckland , New Zealand

[2] Gilbert Basset, JR. and Roger Koenker, June (1982) An Emprical Quantile Function for Linear Models With iid Error Journal of the American Statistical Association ,Volume 77 ,Nvmber 378 Theory and Methods Section

[3] Hogg, R.V (1975) Estimates of percentile regression lines using salary data journal of the American statistical Association 70. 56-59

[4] Koenker, R. and Bassett, G.S., (1978). Regression Quantiles. Econometrica **46**, 33-50.

[5] Roger Koenker June 23, (2002). Quantile Regression Another Introduction Research was Partially supported by NSF grant SBR-9911184.

[6] Roger Koenker October 25, (2000). Quantile regression International Encyclopedia of the social sciences

[7] Roger Koenker November 10, (2000).Quantile Regression Department of Economhcs University of Illinois Urbana-champaign champaign61820 USA

میانگین‌گیری بیزی مدلهای رگرسیونی

افشین فلاح، محسن محمدزاده

گروه آمار دانشگاه تربیت مدرس

چکیده: در روشهای مدلسازی که بسته به معیار انتخاب مدل و نظر تحلیلگر مدلهای متفاوت انتخاب می‌شوند، عدم حتمیت در فرآیند مدلسازی نادیده گرفته می‌شود. میانگین‌گیری بیزی مدلها شیوه‌ای توانمند در مدلسازی است، که در آن به هر مدل به تناسب میزان حمایتی که از جانب داده‌ها صورت می‌پذیرد، وزنی اختصاص داده می‌شود و میانگین وزنی همهٔ مدلها بعنوان مدل نهایی بکار گرفته می‌شود. در این مقاله نشان داده می‌شود که روش میانگین‌گیری بیزی، خطای مدلسازی را کاهش و کارایی را افزایش می‌دهد.

واژه‌های کلیدی: عدم حتمیت، توزیع پیش‌بین، میانگین‌گیری بیزی

۱ مقدمه

معمولاً در مدلسازی، کلاسی از مدلها در نظرگرفته می‌شود، سپس یکی از آنها بر اساس یک یا چند معیار ارزیابی بعنوان بهترین مدل انتخاب می‌شود. در حالی که ممکن است رقبای بسیار خوبی برای مدل انتخابی در فضای مدل وجود داشته باشند. این شیوهٔ مدلسازی دارای نارسایی‌هایی است که مهمترین آنها در نظر نگرفتن عدم حتمیّت مدلهای انتخابی در فرآیند مدلسازی است. میانگین‌گیری بیزی مدلها[1] (BMA)، شیوه‌ای است که در آن از تمام مدلهای موجود در فضای مدل برای دستیابی به مدلی مناسب استفاده می‌شود. در این روش بر اساس میزان حمایت داده‌ها از هر مدل، وزنی به آن اختصاص داده می‌شود. سپس مدل حاصل از میانگین وزنی همهٔ مدلها برای انجام استنباط و پیش‌بینی به کارگرفته می‌شود. ریشهٔ تاریخی این روش به مقالهٔ لیمر (۱۹۷۸) برمی‌گردد، که به دلیل محاسبات دشوار و پیچیده در آن زمان چندان مورد توجه قرار نگرفت. با پیشرفت رایانه‌ها و فنون محاسبات تقریبی در اوایل دههٔ نود این روش دوباره مطرح و درکانون توجهات قرارگرفت. مادیگان و رفتری (۱۹۹۴) و مادیگان و یورک (۱۹۹۵) دو روش پایه‌ای برای اجرای این شیوهٔ مدلسازی ارائه کردند. هوئتینگ و همکاران (۱۹۹۴، ۱۹۹۶، ۱۹۹۹) به نحوهٔ انجام محاسبات و اجرای این روش پرداختند. روبرت (۲۰۰۰) و لیپکوویچ (۲۰۰۲) میانگین‌گیری بیزی را در حالت چند متغیره مورد توجه قرار دادند. در این مقاله روش میانگین‌گیری بیزی مدلها، نحوهٔ محاسبهٔ مؤلفه‌های و روشهای اجرای آن در بخش ۲ شرح داده

1) Bayesian Model Averaging

می‌شود و سپس در بخش ۳ به ارزیابی روش BMA پرداخته و نهایتاً بحث و نتیجه‌گیری در بخش ۴ ارائه خواهد شد.

۲ میانگین‌گیری بیزی مدلها

فرض کنید برای متغیر وابستهٔ Y و مجموعه‌ای از متغیرهای پیش‌بین $X_1, \ldots, X_k$، هدف یافتن بهترین مدل از بین همهٔ مدلهای خطی

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon \qquad (1)$$

باشد، که در آن $X_{i1}, \ldots, X_{ip}$ زیرمجموعه‌ای از متغیرهای $X_1, \ldots, X_k$ هستند. بسته به حضور یا عدم حضور هر متغیر پیش‌بین در مدل، 2^k مدل رگرسیونی وجود دارد، که فرض می‌شود همگی در فضای مدل $\mathcal{M}$ قرار دارند. اگر $\mathcal{M} = \{M_1, \ldots M_T\}$ نشان دهندهٔ فضای مدل و Δ کمیّتی مورد علاقه، مانند مشاهده‌ای در آینده باشد. در اینصورت استنباط بیزی بر اساس توزیع پسین Δ صورت می‌پذیرد، که با فرض مشاهدهٔ مجموعهٔ دادهٔ D، بنا بر قاعدهٔ احتمال کل، آمیخته‌ای از احتمالهای پسین همهٔ مدلها بصورت

$$Pr(\Delta|D) = \sum_{k=1}^{T} Pr(\Delta|M_k, D).Pr(M_k|D) \qquad (2)$$

است. میانگین و واریانس پسین Δ به ترتیب بصورت

$$\begin{aligned} E[\Delta|D] &= \sum_{k=1}^{T} E[\Delta|M_k, D].Pr(M_k|D) \\ &= \sum_{k=1}^{T} \hat{\Delta}_k.Pr(M_k|D) \qquad (3) \end{aligned}$$

و

$$\begin{aligned} Var[\Delta|D] &= E_M[Var(\Delta|D, M)] + Var_M[E(\Delta|D, M)] \\ &= \sum_{k=1}^{T} (Var[\Delta|D, M_k] + \hat{\Delta}_k^2) \times Pr(M_k|D) - E[\Delta|D]^2 \qquad (4) \end{aligned}$$

هستند (رفتری، ۱۹۹۳)، که در آن مؤلفهٔ $Var_M[E(\Delta|D, M)]$ عدم حتمیّت بین مدلها را نشان می‌دهد. عبارت (۲) یک میانگین وزنی احتمال پسین است، که در آن هر مدل با احتمال

پسین متناظر خود، یعنی $P(M_k|D)$، وزن‌دار شده است و توزیع پیش‌بین Δ به شرط مدل M_k بصورت

$$Pr(\Delta|M_k, D) = \int Pr(\Delta|\theta_k, M_k, D).Pr(\theta_k)\ d\theta_k \qquad (۵)$$

است، که در آن θ_k بردار پارامترهای مدل M_k را نشان می‌دهد. احتمال پسین مدل M_k با استفاده از قاعدهٔ بیز بصورت

$$Pr(M_k|D) = \frac{Pr(D|M_k).Pr(M_k)}{\sum_{j=1}^{T} P(D|M_j).Pr(M_j)} \qquad (۶)$$

بدست می‌آید، که در آن $Pr(M_k)$ احتمال پیشین درست بودن مدلM_k و

$$Pr(D|M_k) = \int Pr(D|\theta_k, M_k).Pr(\theta_k|M_k)\ d\theta_k \qquad (۷)$$

درستنمایی جمع بستهٔ[۲] مدلM_k و $Pr(D|\theta_k, M_k)$ درستنمایی مدل M_k است. برای محاسبهٔ $Pr(\Delta|D)$ باید مؤلفه‌های تشکیل دهندهٔ آن را مشخص و جایگزین نمود. بنابراین لازم است توزیع پیشین پارامترها، احتمالهای پیشین و پسین هر مدل و توزیع پیش‌بین کمیّت مورد علاقه را برای همهٔ مدلهای موجود در فضای مدل، مشخص کرد.

الف ــ تعیین احتمال پیشین پارامترهای مدل: یکی از مسائل دشوار در BMA تخصیص پیشین به پارامترهای مدل است. استفاده از پیشین‌های نامناسب[۳]، منجر به توزیع‌های پسین نامناسب می‌شود، که در اینصورت نمی‌توان این احتمالها را به عنوان احتمال مدل و نسبت آنها را به عنوان بیز فاکتور تعبیر نمود. به همین دلیل بسیاری از محققان گونه‌های مختلفی از پیشین‌های آگاهی بخش[۴] را پیشنهاد کرده‌اند. هوئتینگ (۱۹۹۴) استفاده از پیشین‌های مناسبی[۵] را پیشنهاد کرده است که در قسمتهایی از فضای پارامتر با درستنمایی بزرگ هموار باشند. مدل(۱) را می‌توان بصورت

$$Y = X\beta + \varepsilon \qquad (۸)$$

نوشت، که در آن $X_{n\times(p+1)}$ ماتریس مشاهدات، Y بردار n بعدی متغیرهای وابسته و ε بردار خطا است. فرض می‌شود εها برای مشاهدات مختلف، مستقل و دارای توزیع نرمال با میانگین صفر و واریانس σ^2 هستند. بردار $\beta = (\beta_0, \ldots, \beta_p)$ و پارامتر σ^2 نامعلوم هستند. توزیعهای پیشین این پارامترها بایستی بگونه‌ای تعیین شوند که عدم حتمیّت آنها را به خوبی منعکس سازند. هوئتینگ (۱۹۹۴) ردهٔ پیشین‌های مزدوج نرمال‌ـگاما را بصورت

$$\beta \sim N_{p+1}(\mu, \sigma^2 V) \quad , \quad \frac{\nu.\lambda}{\sigma^2} \sim \chi^2_\nu \qquad (۹)$$

2) Integrated Likelihood 3) Improper Prior 4) Informative 5) Proper Prior

در نظرگرفت، که در آنها ν، λ، ماتریس $V_{(p+1)\times(p+1)}$ و بردار $p+1$ بعدی μ، ابر پارامترهایی هستند که باید برآورد شوند.

ب ــ تعیین احتمال پیشین هر مدل: وقتی هیچ اطلاع پیشینی در مورد مدلها وجود ندارد یا میزان اطلاعات پیشین در مورد مدلها اندک است، معمولاً فرض می‌شود توزیع مدلها یکنواخت است. یعنی $Pr(M_j) = \frac{1}{T}$، در اینصورت احتمال پسین مدل M_k بصورت

$$Pr(M_k|D) = \frac{Pr(D|M_k)}{\sum_{j=1}^{T} Pr(D|M_j)} \qquad (10)$$

خواهد بود. اگر بعضی از مدلها در مقایسه با سایرین دارای احتمال بیشتری باشند یا اطلاعات خوبی در مورد آنها در دسترس باشد، لازم است این اطلاعات برای تعدیل احتمالهای پیشین مدلها بکار گرفته شوند تا از پیشین‌های آگاهی بخش‌تر استفاده شود. در مسائل مربوط به انتخاب متغیرهای پیش‌بین، اطلاعات پیشین بشکل شواهد قبلی برای در نظرگرفتن یک متغیر مورد استفاده قرار می‌گیرند. فرض کنید تنها این نوع از اطلاعات پیشین در اختیار باشد و مدل M_k توسط بردار $(\delta_{k1}, \ldots, \delta_{kp})$ مشخص شود، که در آن

$$\delta_{ki} = \begin{cases} 1 & X_i \in M_i \\ 0 & X_i \notin M_i \end{cases} \qquad i = 1, \ldots, p$$

توابع نشانگر هستند. حال اگر π_i احتمال مؤثر بودن متغیر X_i بر Y را نشان دهد و بپذیریم که اطلاعات پیشین در مورد متغیرهای متفاوت تقریباً مستقل هستند، می‌توان

$$Pr(M_k) = \prod_{i=1}^{p} [\pi_i^{\delta_{ki}} \times (1-\pi_i)^{1-\delta_{ki}}] \qquad (11)$$

را به عنوان احتمال پیشین صحیح بودن مدل M_k در نظرگرفت. چون این توزیع به متغیری که مهمتر است احتمال بزرگتری تخصیص می‌دهد، توسط مادیگان و رفتری (۱۹۹۱) توزیع پیشین متغیر[6] نامیده شده است.

ج ــ تعیین احتمال پسین هر مدل: برای محاسبه احتمال پسین هر مدل لازم است $Pr(D|M_k)$ از رابطهٔ (۷) که یک انتگرال با بعد برابر با تعداد پارامترهای مدل M_k است، محاسبه شود. بنابراین محاسبهٔ دقیق این احتمال به دلیل پیچیده بودن انتگرال مربوطه تنها در حالات بسیار خاص و ساده امکان پذیر است و در سایر موارد از روشهای تقریبی و محاسباتی استفاده می‌شود. در اینجا از معیار اطلاع بیز[7] (BIC) برای این منظور بهره می‌بریم. معیار BIC برای رگرسیون خطی بصورت

$$BIC_j = n \log(1 - R_j^2) + k_j \log n \qquad (12)$$

6) Variable Prior 7) Bayesian Information Criteria

تعریف می‌شود، که در آن n تعداد مشاهدات، R_j^2 ضریب تعیین تعدیل شده مدل و k_j تعداد متغیرهای پیش‌بین موجود در مدل jام و را نشان می‌دهد. در اینصورت درستنمایی جمع بستهٔ مدل jام را می‌توان بصورت تقریبی

$$Pr(D|M_j) \propto e^{-0/5BIC_j} = e^{-0/5(n\log(1-r_j^2)+k_j\log n)} \qquad (۱۳)$$

خواهد بود. بر این اساس احتمال پسین مدل kام با استفاده از قاعدهٔ بیز بصورت

$$Pr(M_k|D) \quad = \quad \frac{\exp\{-0/5BIC_k\}\ .\ Pr(M_k)}{\sum_{j=1}^{T}\exp\{-0/5BIC_j\}\ .\ Pr(M_j)} \qquad (۱۴)$$

نوشت. چون معیار BIC برای بسیاری از مدلها دارای شکل بسته و ساده‌ای است، استفاده از این تقریب موجب سهولت محاسبه و افزایش سرعت می‌شود.

د ــ تعیین توزیع پیش‌بین: انتگرال تشکیل دهندهٔ توزیع پیش‌بین (۵) از دو جزء تشکیل شده است. معمولاً برای جزء دوم آن از تقریب

$$Pr(\Delta|M_k, D) \approx Pr(\Delta|M_k, \hat{\theta}_k, D) \qquad (۱۵)$$

استفاده می شود (رفتری و همکاران، ۱۹۹۴)، که در آن $\hat{\theta}_k$ برآورد حداکثر درستنمایی بردار پارامترهای مدل M_k است.

اکنون با فرض مشخص بودن همهٔ مولفه‌های لازم برای اجرای روش BMA محاسبهٔ مجموع (۲) بواسطه تعداد زیاد جملات آن عملاً امکان‌پذیر نیست و لازم است زیر مجموعه‌ای از محتمل‌ترین مدلها انتخاب شود. مادیگان و رفتری (۱۹۹۴) روش پنجره اوکام[8] را برای این منظور پیشنهاد کرده‌اند، که از دو اصل کلی پیروی می‌کند. بنابر اصل اول مدلهایی که در مقایسه با محتمل‌ترین مدل خیلی کم شانس هستند، کنارگذاشته می‌شوند. بنابر اصل دوم که تیغ اوکام[9] نامیده می‌شود، مدلهایی که نسبت به زیرمدلهای ساده‌تر خود کمتر از جانب داده‌ها حمایت می‌شوند، کنارگذاشته می‌شوند. با اجرای اصل اول مدلهایی که در مجموعهٔ

$$\mathcal{A}' = \left\{ M_k : \frac{Max_{M_k\in\mathcal{M}}\{Pr(M_l|D)\}}{Pr(M_k|D)} > C_1 \right\}$$

قرار دارند و همچنین بنابر اصل تیغ اوکام، مدلهایی که در مجموعه

$$\mathcal{B} = \left\{ M_k : \exists M_l \in \mathcal{A}\ s.t\ M_l \subset M_k, \frac{Pr(M_l|D)}{Pr(M_k|D)} > C_2 \right\}$$

8) Occam's Window 9) (OW)Occam's Razor

باشند، از مجموع (۲) خارج می‌شوند، که در آن C_1 و C_2 توسط تحلیل‌گر تعیین می‌شوند. در اینصورت مجموع (۲) را می‌توان بصورت

$$Pr(\Delta|D) = \frac{\sum_{M_k \in \mathcal{A}} Pr(\Delta|M_k, D).Pr(D|M_k).Pr(M_k)}{\sum_{M_k \in A} Pr(D|M_k).Pr(M_k)} \quad (۱۶)$$

بازنویسی کرد، که در آن $\mathcal{A} = \mathcal{A}' - \mathcal{B} \in \mathcal{M}$ مجموعهٔ پذیرش است.

۳ مثال کاربردی

در این بخش کارایی روش BMA در تشخیص متغیرهای پیش‌بین مؤثر در مدل و برآورد پارامترهای مدل با روش مدلسازی گام به گام بعنوان یکی از روشهای مدلسازی مرسوم مورد مقایسه قرار می‌گیرد. با توجه به اینکه بطور طبیعی کارایی روش BMA با افزایش تعداد متغیرهای پیش‌بین، افزایش می‌یابد، برای نمایش بهتر کارایی روش BMA از مثالی با تعداد متغیرهای پیش‌بین کم استفاده شده تا حتی‌المکان از بالا بودن کارایی این روش بدلیل تعداد زیاد متغیر پیش‌بین اجتناب شود.

جدول ۱: داده‌های مربوط به کارایی شغلی پرستاران

Y	۴۵	۶۵	۷۳	۶۳	۸۳	۴۵	۶۰	۷۳	۷۴	۶۹	۶۶	۶۹	۷۱	۷۰	۷۹
X_1	۷۴	۶۵	۷۱	۶۲	۷۹	۵۶	۶۸	۷۶	۸۳	۶۲	۵۲	۶۱	۶۳	۸۴	۷۸
X_2	۲۹	۵۰	۶۷	۴۴	۵۵	۴۸	۴۱	۴۹	۷۱	۴۴	۵۲	۴۶	۵۶	۸۲	۵۳
X_3	۴۰	۶۴	۷۹	۵۷	۷۶	۵۴	۶۶	۶۵	۷۷	۵۷	۶۷	۶۶	۶۷	۶۸	۸۲
X_4	۶۶	۶۸	۸۱	۵۹	۷۶	۵۹	۷۱	۷۵	۷۶	۶۷	۶۳	۶۴	۶۰	۶۴	۶۴
X_5	۹۳	۷۴	۸۷	۸۵	۸۴	۵۰	۶۹	۶۷	۸۴	۸۱	۶۸	۷۵	۶۴	۷۸	۷۸
X_6	۴۷	۴۹	۳۳	۳۷	۳۳	۴۲	۳۷	۴۳	۳۳	۴۳	۳۶	۴۳	۳۵	۳۷	۳۹
Y	۸۳	۷۵	۶۷	۶۷	۵۲	۵۲	۶۶	۵۵	۴۲	۶۵	۶۸	۸۰	۵۰	۸۷	۸۴
X_1	۶۵	۸۶	۶۱	۷۱	۵۹	۷۱	۶۲	۶۷	۶۵	۵۵	۷۸	۷۶	۵۸	۸۶	۸۳
X_2	۴۹	۶۳	۶۴	۴۵	۶۷	۳۲	۵۱	۵۱	۴۱	۴۱	۶۵	۵۷	۴۳	۷۰	۳۸
X_3	۸۲	۷۹	۷۵	۶۷	۶۴	۴۴	۷۲	۶۰	۴۵	۵۸	۷۳	۸۴	۵۵	۸۱	۸۳
X_4	۸۴	۶۵	۸۴	۶۰	۸۰	۶۹	۴۸	۷۱	۶۸	۵۵	۷۱	۹۳	۸۵	۵۶	۸۲
X_5	۵۵	۸۰	۸۱	۸۶	۷۹	۶۵	۸۱	۸۱	۵۸	۷۶	۷۷	۷۹	۸۴	۷۵	۷۹
X_6	۳۸	۴۱	۴۵	۴۸	۵۴	۴۳	۴۳	۳۹	۵۱	۳۵	۴۲	۳۵	۴۰	۳۰	۴۱

جدول ۱ مشاهدات یک مطالعهٔ رگرسیونی را نشان می‌دهد، که هدف آن بررسی تأثیر ۶ متغیر مستقل شامل قاطعیت (X_1)، علاقه‌مندی (X_2)، بلند همتی (X_3)، روابط اجتماعی (X_4)، هنرمندی درحل مشکلات (X_5) و ابتکار (X_6) بر کارایی شغلی پرستاران است. بدلیل وجود ۶ متغیر پیش‌بین، فضای مدل شامل $۶۴ = ۲^۶$ مدل است. روش پنجرهٔ اوکام ۱۸ مدل را بر اساس مقایسه احتمالهای پسین مدلها به عنوان برترین مدلهای موجود مورد استفاده قرار می‌دهد. مدلهای حاصل به همراه احتمال پسین، مقدار R^2_{adj} به درصد و معیار اطلاع بیز آنها در جدول ۲ ارائه شده‌اند. تعداد زیاد مدلهایی که در پنجرهٔ اوکام قرار گرفته‌اند (بیش از $\frac{1}{4}$ مدلها)، نشان

جدول ۲: مدلهای انتخابی توسط روش پنجرهٔ اوکام

شماره مدل	مدل	احتمال پسین	R^2_{adj}	BIC
۱	X_3 X_6	۰٫۱۹۷۶	۸۲٫۲۰۴	-۴۱٫۷۵۹۷
۲	X_2 X_3 X_6	۰٫۱۴۷۳	۸۲٫۹۲۲	-۴۱٫۱۵۲۱
۳	X_1 X_2 X_3 X_6	۰٫۱۱۷۵	۸۴٫۵۵۵	-۴۰٫۶۰۰۵
۴	X_1 X_3 X_6	۰٫۰۹۸۹	۸۲٫۴۴۷	-۴۰٫۳۵۶۵
۵	X_3	۰٫۰۶۸۸	۷۷٫۲۹۶	-۳۹٫۶۲۸۹
۶	X_3 X_5 X_6	۰٫۰۴۸۱	۸۱٫۵۵۲	-۳۸٫۹۱۴۳
۷	X_2 X_3 X_5 X_6	۰٫۰۴۲۱	۸۳٫۴۲۳	-۳۸٫۶۴۸۳
۸	X_3 X_4 X_6	۰٫۰۳۸۳	۸۱٫۲۶۱	-۳۸٫۴۶۰۴
۹	X_2 X_3	۰٫۰۳۵۰	۷۸٫۸۲۰	-۳۸٫۲۷۶۶
۱۰	X_1 X_3	۰٫۰۳۴۹	۷۸٫۸۱۶	-۳۸٫۲۷۱۲
۱۱	X_2 X_3 X_4 X_6	۰٫۰۳۰۳	۸۳٫۰۴۱	-۳۷٫۹۸۷۶
۱۲	X_1 X_2 X_3	۰٫۰۲۶۱	۸۰٫۷۵۴	-۳۷٫۶۸۶۲
۱۳	X_1 X_2 X_3 X_5 X_6	۰٫۰۲۵۰	۸۷٫۷۰۰	-۳۷٫۶۰۵۷
۱۴	X_1 X_2 X_3 X_4 X_6	۰٫۰۲۲۳	۸۴٫۵۸۲	-۳۷٫۳۸۲۹
۱۵	X_1 X_3 X_5 X_6	۰٫۰۱۹۷	۸۲٫۵۳۴	-۳۷٫۱۳۳۳
۱۶	X_1 X_3 X_4 X_6	۰٫۰۱۹۱	۸۲٫۴۹۵	-۳۷٫۰۶۸۶
۱۷	X_3 X_5	۰٫۰۱۶۲	۷۷٫۶۶۳	-۳۶٫۷۳۴۳
۱۸	X_3 X_4	۰٫۰۱۲۸	۷۷٫۳۰۴	-۳۶٫۲۷۱۸

دهندهٔ زیاد بودن عدم حتمیّت است، یعنی بیش از یک مدل صحیح وجود دارد، بنابراین انتخاب یک مدل به عنوان مدل نهایی، منطقی بنظر نمی‌رسد. ملاحظه می‌شود که مقادیر R^2_{adj} و معیار اطلاع بیز BIC، برای مدلهایی که در پنجره اوکام قرار گرفته‌اند، نزدیک به یکدیگر هستند. بیشترین مقدار R^2_{adj} مربوط به مدل شماره ۱۳ است، در حالی که بر اساس مقدار کمینهٔ معیار اطلاع بیز، مدل ۱ بهترین است. انتخاب یک مدل و مبنا قرار دادن آن به معنی آن است که احتمال صحیح بودن آن یک می‌باشد، در حالی که بیشترین احتمال پسین مربوط به مدل شمارهٔ ۱ و حدود ۰٫۲ است و احتمال پسین سایر مدلها همگی از ۰٫۲ کمتر است. جدول ۲ نشان می‌دهد که متغیر X_3 در همهٔ مدلها و متغیر X_6 در ۱۲ مدل منتخب روش پنجرهٔ اوکام حضور دارند، که نشان دهندهٔ تأثیرگذار بودن این دو متغیر است. برآورد ضرایب متغیرهای مستقل که در جدول ۳ ارائه شده نیز اهمیت دو متغیر X_3 و X_6 را تأئید می‌کند. همانطور که در جدول ۳ ملاحظه می‌شود، برآوردهای بیز ضرایب این دو متغیر که با استفاده از روش BMA بدست آمده‌اند، به خوبی اهمیت این دو متغیر را منعکس می‌سازند. جدول ۴ برآورد پارامترها و خطای معیار آنها را که به روش گام به گام بدست آمده‌اند را نشان می دهد. همانطور که ملاحظه می شود، خطای معیار متناظر با برآوردگرهای این دو پارامتر در روش BMA بصورتی قابل ملاحظه از مقادیر مشابه در روش گام به گام بزرگتر هستند. این یکی از برتریهای روش BMA است که میزان واقعی عدم حتمیت را نشان می‌دهد. مطلب دیگری که می‌توان به آن اشاره کرد این است

جدول ۳: برآورد ضرایب در روش BMA

پارامتر رگرسیونی	$Pr[\beta_i \neq 0 \mid D]$	میانگین پسین	انحراف معیار پسین
β_1	۰/۳۶۴	۰/۰۶۲۸	۰/۱۱۰۶
β_2	۰/۴۴۶	-۰/۰۷۴۸	۰/۱۰۷۸
β_3	۱	۰/۸۰۷۶	۰/۱۲۵۸
β_4	۰/۱۲۳	۰/۰۰۱۹	۰/۰۴۹۷
β_5	۰/۱۵۱	۰/۰۱۰۳	۰/۰۴۸۲
β_6	۰/۸۰۶	۰/۳۷۹۱	۰/۲۵۸۱

جدول ۴: برآورد ضرایب در مدل کامل($R^2_{adj} = 0/8484$) و روش گام به گام

مدل کامل	برآورد پارامترها	خطای معیار	روش گام به گام	برآورد پارامترها	خطای معیار
β_0	۲۲/۹۱۳۵	۱۴/۱۶۶۶۵	*	۳۱/۹۶	۱۲/۶۴۸۵
β_1	۰/۲۳۳۳	۰/۱۲۸۳	---	---	---
β_2	-۰/۱۸۹۵	۰/۱۰۱۶	---	---	---
β_3	۰/۸۵۸۹	۰/۱۳۳۲	*	۰/۷۸۷۲	۰/۰۹۶۸
β_4	۰/۰۷۲۱	۰/۱۴۷۰	---	---	---
β_5	۰/۰۲۴۰	۰/۱۱۰۱	---	---	---
β_6	۰/۴۱۷۷	۰/۱۹۷۰	*	-۰/۴۴۶	۰/۲۰۲۷

که بخت حضور متغیر X_2 در مدل بیش از ۰/۸ است، که بر اساس قواعد سرانگشتی جفریز (۱۹۶۱) شواهد مثبتی برای حضور متغیر X_2 در مدل ارائه می‌کند، در حالی که اطلاعات موجود در این متغیر، با حذف آن در مدل حاصل از روش گام به گام نادیده گرفته می‌شوند. برآورد ضرایب متغیرهای در روش BMA در مقایسه با سایر روشها و خصوصاً در مورد متغیرهای X_1 و X_4 کاملاً کوچک و نزدیک صفر هستند. با این وجود کاملاً نادیده گرفته نمی‌شوند و از اطلاعات آنها استفاده می‌شود.

۴ بحث و نتیجه‌گیری

در هنگام مدلسازی غالباً نمی‌توان مدلی یافت که بصورت کامل به داده‌ها برازش داشته باشد. در این حالات انتخاب یک مدل به معنی از بین رفتن اطلاعات سایر مدلها است. میانگین‌گیری بیزی مدلها از اطلاعات همهٔ مدلها، یا دسته‌ای از بهترین مدلها استفاده می‌کند و برخلاف روشهای مرسوم عدم حتمیّت را به خوبی منعکس می‌کند. از نقطه نظر کارایی پیش‌بین نیز نتیجهٔ حاصل از روش BMA از نتایج حاصل از هر یک از مدلهای موجود در فضای مدل مطلوب‌تر است.

عملکرد این روش با افزایش عدم حتمیّت بهبود می‌یابد. مثلاً هر چه تعداد متغیرهای پیش‌بین در مدل رگرسیونی بیشتر باشد، کارایی این روش بیشتر می‌شود. در بدترین حالت که عدم حتمیت اندک باشد، بهتر بودن روش BMA نسبت به سایر روشها چشمگیر نمی‌باشد و بکارگیری روشهای مرسوم بدلیل سادگی مقرون به صرفه‌تر است.

مراجع

[1] Good, I. J., (1950). "Probability and The Weighing of Evidence", Griffin London.

[2] Hoeting, J., (1994). "Accounting for Uncertainty in Linear Regression Models", Ph.D Disertation, Department of Statistics, University of Washington.

[3] Hoeting, J., Raftery, A. and Madigan, D., (1996). "A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression Models". J. Comput, Statist. **22**, 251-271.

[4] Hoeting, J., Raftery, A. and Madigan, D., (1999). "Bayesian Simultaneous Variable Selection and Transformation Selection in Linear Regression Models". Tecnical Report 9905, Dept. Statistics, Colorado State Univ. Avalable at www. Colostate. edu.

[5] Lipkovich, I, A., (2002). "Bayesian Model Averaging and Variable Selection in Multivariate Ecological Models", Ph.D Disertation, Faculty of The Virginia Polytechnic Institue and State University, Blacksburg, Virginia.

[6] Madigan, D., Gavrin, J. and Raftery, A. E., (1995). "Elicting Prior Information To Enhace The Preformance of Bayesian Graphical Models", Comm. Statist. Thory Methods, **24**, 2271-2292.

[7] Madigan, D and Raftery, A., (1991). "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window", Technical Reports 213, Univ. Washington, Seattle.

[8] Madigan, D. and Raftery, A., (1994). "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window", Journal of American Statistical Association, **89**, 1535-1546.

[9] Miller A., (1990). "Subset Selection in Regression Variables", New York, Chpman-hal.

[10] Raftery, A. E., (1995). "Bayesian Model Selection in Social Research(With Discussion)", in Sociological Methodology 1995(P. V. Marsden, ed.) 111-195. Blakwell. Cambridge, MA.

[11] Raftery, A. E., (1996). "Approximate Bayes Factor and Accounting for Model Uncertainty in Generalized Linear Models", Technical Report, Biometrika, **83**, 351-266.

[12] Robert, B. Nobel, J. (2000). " Multivariate applications of Bayesian Model averaging", Ph.D disertation, Faculty of the virginia polytechnic institue and state university, Blacksburg, virginia.

نمونه‌های گردان[1]، جایگزینی برای سرشماری‌های جمعیتی

زهره فلاح محسن‌خانی، فاطمه هرندی، فرشید جمشیدی

گروه پژوهشی طرحهای فنی و روشهای آماری پژوهشکده آمار

چکیده: در آمارگیری‌های نمونه‌ای متداول، معمولا نمونه به یک جامعه مشخص و به یک مقطع زمانی خاص منتسب می‌شود. در سالهای اخیر طرح نمونه‌گیری احتمالی با عنوان طرح نمونه‌گیری گردان معرفی شده‌است که برآوردهای حاصل از نمونه‌های این طرح نتایجی فراتر از نتایج آمارگیری‌های نمونه‌ای متداول ارائه می‌کنند تا آنجا که انجام این نمونه‌گیری در طی زمان بعنوان جایگزینی برای سرشماری‌ها معرفی شده است.

در یک طرح نمونه‌گیری گردان، k نمونه احتمالی غیر متداخل همزمان انتخاب می‌شوند. اندازه نمونه‌ها، یکسان و برابر با کسر ثابتی از کل جامعه می‌باشد. از هر نمونه در یک دوره زمانی محدود مصاحبه به‌عمل می‌آید و ضمن آن که اطلاعات آن دوره برآورد می‌شود، با متوسط گیری از اطلاعات دوره‌های زمانی مختلف، برآوردهایی برای کل دوره‌ها نیز ارائه می‌شود. سئوالی که در این روش مطرح است این است که با وجود اختلاف بین دوره‌ها چطور می‌توان از نمونه‌های دوره‌های مختلف زمانی، متوسط گیری کرد و به نتایج حاصل اطمینان نمود؟

در این مقاله طرح نمونه‌گیری گردان بعنوان دیدگاهی جدید در تولید آمارهای رسمی بویژه بعنوان جایگزینی برای سرشماری‌های دهساله ارائه می‌شود.

واژه‌های کلیدی: نمونه‌گیری احتمالی، نمونه‌گیری ادواری، نمونه‌های گردان، نمونه احتمالی متداخل، نمونه احتمالی غیر متداخل

۱ مقدمه

نمونه‌های گردان به وسیلهٔ کیش[2] در یک سری از مقالات شرح داده شده است. (کیش ۱۹۷۹، ۱۹۸۱، ۱۹۸۳، ۱۹۸۷، ۱۹۹۰، ۱۹۹۷، ۱۹۹۸ و کیش و ورما[3] ۱۹۸۳ و ۱۹۸۶). کیش در مقاله‌های مذکور اصول تجمع [4] اطلاعات در فضا و زمان را با استفاده از داده‌های حاصل از نمونه‌گیری گردان تشریح کرده است وی معتقد به استفاده از این روش برای اهداف گوناگون (کیش ۱۹۹۹)، به‌خصوص درکشورهای در حال توسعه (کیش ۱۹۷۹) و نیز در سرشماری ایالات متحده (کیش ۱۹۸۱) است. استفاده او از مفهوم نمونه‌های گردان، حداقل به سال ۱۹۵۸ تحت عنوان «نمونه‌گیری مستمر» (کیش، لاوجوی[5] و راکوف[6]، ۱۹۶۱) باز می‌گردد. وی همچنین پیشنهاد

1) Rolling Sample 2) Kish 3) Verma 4) Cumulating 5) Lovejoy 6) Rackow

نموده است نمونه‌های مذکور برای سرشماری‌های نمونه‌ای ملی توسط اداره سرشماری آمریکا استفاده شوند. (الکساندر[7] ۲۰۰۰).

آمارگیری جامعه آمریکا[8] (ACS)، که به عنوان جایگزینی برای سرشماری درنظرگرفته شده است، از یک نمونه‌گیری گردان استفاده می‌کند. این مقاله چگونگی به‌کارگیری مفهوم نمونه‌های گردان را با استفاده از ملاحظات عملیاتی و اهداف خاص در ACS شرح می‌دهد.

۲ نمونه‌های گردان

در یک طرح «نمونه گردان» همزمان k نمونه احتمالی غیر متداخل (پانل) انتخاب می‌شود که هر یک شامل $1/F$ کل جامعه می‌باشد. هر پانل در یک دوره زمانی، مورد مصاحبه قرار می‌گیرد تا اینکه همه نمونه‌ها بعد از k دوره مصاحبه شوند. بر حسب دقت مورد نیاز، اطلاعات یک پانل $1/F$، می‌تواند برای به‌دست آوردن برآوردهای خوبی از جامعه و چند محدوده وسیع کافی باشد. برای محدوده‌های کوچکتر یا برای حصول دقت بیشتر در محدوده‌های بزرگ، از تجمع چند پانل متوالی تا میزان k/F جمعیت، استفاده می‌شود. یک طرح نمونه گردان که در آن $k = F$ باشد، «سرشمای گردان[9]» نامیده می‌شود. برای یک نمونه‌گردان با دوره‌های ماهانه، طبیعی است که F مضربی از دوازده و تجمع‌ها، فصلی، نیمه‌ساله، سالانه و چند ساله می‌باشد.

۳ آمارگیری جامعه آمریکا (ACS)

سرشماری ده‌سال یکبار همراه با نمونه‌گیری مسکن با نرخهای ۱/۶ تا ۱/۱۰، منبع اصلی جمع‌آوری داده‌های ایالتی در زمینه مشخصه‌های جمعیتی و مسکن ایالات متحده است. همچنین آمارگیری جاری جمعیت (CPS) که آمارگیری نیروی کار ماهانه ایالات متحده است، با نرخ نمونه‌گیری یک از هزار، ماهانه انجام می‌شود، اما طرح آمارگیری آن بگونه‌ای است که نمونه نمی‌تواند به شکل قابل استفاده، آن چنان که در نمونه‌گردان مشاهده می‌شود در زمان تجمع پیدا کند (به دلیل حجم قابل توجه تداخل نمونه بین دوره‌ها). بنابراین دولت فدرال، در فاصله بین دو سرشماری، اطلاعات نسبتاً اندکی راجع به مشخصه‌های جمعیتی در سطح نواحی کشور ارائه می‌نماید. بعد از بحث‌ها و نشست‌های مفّصل در فاصله زمانی سالهای ۱۹۹۱ تا ۲۰۰۲ بنا شد در صورت تأمین منابع مالی توسط کنگره، آمارگیری جامعه آمریکا (ACS) توسط دفتر سرشماری در سال ۲۰۰۳ با نمونه ماهانه‌ای حدود ۲۵۰۰۰۰ آدرس، آغاز شود و در صورت موفقیت، جایگزین سرشماری دهسال یکبار گردد. ACS همان موضوعات سرشماری را با فراهم کردن جزئیات بیشتر در زمینه اجتماعی، اقتصادی و ویژگی‌های مسکن از کل ایالات متحده پوشش می‌دهد.

7) Alexander 8) American Community Survey 9) Rolling Census

ACS یک آمارگیری مبتنی بر نمونه گردان است (کیش ۱۹۹۸) و مطابق با یک نمونه گردان ماهانه با نرخ متوسط تقریباً $F = ۴۸۰$ یا یک نمونه سالانه با $F = ۴۰$ است. آمارگیری از $k = ۶۰$ استفاده می‌کند که کوتاهترین تجمع قابل انتشار آن، برآوردهای سالانه است. نمونه‌گردان با مجموعه متفاوتی از آدرسها در هر ماه تا زمانی که در تمام جامعه گردش پیدا کند، تماس برقرار می‌کند. آدرسها یک نمونه سیستماتیک از آدرس شهروندان هستند که از فایلهای بهنگام شدهٔ آدرسهای اصلی استخراج می‌شوند که در تمام مناطق در هرماه تهیه می‌شود و در واقع بدون وجود آن اجرای طرح ACS امکان‌پذیر نمی‌باشد. نمونهٔ سالانه تقریباً شامل ۳ درصد جامعه می‌شود که در طی یک دورهٔ ۵ ساله به حدود ۱۵ درصد بالغ می‌گردد. ACS از طریق پست اجرا و بی‌پاسخی‌ها به‌وسیلهٔ تلفن پیگیری می‌شود. یک نمونه تصادفی شامل یک سوم بی‌پاسخی‌های باقیمانده، به‌صورت حضوری پیگیری می‌شود.

در محدوده‌هایی با نرخ پاسخگویی متوسط و F ماهانه برابر ۴۸۰، خطای معیار برای یک برآورد متوسط ۵ ساله حاصل ازACS، قدری بیشتر از برآورد متناظر حاصل از نمونه‌گیری همراه سرشماری خواهد بود (حدوداً ۱٫۳۳ برابر بزرگتر). با توجه به امتیاز به‌موقع بودن و اینکه انتظار می‌رود نرخ داده‌های گم شده بدلیل استفاده از آمارگیران دائمی پائین باشد، پیش‌بینی می‌شود که در اکثر موارد برآوردها به قدرکافی نزدیک به واقعیت باشند. در نواحی با نرخهای پاسخ پائین‌تر از متوسط نرخ پاسخ زیر نمونه‌گیری برای پیگیری بی‌پاسخیها، اندازه نمونه مؤثر را کاهش خواهد داد. این امر نه‌تنها به دلیل کاهش تعداد مصاحبه‌ها بلکه به این دلیل است که وزنهای نامساوی نوعاً باعث بالا بردن اثر طرح[10] می‌شود (کیش ۱۹۶۵). برای جبران این مسئله، ACS از یک نرخ نمونه‌گیری بالاتر برای بی‌پاسخی در محدوده‌هایی با نرخ پاسخ پایین‌تر از حد متوسط استفاده می‌کند. [چارلز اچ الکساندر[11] ۲۰۰۱]

۴ استفاده از متوسط‌ها در ACS

مسئله مهمی که در ارتباط با ACS مطرح می‌شود درک مسائلی است که استفاده از متوسط‌های چند ساله در موارد زیر ایجاد می‌کند:

۱) تفسیر و استفاده از متوسط‌های چند ساله برای توصیف و مقایسه نواحی
۲) استفاده از متوسط‌های چند ساله برای تعیین مقدار تخصیص بودجه
۳) استفاده از متوسط‌های چند ساله برای اندازه‌گیری تغییرات و روندها

توصیه شده که برای تولید آمارهای توصیفی پایه برای نواحی کوچک، کاربران داده‌ها را برای چند سال (بسته به جمعیت نواحی)، انباشته کنند. این توصیه برای بسیاری از کاربران ACS خصوصاً آن دسته ازکاربران که عادت به استفاده از داده‌های سرشماری دارند، مفید است. برای پیش‌بینی مقدار تخصیص بودجه نیز متوسط‌ها مفید هستند، گر چه بررسیهای اولیه، تأکید بیشتری

10) Design effect 11) Charles H. Alexander

بر استفاده از متوسط‌های سه سالانه، حتی برای نواحی کوچک، دارند. برای اندازه‌گیری تغییرات در طول زمان، کاربر باید در کل سری‌های زمانی برآوردهای سالانه را تحلیل کند.

۱.۴ مدل‌هایی برای استفاده‌های توصیفی اساسی از داده‌های ACS

سه مدل اساسی وجود دارند که تحت آنها متوسط مقادیر سال‌های گذشته قابل استفاده است و کاربر داده‌ها بایستی استفاده و تفسیر متوسط‌ها را با توجه به یکی از این مدلها انجام دهد. این مدلها عبارتند از:

۱ــ میانگین برای دوره خاص زمانی

۲ــ مدل «کاربرد نوعی سرشماری[12]»

۳ــ مدل «اغتشاش تصادفی[13]».

برای تشریح این سه مدل، هم از داده‌های سرشماری ۲۰۰۰ و هم از داده‌های ۲۰۰۲ـ۱۹۹۸ در سال ۲۰۰۳ استفاده می‌شود. فرض کنید X_t مقدار واقعی ناحیه‌ای خاص در سال t باشد، برای سال جاری، $t = ۲۰۰۳$ است.

به جای استفاده از مقدار واقعی، آنچه مشاهده می‌کنیم $\hat{X}_t = X_t + \epsilon_t$ است، که در آن ϵ_t خطای نمونه‌گیری است. در سرشماری، $\hat{X}_t$ تنها هر ۱۰ سال یک بار مشاهده می‌شود اما دارای واریانس نمونه‌گیری کوچکتری در مقایسه با برآورد سالانه ACS می‌باشد.

۱.۱.۴ مدل ۱: میانگین برای دوره‌ی خاص زمانی

در این مدل هدف برآورد میانگین زیر است:

$$(X_{۱۹۹۸} + \cdots + X_{۲۰۰۲})/۵$$

در این حالت، متوسط ACS برآوردگر آشکاری است و می‌تواند تحت یک رده از مدلهای سری زمانی $\{X_t\}$ توجیه گردد.

گرچه این کاربرد مهمی است، اما غیر معمول می‌باشد و کاربرد معمول، بررسی «وضعیت کنونی» است که در مدل ۲ منعکس شده است.

۲.۱.۴ مدل ۲: مدل «کاربرد متداول سرشماری »

به‌عنوان یک مثال درباره‌ی استفاده‌ی «متداول » از داده‌های سرشماری، نوع نمایش‌های معمول در نقشه محدوده‌های سرشماری را در نظر بگیرید که درصد مردم یا واحدهای مسکونی دارای مشخصه‌های خاص را در محدوده مورد نظر، با رنگهای مختلف نشان می‌دهد. چنانچه هدف نمایش محل تجمع نژادهای مختلف در شهرستان باشد، سئوال این است که در سال ۲۰۰۳

12) Typical Census Users 13) Random Noise

از نقشه‌های سرشماری ۲۰۰۰ استفاده شود، چه فرضیاتی راجع به تغییر طی زمان باید به‌کار گرفته شود و اگر به‌جای اطلاعات سرشمادی ۲۰۰۰ از متوسط اطلاعات سال‌های ۲۰۰۲ـ۱۹۹۸ استفاده شود تفسیرها چه تفاوتی می‌کند؟ با توجه به اینکه موضوع مورد نظر «وضعیت کنونی» است، از برآوردهای سرشماری در مدل استفاده می‌شود با این فرض که این برآوردها، وضعیت کنونی را توصیف می‌کنند. هرچند ممکن است برای بسیاری از نواحی، مقادیر از سال ۲۰۰۰ کاهش یا افزایش یافته باشد اما هیچ تعدیل مشخصی برای تصحیح این تغییرات انجام نمی‌شود.

به‌بیان دیگر مدل ۲ عبارتست از:

- فرض مجازی (مدل پیش گزینه)

$$X_{۱۹۹۸} = X_{۱۹۹۹} = \cdots = X_{۲۰۰۳} = \mu$$

- گزینه ۱ (روند)

$$X_t = X_{۱۹۹۸} + C(t - ۱۹۹۸) \quad ; \quad C \neq ۰$$

- گزینه ۲ (جهش ناگهانی)

$$\begin{array}{rcl} X_t = X_{۱۹۹۸} & ; & t < I \\ X_t = X_{۱۹۹۸} + C & ; & t \geq I \end{array}$$

ACS دارای این مزیت است که با استفاده از سری‌های برآورد تک تک سالها، اطلاعاتی راجع به نواحی‌ای که دارای روند و جهش هستند، ارائه می‌دهد.

۳.۱.۴ مدل ۳: مدل «اغتشاش تصادفی»

گاهی تغییرات بی‌قاعده‌تراز «روندها» و «جهش‌های ناگهانی» هستند. فرض کنید که مقدار واقعی برای ناحیهٔ مورد بررسی به‌صورت زیر باشد

$$\begin{array}{rcl} X_t & = & \mu + \eta_t \\ E(\eta_t) & = & ۰ \end{array}$$

و ηها غیر همبسته باشند.

می‌توان یک شهرکوچک شامل ۲۰ واحد مسکونی را درنظرگرفت که در سال اول ۳ خانوار، سال دوم ۶ خانوار، سال سوم ۴ خانوار و... در آن فقیر باشند. در این حالت، متوسط ۵ ساله به‌عنوان برآورد μ و η به‌عنوان «اغتشاش» غیر دلخواه تعبیر می‌شود که مربوط به تغییرات مقدار واقعی در جامعه است.

۲.۴ استفاده از متوسط‌های چند سالانه برای اندازه‌گیری مقدار تخصیص بودجه

بسیاری از فرمولها که برای تخصیص بودجه برنامه‌های دولت فدرال بکار برده می‌شود از داده‌های آخرین سرشماری استفاده می‌کنند. به این منظور از رویکردهای گوناگونی استفاده می‌شود که در این جا مدل ساده زیر ارائه شده است. نیاز واقعی به بودجه در سال جاری بوسیلهٔ متغیر X_t برای ناحیه جغرافیایی خاص اندازه‌گیری می‌شود. از داده‌های در دسترس برای ارزیابی نیاز جاری به‌صورت زیر استفاده می‌شود:

$$\hat{A}_t = f(..., \hat{X}_{t-i}, ..., \hat{X}_{t-۱})$$

که در آن f یک تابع و $\hat{X}_{t-i}$ برآورد نمونه در سال $t-i$ می‌باشد. فرض شده است که برآورد سال جاری $\hat{X}_t$ در زمانی که به آن نیاز است در دسترس نیست. توزیع سری زمانی $\{X_t\}$ برای هر ناحیه نامعلوم است و ممکن است دارای الگوهای متفاوتی در مناطق مختلف و برای متغیرهای مختلف باشد.

اگر توزیع سری زمانی $\{X_t\}$ را بتوان برای ناحیه و متغیر مورد نظر تعیین کرد، تابع پیش‌بینی بهینهٔ f تعیین می‌گردد. گرچه مقدار بهینهٔ f ممکن است بسته به ناحیه تغییرکند. مقدار متوسط می‌تواند به‌عنوان یک پیش‌بینی کنندهٔ «توافقی[14]» برای همهٔ ناحیه‌ها استفاده شود. گزینه‌های زیر را می‌توان برای برآورد نیاز جاری به کمک متوسط‌های ACS درنظر گرفت.

۱ ساله: $\hat{A}_t = \hat{X}_{t-۱}$
۳ ساله: $\hat{A}_t = ۱/۳(\hat{X}_{t-۳} + \hat{X}_{t-۲} + \hat{X}_{t-۱})$
۵ ساله: $\hat{A}_t = ۱/۵(\hat{X}_{t-۵} + \cdots + \hat{X}_{t-۱})$

اما در صورت عدم استفاده از متوسط‌های ACS و استفاده از سرشماری قبلی، $\hat{A}_t = \hat{X}_۰$ $t = ۲, ..., ۱۱$ که در آن $t = ۰$ مشخص کنندهٔ سال آخرین سرشماری است.

بسیاری پیشنهاد کرده‌اند که بهتر است بجای میانگین ساده از میانگین وزنی با وزنهای بزرگ‌تر برای سالهای اخیر استفاده شود. (کارلز[15] ۲۰۰۱)

۳.۴ اندازه‌گیری تغییرات در طی زمان

ACS دارای کاربردهای بالقوه‌ای برای اندازه‌گیری تغییرات در طول زمان است. در صورتی که بدون داشتن نمونه‌های گردان امکان اندازه‌گیری این تغییرات وجود ندارد. در نواحی کوچک با نمونه‌های غیرمتداخل، احتیاط در تعبیر نتایج به دلیل انحراف استاندارد زیاد، لازم است. همچنین بدیهی است که در اندازه‌گیری تغییرات در طی زمان، تغییرات ناشی از مسایلی همچون تورم (در متغییر درآمد)، و گذر زمان (در متغییر سن) و ... باید از تغییرات واقعی جدا شوند.

14) Compromise 15) Charles

۵ مزایا و معایب سرشمای گردان در مقایسه با شیوه سنتی سرشماری جمعیتی

همان گونه که قبلاً توضیح داده شد یک طرح نمونه‌گیری گردان که درآن $k = F$ است، سرشماری گردان نامیده می‌شود. در این قسمت مزایا و معایب این نوع سرشماری در مقایسه با سرشماری سنتی جمعیت تشریح می‌شود. مزایا و معایب سرشماری گردان نسبت به شیوه سنتی سرشماری جمعیتی را از جنبه‌های: خروجی‌های سرشماری، پذیرش دیدگاه، کاربر، استمرار عملیات، پوشش، قوانین و هزینه‌ها می‌توان بررسی نمود.

۱.۵ خروجی‌ها

در سرشماری گردان، بدلیل وجود اطلاعات دوره‌های زمانی متوالی، کیفیت آمارهای جمعیتی بهبود می‌یابد. عیب اصلی خروجی‌ها این است که نمی‌تواند تصویری لحظه‌ای از کل جامعه ارائه کند. دو مزیت مهم خروجی این نوع سرشماری را به‌صورت ذیل می‌توان بیان کرد:

۱ــ کیفیت برتر برآوردها در مقایسه با برآوردهای حاصل از طرحهای آمارگیری معمول در فواصل بین سرشماری‌ها

۲ــ سطح مناسب برای مدل‌بندی داده‌ها به نحوی که با برآوردهای کل جامعه سازگار شود.

مزیت دیگر خروجی‌های این سرشماری، امکان پذیر بودن تغییر محتوی سئوالات در طی دوره است که موجب انطباق بهتر نتایج آن با نیازهای در حال تغییر کاربران می‌باشد.

۲.۵ پذیرش دیدگاه کاربر

بدلیل امکان‌پذیر بودن تغییر محتوای سئوالات در طی دوره، دولتهای محلی از جمله مخالفان اصلی این نوع سرشماری هستند زیرا معتقدند که با این روش، مقایسه پذیری بین نواحی از بین می‌رود. برای جلب موافقت آنها لازم است آگاهی آنها از این روش را بالا برد.

۳.۵ استمرار عملیات

با توجه به تجربه آمارگیری جاری جمعیت آمریکا[۱۶](CPS) استمرار عملیات، امکان استفاده از نیروی میدانی ثابت یا نیمه ثابت را فراهم می‌آورد که موجب انباشتگی تجربه و توان کارشناسی و در نتیجه بهبود کیفیت اجرا می‌شود.

16) Current Population Survey

۴.۵ پوشش

پوشش در سرشماری‌های گردان با مشکلات متفاوتی به‌صورت ذیل مواجه است:

۱ــ شمول بیش از یکبار خانوارها در دوره ۱۰ ساله بدلیل عدم ایستایی جمعیت

۲ــ عدم شمول بعضی از خانوارها در دوره ۱۰ ساله بدلیل عدم ایستایی جمعیت

۳ــ پوشش مضاعف یا عدم پوشش خانوارهای دارای بیش از یک اقامتگاه

۴ــ مشکلات پوشش ناشی ازکم توجهی عمومی به این نوع سرشماری در مقایسه با سرشماری سنتی که به دلیل منحصر به فرد بودن آن یک واقعه‌ی مهم ملی تلقی می‌شود.

۵.۵ قوانین

این نوع سرشماری ممکن است نیازمند وضع قوانین جدید یا اصلاح قوانین موجود در زمینه فواصل زمانی اجرا، پوشش موضوعی و جغرافیایی و ... باشد.

۶.۵ هزینه

با توجه به بررسیهای اولیه انجام شده، سرشماری گردان پرهزینه‌تر از سرشماری سنتی است (حدود ۱۰ درصد بیشتر) البته در طولانی مدت و با استمرار عملیات، صرفه اقتصادی بخصوص با تغییر شیوه‌های اجرایی (از جمله آمارگیری به روش پستی) قابل مشاهده خواهد بود.

با توجه به مطالب فوق مزایا و معایب سرشماری گردان را می‌توان به شرح زیر برشمرد.

مزایا

- وجود اطلاعات مستمر
- انعطاف پذیری سرفصل‌های اطلاعات
- کارایی ناشی از استمرار عملیات
- استفاده ازکارکنان میدانی ثابت و مکان ثابت برای پردازش داده‌ها

معایب

- از دست رفتن تصویر لحظه‌ای جمعیت
- تضعیف اهمیت سرشماری از دیدگاه جامعه
- افزایش خطر تأثیر سؤ جریانات خارجی برکار میدانی
- نیاز به متقاعد نمودن کاربر و وضع قوانین جدید
- داشتن هزینه بیش از سرشماری‌های سنتی

۶ نتیجه‌گیری

با توجه به مزیت‌هایی که این شیوه نسبت به سرشماری سنتی جمعیت داراست و با توجه به این که تجربیات ناشی از اجرای چندسال این شیوه با تغییر شیوه‌های اجرایی می‌تواند بسیاری از معایب

این روش را کاهش داده یا برطرف نماید، انتظار می‌رود این روش بتواند جایگزینی مناسب برای سرشماری سنتی جمعیت باشد.

مراجع

[1] Alexander, C. H. (1998)," Recent Developments In the American Community Survey ", Proccedings of the Survey Research Methods Section, American Statistical Association, pp. 92-100.

[2] Alexander, C. H. and Wetrogan, S (2000)," Integrating the American Community Survey and the Intercensal Demographic Estimates Program.",Integrating the American Community Survey and the Intercensal Demographic, American Statistical Association, pp. 295-300.

[3] Alexander, C. H. (2001)," Still Rolling: Leslie Kish 'Rolling Samples' And The American Community Survey", Estimates Program.", Proccedings of Statistics Canada Symposium, pp. 1-10.

[4] Kish, L., Lovejoy, W., and Rackow, P. (1961),"A Multi-Stage Probability Sample for Sample for Traffic Surveys", Procedings of the Social Statistics Section, American Statistical Association, pp. 227-230

[5] Kish, L. (1965), Survey Sampling .John Wiley and Sons, New York.

[6] Kish, L.(1979),"Rolling Samples Instead of Census", Asia and Pacific Census Forum, G(1), August 1979, pp.12-13.

[7] Kish, L.(1981),"Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau", Washington, D.C., U.S. Government Printing Office.

[8] Kish, L.(1983),"Data Collection for Details over Space and Time",In T. Wright,(ed.), Statistical Methods and the Improvement of Data Quality, New York: Academic, pp. 72-84.

[9] Kish, L., and Verma, V (1983), Census Plus Samples: Combined Uses and Designs, Bulletin of the International Statistical Institute, pp. 66-82.

[10] Kish, L., and Verma, V., (1986), "Complete Census and Samples", Journal of Offical Statistics, 2, pp. 381-93. the Census Bureau", Washington, D.C., U.S. Government Printing Office.

[11] Kish, L.(1987),"Statistical Research Design, John Wiley and Sons, New York.

[12] Kish, L.(1990),"Rolling Samples and Censues", Survey Methodology, 16, 63-79.

[13] Kish, L.(1997),"Periodic and Rolling Samples and Census", Chapter 7 In Statistics and Public Policy, Bruce D. Spencer, ed., Clarendon Press, Oxford.

[14] Kish, L.(1998),"Space/Time Variations and Rolling Samples",Journal of Official Statistics, 14, 1, 1998, pp. 31-46.

فواصل پیش‌بینی برای مدلهای اتورگرسیو خطی با استفاده از تکنیکهای بوت استرپ

ملیحه قابل[۱]، صادق رضایی[۲]

[۱] دانشجوی کارشناسی ارشد دانشگاه صنعتی امیرکبیر
[۲] گروه آمار دانشگاه صنعتی امیرکبیر

چکیده: در این مقاله ابتدا به محاسبهٔ فواصل پیش‌بینی برای مدلهای اتورگرسیو خطی می‌پردازیم و روشهای بهبود این فواصل را بررسی می‌کنیم. سپس تکنیکهای بوت استرپ را برای برآورد پارامتر نامعین و کاهش اریبی این برآورد در نمونه‌های کوچک بکارمی‌بریم. همچنین تصحیح اریبی پارامتر برآورد شده را برای محاسبهٔ فواصل پیش‌بینی بکارگرفته و در پایان بوسیلهٔ شبیه‌سازی مونت کارلو برای مدل AR(1) فواصل اطمینان معرفی شده را با سطوح پوشش متفاوت مقایسه می‌کنیم.

واژه‌های کلیدی: فواصل پیش‌بینی، اتورگرسیو، بوت استرپ، تصحیح اریبی

۱ مقدمه

فاصلهٔ پیش‌بینی یک حد پایین و بالایی است با این احتمال که فرآیند تحقق یافتهٔ بعدی باید بین این دو حد قرار بگیرد. یک فاصله باید روی نمونهٔ مشاهده شده تحقق یافته شرطی شود. بعضی از موضوعات وابسته به هم، در ارتباط با محاسبهٔ فواصل پیش‌گویی با سطوح پوشش متناسب می‌باشند. که شامل برآورد پارامتر نامعین (یعنی پارامترهای مدل برآورد شده متغیر تصادفی‌اند) اریبی نمونهٔ کوچک از پارامتر برآورد شده، پیش‌بینی توابع غیر خطی از پارامتر برآوردشده و . . . . نشان خواهیم داد که این بررسی‌ها با تکنیکهای بوت استرپ متناسب می‌توانند در ارتباط باشند.

روش فاصله اطمینان BOX–JENKINS که توسط Jenkins و Box در سال ۱۹۷۶ اعمال شده شاید ساده‌ترین روش باشد. که در بخش دوم این مقاله به معرفی این روش پرداخته‌ایم. سپس فاصله اطمینان چندک Efron و فاصله اطمینان چندک Hall و فواصل t-چندک را معرفی کرده و این فواصل را در مدلهای اتورگرسیو بکار می‌بریم. در سال ۲۰۰۱، Ruiz، Romo و Pascual شیوه‌ای را اعمال کردند که با وجود نمایش رو به عقب فرآیند، نیازی به دوباره نمونه‌گیری نداشت. بنابراین این شیوه برای مدلهایی با مولفه‌های میانگین متحرک به سرعت بکارگرفته شد. در پایان بخش دوم فواصل اطمینان PRRرا معرفی می‌کنیم. بخش سوم این مقاله شامل تجویز برآورد نامعین و تصحیح اریبی پارامتر برآورد شده می‌باشد. و بخش چهارم شامل شبیه‌سازی

مونت کارلو برای فواصل پیش‌بینی PRR با تصحیح اریبی پارامتر برآورد شده می‌باشد و در پایان تخمین مونت کارلو را برای مدل AR(1) شبیه‌سازی کرده و با سطوح پوشش متفاوت آنها را مقایسه می‌کنیم.

۲ محاسبهٔ فواصل پیش‌بینی

مدل AR(p) را در نظر بگیرید.

$$Y_t = \delta + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + a_t \qquad (1)$$

که a_t یک دنبالهٔ iid با تابع توزیع F_a, $E[a_t] = 0$, $E[a_t^2] = \sigma_a^2 < \infty$ مینیمم میانگین مربع خطا(MMSE) پیش‌گویی k مرحلهٔ بعدی Y_{t+k} وقتی پارامترهای مدل معلوم هستند بصورت:

$$\tilde{y}_{T+k} = \delta + \phi_1 \tilde{y}_{T+k-1} + \ldots + \phi_p \tilde{y}_{Y+k-p} \qquad (2)$$

می‌باشد که $\tilde{y}_{T+i} = y_{T+i}$ برای $i \leq 0$. که y_t یک مقدار تحقق یافته از Y_T است.
فاصلهٔ پیش‌بینی استاندارد Box–jenkins (BJ) در سطح $(1-\alpha)$ بصورت زیر تعریف می‌شود:

$$\{\tilde{y}_{T+K} + z_{\frac{\alpha}{2}}\hat{\sigma}_k, \tilde{y}_{T+K} - z_{1-\frac{\alpha}{2}}\hat{\sigma}_k\} \qquad (3)$$

که $\hat{\sigma}_k$ برآورد واریانس خطای پیش‌بینی kـ مرحلهٔ بعدی است و $\gamma = \phi(Z_\gamma)$ که ϕ تابع توزیع نرمال استاندارد است. بنابراین روش BJیک فاصله را روی نمونهٔ مشاهده شدهٔ تحقق یافته شرطی می‌کند. حال در مدل بالا اگر $\{\delta, \phi_1, \ldots, \phi_p, \sigma_a^2\}$ معلوم باشند و F_a تابع توزیع گوسین باشد آنگاه چگالی پیش‌گویی یک مرحلهٔ بعدی در پریود $T+1$ به شرط اطلاعات قابل دسترس تا پریود T بصورت $Y_{T+1} \mid I_T = N(\tilde{y}_{T+1}, \sigma_a^2)$ می‌باشد که

$$I_T = [Y_T = y_t, \ldots, Y_{T-p+1} = y_{T-p+1}], \tilde{y}_{T+1} = \delta + \phi_1 y_T + \ldots + \phi_p y_{T-p+1}$$

در این حالت فاصلهٔ پیش‌بینی $(1-\alpha)$، BJ بصورت:$\{\tilde{y}_{T+1} \pm Z_{\frac{\alpha}{2}} \cdot \sigma_a\}$ می‌باشد.

$$P_r[Y_{T+1} \in \tilde{y}_{T+1} \pm Z_{\frac{\alpha}{2}} \cdot \sigma_a] = 1 - \alpha$$

$$P_r[Z_{\frac{\alpha}{2}} < \frac{Y_{T+1} - \tilde{y}_{T+1}}{\sigma_a} < Z_{\frac{\alpha}{2}}] = 1 - 2\phi(Z_{\frac{\alpha}{2}}) = 1 - \alpha$$

فواصل چند مرحله‌ای به روش مشابه بدست می‌آیند.

۱.۲ پیش‌گویی‌های بوت استرپ

چگونگی یک روش بوت استرپ استاندارد را بررسی می‌کنیم. مدل (۱) را برای $t = p+۱, \ldots, T$ با داشتن مقادیر $\hat{\delta}, \hat{\phi}_۱, \ldots, \hat{\phi}_p, \hat{\sigma}_a^۲$ و باقیمانده‌های $\hat{a}_t$ برآورد می‌کنیم. اگر باقیمانده‌ها دارای میانگین صفر نباشند آنها را مرکزی کرده و در صورت امکان با عامل $\sqrt{(T-p)/(T-۲p)}$ دوباره مقیاس‌بندی می‌کنیم. تابع توزیع تجربی باقیمانده‌های مرکزی شده و مقیاس‌بندی شده را با $F_{\hat{a}}$ نمایش می‌دهیم. تابع توزیع تجربی برای هر مقدار $\{\hat{a}_t\}$ و مقادیر $t = p+۱, \ldots, T$ برابر با $\frac{۱}{T-p}$ است. بطوریکه مستقلاً از $F_{\hat{a}}$ بصورت با جایگذاری متناظر با نمونه‌گیری خطاها استخراج شده‌اند. یک نمونهٔ بوت استرپ توسط نمونه‌گیری از $F_{\hat{a}}$ تولید کرده که باقیمانده‌ها بصورت $\{a_{b,t}^*\}$ برای $t = T+۱, \ldots, T+k$ مشخص می‌شوند. معادلهٔ اتورگرسیو در این حالت بصورت زیر می‌باشد:

$$y_{b,T+k}^* = \hat{\delta} + \hat{\phi}_۱ y_{b,T+k-۱}^* + \ldots + \hat{\phi}_p y_{T+k-p}^* + a_{b,T+k}^* \qquad (۴)$$

که $y_{b,T+s}^* = y_{T+s}$ برای $s \leq ۰$. با تکرار این عملیات B مرتبه یک نمونهٔ بوت استرپ y_b^* برای $b = ۱, \ldots, B$ خواهیم داشت که $y_b^* = [y_{b,T+۱}^*, \ldots, y_{b,T+k}^*]'$.

۲.۲ فاصله اطمینان چندک

افرون (Efron 1979) روش فاصله اطمینان چندک را بصورت زیر محاسبه کرد. فرض کنید یک مجموعهٔ دادهٔ X^* طبق رابطهٔ $\hat{P} \to X^*$ را تولید کرده‌ایم. تکرارهای بوت استرپ $\hat{\theta}^* = S(X^*)$ محاسبه شده‌اند. اگر $\hat{G}$ تابع توزیع تجمعی $\hat{\theta}^*$ باشد، فاصلهٔ چندک $۱ - ۲\alpha$ توسط چندکهای α و $۱ - \alpha$ از $\hat{G}$ بصورت زیر تعریف می‌شود:

$$[\hat{\theta}_{Lo}, \hat{\theta}_{Up}] = [\hat{G}^{-۱}(\alpha), \hat{G}^{-۱}(۱-\alpha)] \qquad (۵)$$

چون طبق تعریف $\hat{G}^{-۱}(\alpha) = \hat{\theta^*}_{\beta}^{(\alpha)}$ و $\alpha \cdot ۱۰۰۰$ امین چندک تجربی از مقادیر $\hat{\theta^*}(b)$ است. یعنی مقدار $\alpha \cdot \beta$، آنگاه فاصلهٔ چندک $(۱ - ۲\alpha)$ بصورت تقریبی برابر است با:

$$[\hat{\theta}_{Lo}, \hat{\theta}_{Up}] \approx [\theta^{*(\alpha)}_{\beta}, \theta^{*(۱-\alpha)}_{\beta}] \qquad (۶)$$

بخاطر محبوبیت روش چندک افرون، مختصراً روش چندک Hall را معرفی می‌کنیم. Hall(1988)، هفت روش را برای محاسبهٔ فواصل پیش‌بینی بررسی کرد. Hallروش چندک دیگری را معرفی کرد و توزیع تجربی چندکها را برای تعریف نقاط بحرانی بالا و پایین فواصل پیش‌بینی بکار گرفت. فرض کنید F_k^*، تابع توزیع تجربی از $\{y_{b,T+k,b=۱,\ldots,B}^*\}$ باشد آنگاه فاصله بصورت

$[F_k^{*-1}(\frac{\alpha}{2}), F_k^{*-1}(\frac{1-\alpha}{2})]$ می‌باشد. به عنوان مثال یک فاصله با پوشش ۹۵ درصد و $B = 1000$ تکرار بصورت زیر می‌باشد.

$$\alpha = 0/05, F_k^{*-1}(0/025) = L_k = y_{T+k}^{*(25)}\ F_k^{*-1}(0/975) = U_k = y_{T+k}^{*(975)}$$

که مقادیر L_k, U_k متناسب با آماره‌های ترتیبی از توزیع تجربی می‌باشند. بنابراین فاصلهٔ چندک Hall بصورت زیر محاسبه می‌شود:

$$[\hat{y}_{T+k} - t_{1-\frac{\alpha}{2}}, \hat{y}_{T+k} - t_{\frac{\alpha}{2}}] \qquad (7)$$

که $\hat{y}_{T+k}$ پیش‌بینی k مرحلهٔ بعدی بر اساس برآوردهای پارامتر نمونه و $y_T, \ldots, y_{T-p+1}$ است یعنی:

$$\hat{y}_{T+k} = \hat{\delta} + \hat{\phi}_1 \hat{y}_{T+k-1} + \ldots + \hat{\phi}_p \hat{y}_{T+k-p} \qquad (8)$$

که $\hat{y}_{T+s} = y_{T+s}$ برای $s \leq 0$. و $t_{\frac{\alpha}{2}}$ و $t_{1-\frac{\alpha}{2}}$ چندکهای توزیع بوت‌استرپ از $Y_{T+k}^* - \hat{y}_{T+k}$ است.
(مقادیر تحقق یافتهٔ Y_{T+k}^* بصورت $\{y_{b,T+k}^*, b = 1, \ldots, B\}$ است.) محاسبهٔ فواصل پیش‌بینی و تصحیح اریبی برآوردهای پارامتر با حل معادلهٔ زیر می‌تواند بررسی شود.

$$E[f_t(F_0, F_1) \mid F_0] = 0 \qquad (9)$$

که F_0 تابع توزیع جامعه و F_1 تابع توزیع تجربی از نمونهٔ $\chi = \{y_1, \ldots, y_T\}$ است. f_T تابعی از کلاس $\{f_t, t \in \tau\}$ می‌باشد که $t = \{t_1, t_2\}$ جوابهای معادله‌اند.
امید نسبت به جامعهٔ F_0 شرطی شده است. برای محاسبهٔ فواصل اطمینان $(1-\alpha)$ درصد برای آمارهٔ $\theta_0 = \theta(F_0)$، f_t می‌تواند بصورت زیر باشد.

$$f_t(F_0, F_1) = I\{\theta(F_1) - t_1 \leq \theta(F_0) \leq \theta(F_1) + t_2\} - (1-\alpha)$$

که $I\{A\}$ تابع نشانگر است و مقدار ۱ را می‌گیرد اگر پیشامد A رخ دهد و در غیر اینصورت برابر صفر است. $\theta_0 = Y_{Y+k}$ و $\theta_1 = \theta(F_1)$ برآوردی از پیش‌بینی k-مرحلهٔ بعدی براساس نمونهٔ داده‌ها بصورت $\hat{y}_{T+k}$ است. بنابراین فاصلهٔ اطمینان بصورت $\{\theta_1 - t_1, \theta_1 + t_2\}$ یا $\{\hat{y}_{T+k} - t_1, \hat{y}_{T+k} + t_2\}$ می‌باشد. به جای حل معادلهٔ جامعه که شامل مجهولهای $t, \theta(F_0), F_0$ می‌باشد، معادلهٔ نمونه را حل می‌کنیم. یعنی:

$$E[f_t(F_1, F_2) \mid F_1] = 0 \qquad (10)$$

که با جایگذاری F_0 بوسیلهٔ F_1 و F_1 بوسیلهٔ F_2 بدست آمده که F_2 تابع توزیع تجربی از داده‌های χ که دوباره نمونه‌گیری شده‌اند، می‌باشد که جواب معادلهٔ بالا $\hat{t} = \{\hat{t}_1, \hat{t}_2\}$ است، که در این حالت تابع f_t بصورت:

$$f_{\hat{t}}(F_1, F_2) = I\{\theta(F_2) - \hat{t}_1 \leq \theta(F_1) \leq \theta(F_2) + \hat{t}_2\} -)(1-\alpha)$$

و فاصلهٔ اطمینان بصورت $\{\theta_۱ - \hat{t}_۱, \theta_۱ + \hat{t}_۲\}$ می‌باشد. که جواب معادلهٔ نمونه باید تقریب خوبی برای معادلهٔ جامعه باشد. که Hall آن را به عنوان قاعدهٔ بوت استرپ معرفی می‌کند. که در حالت پیش‌بینی معادلهٔ زیر را حل می‌کنیم.

$$P(\{Y^*_{T+k} - \hat{t}_۱ \leq \hat{y}_{T+k} \leq Y^*_{T+k} + \hat{t}_۲\}) = ۱ - \alpha$$

برای یک فاصله با دمهای برابر داریم:

$$P_r(Y^*T + k - \hat{y}_{T+k} \leq \hat{t}_۱) = ۱ - \frac{\alpha}{۲}$$

$$P_r(Y^*T + k - \hat{y}_{T+k} \leq -\hat{t}_۲) = \frac{\alpha}{۲}$$

بنابراین $\hat{t}_۱$ چندک $(۱ - \frac{\alpha}{۲})$ از توزیع بوت‌استرپ $Y^*_{T+k} - \hat{y}_{T+k}$ و $-\hat{t}_۲$ چندک $\frac{\alpha}{۲}$ام می‌باشد.

۳.۲ فواصل t ـ چندک

علاوه بر فواصل چندک Hall می‌توان فواصل t ـ چندک را نیز محاسبه کرد. فاصلهٔ اطمینان t ـ چندک (Efron,Tibshirani 1993) بصورت$(\hat{\theta} - \hat{t}^{۱-\alpha} \cdot \hat{Se}, \hat{\theta} - \hat{t}^{\alpha} \cdot \hat{Se})$ می‌باشد. که فاصلهٔ t ـ چندک در مدل اتورگرسیو بصورت زیر محاسبه می‌شود.

$$[\hat{y}_{T+k} - t_{۱-\frac{\alpha}{۲}} \cdot \hat{\sigma}(\hat{y}_{T+k}), \hat{y}_{T+k} - t_{\frac{\alpha}{۲}}\hat{\sigma}(\hat{y}_{T+k})] \quad (۱۱)$$

که $t_{\frac{\alpha}{۲}}, t_{۱-\frac{\alpha}{۲}}$ در اینجا چندکهای بوت استرپ از $\frac{Y^*_{T+k} - \hat{y}_{T+k}}{\hat{\sigma}(Y^*_{T+k})}$ می‌باشند. یک ایده اینست که بوت‌استرپ استودنت شده $\frac{\hat{y}_{T+k} - Y_{T+k}}{\hat{\sigma}(\hat{y}_{T+k})}$ بهتر از $(\hat{y}_{T+k} - Y_{T+k})$ است. هر چند فواصل اطمینان بر اساس آماره‌های محوری، ممکن است بطور مجانبی دقیق‌تر از آماره‌های غیر محوری نباشند. حال بررسی می‌کنیم که آیا استودنت کردن دقت فواصل پیش‌بینی را بهبود می‌بخشد؟ از رابطهٔ (۸) پیش‌بینی k- مرحلهٔ بعدی برای $p = ۱$ می‌تواند بصورت زیر نوشته شود.

$$\hat{y}_{T+k} = \sum_{i=۰}^{k-۱} \hat{\phi}^i \hat{\delta} + \hat{\phi}^k y_T$$

که $\hat{y}_{T+s} = y_{T+s}$ برای $S \leq ۰$. و خطای استاندارد پیش‌بینی بصورت

$$\sigma(\hat{y}_{T+k}) = \sqrt{E[(\hat{y}_{T+k} - E(\hat{y}_{T+k})^۲]}$$

می‌باشد. با استفاده از تقریب $E[\hat{y}_{T+k}] \approx \sum_{i=0}^{k-1} \phi^i \delta + \phi^k y_T$ خواهیم داشت $\hat{y}_{T+k} - E(\hat{y}_{T+k}) = \sum_{i=0}^{k-1}(\hat{\phi}^i\hat{\delta} - \phi^i\delta) + (\hat{\phi}^k - \phi^k)y_T$ (Clement,Hendry) در سال ۱۹۹۸ رابطهٔ زیر را اثبات کردند.

$$\sigma^2(\hat{y}_{T+k}) = \sigma_a^2(\frac{1-\phi^k}{1-\phi})^2 T^{-1} + k^2\phi^{2(k-1)}(1-\phi^2)y_T^2 T^{-1} \quad (۱۲)$$

برای فرآیندهای ایستا و $\delta = 0$ یک عرض از مبداء برآورد شده است. برای ساختن این عملگر σ_a^2 را با یک برآورد نااریب و ϕ را با $\hat{\phi}$ جایگزین می‌کنیم.

$$\hat{\sigma}^2(\hat{y}_{T+k}) = \hat{\sigma}_a^2(\frac{1-\hat{\phi}^k}{1-\hat{\phi}})^2 T^{-1} + k^2\hat{\phi}^{2(k-1)}(1-\hat{\phi}^2)y_T^2 T^{-1} \quad (۱۳)$$

$\hat{\sigma}(Y^*_{T+k})$ بطور مشابه برآوردمی‌شود بطوری که برای نمونه bام بوت استرپ داریم:

$$\hat{\sigma}^2({y^*}_{b,T+k}) = \hat{\sigma}^{2*}_{b,a}(\frac{1-(\phi^*)^k_b}{1-\phi^*_b})^2 T^{-1} + k^2(\phi)^{*2(k-1)}_b(1-\phi^{*2}_b)y_T^2 T^{-1} \quad (۱۴)$$

که ${\phi^*}_b$ برآورد پارامتر بر اساس داده‌های بوت استرپ (دوباره نمونه‌گیری) و ${\sigma_{\hat{b},a}}^{*2}$ واریانس خطای برآورد شده بر اساس داده‌های بوت استرپ می‌باشد. اثر برآورد پارامتر را روی شکل چگالی‌های پیش‌بینی با استفاده از روش بوت استرپ که توسط Pascual ,Romo ,Ruiz در سال ۱۹۸۸ برای مدلهای ARIMA(p,d,q) اعمال شده را به نام فواصل اطمینان PRR معرفی می‌کنیم. ابتدا پارامترها توسط روش شبه درستنمایی ماکزیمم شرطی (QML) برآورد می‌شوند. روش بوت استرپ که توسط Pascual در سال ۱۹۸۸ برای ساخت فواصل پیش‌گویی برای مقادیر بعدی از سری تولید شده در فرآیند ARIMA(p,d,q) در نظر می‌گیریم.

$$\nabla^d y_t = \phi_0 + \phi_1\nabla^d y_t - 1 + \ldots + \phi_p\nabla^d y_{t-p} + a_t + \theta_1 a_{t-1} + \ldots + \theta_q a_t - q \quad (۱۵)$$

که a_t یک فرآیند اغتشاش خالص است و ∇ عملگر تفاضلی بصورت $\nabla y_t = y_t - y_{t-1}$ $(\phi_0, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_a)$ پارامترهای مجهول هستند. از سری مشاهده شدهٔ $\{y_1, \ldots, y_T\}$ پارامترها توسط یک برآورد سازگار مانند QML شرطی می‌توانند برآورد شوند. با داشتن مقادیر $(\hat{\phi}_0, \hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_a)$ باقیمانده‌ها بصورت زیر محاسبه می‌شوند.

$$\hat{a}_t = \nabla^d y_t - \hat{\phi}_0 - \hat{\phi}_1\nabla^d y_t - 1 - \ldots + \hat{\phi}_p\nabla^d y_{t-p} - \hat{\theta}_1\hat{a}_{t-1} + \ldots - \hat{\theta}_q\hat{a}_t - q \quad (۱۶)$$

که باقیمانده‌ها متناظر با پریودهایی از زمان $t = 0, -1, -2, \ldots$ مجموعه‌ای برابر صفر هستند. $\hat{F}_a$ را تابع توزیع تجربی از باقیمانده‌های مرکزی شده در نظر می‌گیریم. یک مجموعه از $p+q$

مقدار اولیه از متغیر y_t یعنی $\{y_۱, \ldots, y_{p+d}\}$ داده شده، یک تکرار بوت استرپ از سریهای $\{y^*_۱, \ldots, y^*_T\}$ توسط معادله زیر ساخته می‌شود.

$$\nabla^d_{y^*_t} = \hat{\phi}_۰ + \sum_{j=۱}^{p} \hat{\phi}_j \nabla^d_{y^*_t - j} + \sum_{j=۱}^{q} \hat{\theta}_j \hat{a}^*_{t-j} + \hat{a}^*_t \quad , t = p+d+۱, \ldots, T \qquad (۱۷)$$

که $y^*_t = y_t$ برای $t = ۱, \ldots, p+q$ و $\hat{a}^*_{۱+p+d-p}, \ldots, \hat{a}^*_T$ تصادفاً از $\hat{F}_a$ استخراج شده‌اند. یکبار دیگر پارامترها از این سریهای بوت استرپ برآورد می‌شوند. یعنی $(\hat{\phi}^*_۰, \hat{\phi}^*_۱, \ldots, \hat{\phi}^*_p, \hat{\theta}^*_۱, \ldots, \hat{\theta}^*_q)$ پیش گویی k- مرحله بعدی بصورت زیر می‌باشد.

$$\nabla^d_{y^*_{T+k}} = \hat{\phi}^*_۰ + \sum_{j=۱}^{p} \hat{\phi}^*_j \nabla^d_{y^*_T + k - j} + \sum_{j=۱}^{q} \hat{\theta}^*_j \hat{a}^*_{T+k-j} + \hat{a}^*_{T+k} \quad , k = ۱, ۲, \ldots \qquad (۱۸)$$

که

$$k \le j \quad , y^*_{T+k-j} = y_{T+k-j}$$
$$k \le j \quad , a^*_{T+k-j} = \hat{a}_{T+k-j}$$

یعنی $p+d$ مشاهده آخر از سریها و q باقیمانده آخر در مرتبه‌ای که چگالی پیش‌گویی شرطی روی داده‌های مشاهده شده بدست می‌آید، ثابت است. حدود پیش‌گویی به صورت چندکهای تابع توزیع بوت استرپ از y^*_{T+k} تعریف می‌شوند. یعنی اگر $G^*(h) = P(y^*_{T+k} \le h)$ تابع توزیع y^*_{T+k} باشد و برآورد مونت کارلو بصورت $G^*_B(h) = \sharp \frac{(y^*_{T+k} \le h)}{B}$ می‌باشد که یک فاصله پیش‌بینی $۱۰۰\alpha\%$ برای Y^*_{T+k} بصورت زیر تعریف می‌شود.

$$[L^*_B, U^*_B] = [Q^*_B(\frac{۱-\alpha}{۲}), Q^*_B(\frac{۱+\alpha}{۲})] \qquad (۱۹)$$

که $Q^*_B = G^{*-۱}_B$. این روش را که تغییرات مربوط به برآورد پارامتر را در نظر می‌گیرد PRR می‌نامیم. بنابراین روش بوت استرپ PRR به آسانی برای مدلهایی با میانگین متحرک استفاده می‌شود.

۳ تجویز برآورد نامعین[۱]

در بخش قبل خطاها نرمال فرض نشده بودند و تغییرات در برآوردهای پارامتر نادیده گرفته شده بود. یک نمونه بوت استرپ از T مشاهده با نمونه‌گیری از $F_{\hat{a}}$ بصورت زیر تولید می‌شود.

$$y^*_{b,t} = \hat{\delta} + \hat{\phi}_۱ \hat{y}^*_{b,t-۱} + \ldots + \hat{\phi}_p y^*_{b,t-p} + a^*_{b,t} \qquad (۲۰)$$

1) Allowing Estimation Uncertainty

(خطاها با $\{a^*_{b,t}\}$ برای $t = p+1,\ldots,T$ مشخص می‌شوند). برای $b = p+1,\ldots,T$ به‌شرط مقادیر مشاهده شدهٔ اولیهٔ $\{y_1,\ldots,y_p\}$ بطوریکه $y^*_{b,t} = y_t$ و برای $t = 1,\ldots,p$ که نمونهٔ $\{y^*_{b,p+1},\ldots,y^*_{b,T}\}$ را می‌دهد. با تولید B نمونه به این روش برآورد پارامتر در مدل (۱) بصورت $\theta^{*\prime}_b = \{\hat{\delta}_b, \hat{\phi}_{1,b},\ldots,\hat{\phi}_{p,b}\}$ خواهند بود و معادله بصورت زیر می‌باشد.

$$y^*_{b,T+k} = \hat{\delta}_b + \hat{\phi}_{1,b}\, y^*_{b,T+k-1} + \ldots + \hat{\phi}_{p,b} y^*_{b,T+k-p} + a^*_{b,T+k} \qquad (۲۱)$$

$$k = 1, 2,\ldots, b = 1,\ldots,B$$

که $\{a^*_{b,T+1},\ldots,a^*_{b,T+k}\}$ همانند قبل از $F_{\hat{a}}$ استخراج شده‌اند. توزیع تجربی از $\{y^*_b\}$ برآورد بوت استرپ، از توزیعهای پیش‌بینی مجهول است. اثر برآورد نامعین اینک متشکل از هر یک از تکرارهای بوت استرپ داده قبلی که دوباره استخراج شده و مدل برآورد شده روی آن است.

۱.۳ تصحیح اریبی پارامتر برآورد شده

دیدیم که تکرارهای بوت استرپ از برآورد پارامتر اصلی تولید شدند. Kilian(1998) نشان داد که تصحیح اریبی برآوردهای $\hat{\theta}' = \{\hat{\delta}, \hat{\phi}_1,\ldots,\hat{\phi}_p\}$ در ساخت فواصل اطمینان برای پاسخهای ناگهانی می‌تواند مفید باشد. شباهت بین پیش‌بینی‌های چند مرحله‌ای و پاسخهای ناگهانی یک استراتژی مشابهی را پیشنهاد می‌کند، به‌ویژه برای k بزرگ، پارامترهای با توان بالا ممکن است اریبی را بیشتر کنند. وقتی حجم نمونه کوچک است و ریشه‌ها به ناحیه ناایستا نزدیک هستند، نادیده گرفتن اریبی نمونه کوچک ممکن است در فواصل با پوشش ضعیف نتیجه شوند. فرض کنید $\theta' = \{\delta, \phi_1,\ldots,\phi_p\}$ و ψ مقدار اریبی در برآورد θ باشد، $\psi = E[\hat{\theta} - \theta]$ که یک برآورد بر اساس بوت استرپ بصورت $\hat{\psi} = \bar{\theta^*} - \hat{\theta}$ است که $\bar{\theta^*} = B^{-1}\sum_{b=1}^{B}\theta^*_b$ و θ^*_b برآوردهایی از پارامترهای مدل (۱) برای هر B تکرار بوت استرپ هستند که از $\hat{\theta}$ و $F_{\hat{a}}$ تولید شده‌اند. شکل تابع $f_t(F_0, F_1) = \theta(F_1) - \theta(F_0) - \psi$ در این حالت بصورت $f_t(F_0, F_1)$ می‌باشد که $t = \psi$ اریبی را در برآورد پارامترهای جامعه مشخص می‌کند. پارامترهای جامعه $\theta(F_0)$ و برآوردهای OLS بر اساس نمونهٔ $\chi = \{y_1,\ldots,y_T\}$ بصورت $\hat{\theta} = \theta(F_1)$ هستند. بنابراین معادلهٔ جامعه بصورت زیر می‌باشد:

$$E[\theta(F_1) - \theta(F_0) - \psi \mid F_0] = 0 \qquad (۲۲)$$

که برای به دست آوردن برآوردی از $\hat{\psi}$ که $E[\theta(F_2) \mid F_1] = \bar{\theta^*} = B^{-1}\sum_{b=1}^{B}\theta^*_b$ معادلهٔ نمونه را حل می‌کنیم.

$$E[\theta(F_2) - \theta(F_1) - \psi \mid F_1] = 0 \qquad (۲۳)$$

که $\theta^*_b = \theta(F_{2b})$ و F_{2b} تابع توزیع تجربی از χ^*_b می‌باشد که از F_1 استخراج شده است. امید ریاضی θ^* بوسیلهٔ میانگین تعداد نمونه‌های زیاد بوت استرپ تقریب زده می‌شود. این تصحیح

اگر لازم باشد، در شکل ساده‌ای می‌تواند تکرار شود. به عنوان مثال اگر در (۲۳)، $\theta(F_1)$ با $\tilde{\theta} = \hat{\theta} - \hat{\psi}$ جایگزین شود آنگاه F_{2b} تابع توزیع تجربی از یک نمونه تولید شده از تصحیح اریبی برآوردهاست. برای مدلهایی با ریشه‌های نزدیک به ریشه‌های نزدیک به دایره واحد، باید مطمئن شویم که تصحیح اریبی تعداد ریشه‌های واحد را تغییر نمی‌دهد. در $|\hat{\phi}(z)| = 0$ (که $\hat{\phi} = 1 - \hat{\phi}_1 L - \ldots - \hat{\phi}_p L^p$) فرضاً $\hat{\theta}$ درون ناحیهٔ ایستا قرار بگیرد با تصحیح اریبی $\tilde{\theta} = \hat{\theta} - \hat{\psi}$ آنگاه $\tilde{\theta}$ به سادگی در ۲۰ جایگزین $\hat{\theta}$ برای تولید B نمونهٔ بوت استرپ می‌شود. برآوردهای پارامتر روی این نمونه‌ها به دست می‌آیند و با θ_b^* و تصحیح اریبی $\hat{\psi}$ $\tilde{\theta}_b^* = \theta_b^* - \hat{\psi}$ مشخص می‌شوند. در عمل عوامل تصحیح اریبی را برای هر θ_b^* می‌توان برآورد کرد. در انجام تصحیحهای اریبی باید مطمئن شویم که تصحیح باعث پیش‌بینی‌های ایستا به ناایستا نمی‌شود.

۲.۳ الگوریتم مونت کارلو برای فاصلهٔ پیش‌بینی PRR

بطور خلاصه یک تخمین مونت کارلو از فواصل پیش‌بینی PRR با تصحیح اریبی برآوردهای پارامتر بصورت زیر اعمال می‌شود.

مرحله ۱) شبیه‌سازی سری $\{y_1, \ldots, y_T\}$ به طول T از مدل AR(P)با توزیع خطای گوسین $F(\hat{a})$ شبیه‌سازی R دنباله از سریها به طول K با بیشترین حد پیش‌بینی. بکارگیری مقادیر پارامترکه از توزیع خطای $F(a)$ گرفته شده و روی $y_{T-p+1}, \ldots, y_T$ برای یک فرآیند مرتبهٔ pام شرطی شده است. این دنباله‌ها مقادیر تحقق یافتهٔ آینده هستند و برای برآوردهای فواصل پوشش بکار می‌روند.

مرحله ۲) برآورد $\hat{\theta}$ و F_a روی $\{y_1, \ldots, y_T\}$ و انجام بوت استرپ برای تصحیح اریبی $\hat{\theta}$

مرحله ۳) شبیه‌سازی B تکرار بوت استرپ به طول T با بکارگیری $\tilde{\theta}$ و F_a و برآورد (۱) برای بدست آوردن θ_b^*. برآورد اریبی را از مرحلهٔ ۲ برای تصحیح اریبی این برآوردها برای داشتن θ_b^* بکار می‌بریم.

مرحله ۴) با شرطی کردن روی $\{y_{T-p+1}, \ldots, y_T\}$ برای هر $\tilde{\theta}_b^*$ گرفته شده از $F_{\hat{a}}$ و دنباله‌ای از سریها تولید می‌کنیم. که $\{y^*1, T+k, \ldots, y^*_{B,T+k}\}$ را برای $k = 1, \ldots, K$ می‌دهد. فرض کنید L^*_k وU^*_k فواصل انتهایی در روش چندک Hall را مشخص کنند آنگاه $\lambda^*_k = U^*_k - L^*_k$ طول فاصله و پوشش $\beta^*_k = \sharp\frac{\{L^*k \le y^r_{T+k} \le U^*_k\}}{R}$ که اندیس r, r امین عضو از R دنبالهٔ شبیه‌سازی مرحلهٔ ۱ است.

مرحله ۵) تکرار مراحل ۴ــ۱، M مرتبه، مقادیر بدست آمده روی هر تکرار با m اندیس گذاری می‌کنیم $m = 1, \ldots, M$.

مرحله ۶) محاسبهٔ طول و پوشش برآوردهای مونت کارلو که تغییرات این برآوردها بصورت زیر می‌باشد.

$$\bar{\beta}^*_k = \frac{\sum_{m=1}^{M} \beta^*_{k,m}}{M}$$

$$S.E(\bar{\beta}_k^*) = \{\frac{[\frac{\sum_{m=1}^{M}(\beta_{k,m}^* - \bar{\beta}_k^*)^2}{M-1}]}{M}\}^{\frac{1}{2}}$$

$$\bar{\lambda}_k^* = \frac{\sum_{m=1}^{M} \lambda^* k, m}{M}$$

$$S.E(\bar{\lambda}_k^*) = \{\frac{[\frac{\sum_{m=1}^{M}(\lambda_{k,m}^* - \bar{\lambda}_k^*)^2}{M-1}]}{M}\}^{\frac{1}{2}}$$

۳.۳ شبیه‌سازی مونت کارلو برای یک مدل AR(1)

تخمین مونت کارلو را برای مدل AR(1)، $Y_t = ۰/۹۵Y_{t-1} + a_t$ محاسبه می‌کنیم. برآوردهایی به حجم $T = \{۲۵, ۵۰\}$ با ماکزیمم حد پیش‌بینی $k = ۱۰$ را در نظر می‌گیریم. فواصل اطمینان Box–Jekins (BJ) و فواصل PRR و فواصل PRR تصحیح اریبی شدهٔ kialin (k) را با سطوح پوشش $c = \%۸۰$ و $c = \%۹۰$ شبیه‌سازی و مقایسه می‌کنیم.

مراجع

[1] Michael P. Clements, Nick Taylor.(2001),Bootstrapping prediction intervals for autoregressive models, International Journalof Forecasting 17,247-267.

[2] Pascual, Romo, Ruiz.(2001)Effects of parameter estimation on prediction densities :a bootstrap approach International Journal of Forecasting 17,83-103.

[3] Thombs, Schucany.(1990)Bootstrap prediction intervals for autoregression Journal of the American Statistical Association 85,NO 410.

[4] Efron, Tibshirani.(1993)An introduction to the Bootstrap.

[5] Box,Jenkins.(1976)Time Series Analysis:Forecasting and control.

تعیین تعداد خوشه‌ها در تحلیل‌های آمیخته با استفاده از آنتروپی نرمال شده

محمد قربانی[۱]، محسن محمدزاده[۲]

[۱] گروه آمار دانشگاه تبریز
[۲] گروه آمار دانشگاه تربیت مدرس

چکیده: یکی از مسائل مهم در تحلیل خوشه‌ای، تعیین تعداد خوشه‌هاست. علی‌رغم اینکه این موضوع توسط محققان زیادی مورد بحث واقع شده ولی همچنان به عنوان یک مشکل خوشه‌بندی مطرح است. در این مقاله هدف، تعیین تعداد خوشه‌ها با استفاده از آنتروپی نرمال‌شده در تحلیل آمیخته است که این معیار از ارتباط بین درستنمایی کلی و درستنمایی رده‌بندی به دست می‌آید. همچنین با استفاده از شبیه‌سازی مونت کارلو نشان داده خواهد شد که این معیار بهتر از معیارهایی مانند BIC و AIC عمل می‌کند.

واژه‌های کلیدی: تحلیل آمیخته، تعداد خوشه‌ها، آنتروپی نرمال‌شده، درستنمایی رده‌بندی

۱ مقدمه

یکی از مسائل مهم در بسیاری از مطالعات ژنتیکی و پزشکی، خوشه‌بندی داده‌ها است، که در آن N مشاهده با M ویژگی به g گروه همگن افراز می‌شوند، به طوری که تشابهات درون گروهها ماکسیمم گردند. روشهای مختلفی از جمله، روشهای سلسله مراتبی برای خوشه‌بندی کردن داده‌ها به کار می‌روند که در تعریف «فاصله بین دو خوشه» با هم تفاوت دارند (هارتیگان، ۱۹۷۵). همچنین به دلیل اینکه این روشها براساس مدل خاصی بنا نشده‌اند، استنباط آماری بر اساس آنها امکان‌پذیر نیست و تعداد خوشه‌ها نیز به صورت ابتکاری با تعریف آستانه‌ای دلخواه تعیین می‌شود. برای رفع این مشکلات لازم است روش خوشه‌بندی حتی الامکان مبتنی بر سلیقه محقق نبوده و بر اساس یک مدل یا توزیع احتمالی باشد تا بتوان در مورد آن استنباط آماری انجام داد. معمولاً مجموعه مشاهدات تحت بررسی، همگی از یک جامعهٔ خاص نیستند و برای تشخیص این که هر مشاهده از کدام جامعه آمده است، منطقی است فرض شود که هر مشاهده بر اساس ویژگی‌ها و خصوصیاتش دارای توزیع احتمال خاصی است. بنابراین جامعه‌ای مرکب از چند زیرجامعه، دارای توزیع احتمالی آمیخته بصورت

$$f(x|\psi) = \lambda_1 f_1(x|\theta_1) + \ldots + \lambda_g f(x|\theta_g)$$

است، که در آن برای $j = ۱,\ldots,g$، $f_j(.)$ تابع چگالی مولفه‌ها، $۰ < \lambda_j \leq ۱$ و $\sum_{j=۱}^{g} \lambda_j = ۱$، $\lambda = (\lambda_۱,\ldots,\lambda_g)$ و $\theta = (\theta_۱,\ldots,\theta_g)$ و $\psi = (\lambda,\theta)$ است. معمولاً فرض می‌شود مولفه‌ها دارای توزیع تک مدی $n(\mu_j,\sigma_j^۲)$ است و هدف از خوشه‌بندی، تجزیه مولفه‌های چند بعدی مبهم و آمیخته به مولفه‌های ساده تک مدی است. یکی از مسائل مهم خوشه‌بندی تعیین تعداد خوشه‌ها (مولفه‌ها) است. علی‌رغم اینکه این موضوع توسط محققان زیادی مورد بحث واقع شده ولی همچنان به عنوان یک مشکل خوشه‌بندی مطرح است. در این مقاله از معیاری بنام معیار آنتروپی نرمال شده برای تعیین تعداد خوشه‌ها استفاده می‌شود و بکمک تکنیک شبیه‌سازی مونت کارلو نشان داده خواهد شد این معیار بهتر از معیارهایی مانند BIC و AIC عمل می‌کند.

۲ تعیین معیارهای خوشه‌بندی بر اساس مدل

فرض کنید $X_۱،\ldots،X_n$ یک نمونه تصادفی از جامعه Π با زیر جامعه‌های $\Pi_۱،\ldots،\Pi_g$ باشند و $\phi(x|\mu_j,\Sigma_j)$ تابع چگالی متغیر تصادفی X_i، در زیر جامعه jام باشد. در این صورت اگر λ_j احتمال تعلق X_i به جامعه Π_j باشد، توزیع X_i عبارت است از

$$f(x;\psi) = \sum_{j=۱}^{g} \lambda_j \phi(x|\mu_j,\Sigma_j) \qquad (۱)$$

که در آن $\theta_j = (\mu_j,\Sigma_j)$، $\psi = (\lambda_۱,\ldots,\lambda_g,\theta_۱,\ldots,\theta_g)$، $\sum_{j=۱}^{g} \lambda_j = ۱$ و $\lambda_j > ۰$ و تابع درستنمایی نمونه تصادفی به صورت

$$L(\psi|x) = \prod_{i=۱}^{n} \sum_{j=۱}^{g} \lambda_j f_j(x|\theta_j) \qquad (۲)$$

خواهد بود. برای $C_j = \{i, X_i \in \Pi_j\}$ معیار خوشه‌بندی حاصل از ماکسیمم کردن تابع درستنمایی

$$L(x|C) = \prod_{j=۱}^{g} \lambda_j^{n_j} \prod_{X_i \in C_j} f(x_i|\theta_j) \qquad (۳)$$

معادل معیار حاصل از ماکسیمم کردن (۲) خواهد بود (مک‌لن، ۱۹۸۲ و فرالی و رافتری، ۱۹۹۸ و وحیدی و همکاران، ۱۳۸۱). معمولاً مولفه اصلی X_iها معلوم نیستند و برای مشخص کردن مولفه اصلی X_i متغیرهای گروه‌بندی Z_{ij} به صورت

$$Z_{ij} = \begin{cases} ۱ & X_i \in \Pi_j \\ ۰ & X_i \notin \Pi_j \end{cases}$$

تعریف می‌شوند. بر اساس مشخصه‌های گروه‌بندی لگاریتم تابع درستنمایی را می‌توان به صورت

$$\log L(\psi|x,Z)=\sum_{j=1}^{g}\sum_{i=1}^{n} z_{ij}\log\{\lambda_j f(x_i/\theta_j)\} \qquad (۴)$$

نوشت. اگر z_{ij}ها مشخص باشند، عناصر خوشهjام به صورت
$C_j=\{j;\quad z_{ij}>z_{ij'}\quad j\neq j'\}$ خواهد بود.
ولی اگر z_{ij}ها معلوم نباشند برای انجام تحلیل خوشه‌ای از برآورد مقدار مورد انتظار آنها استفاده می‌شود. پس تحلیل خوشه‌ای را می‌توان به عنوان برآورد مقدار مورد انتظار z_{ij}، یعنی

$$E(Z_{ij})=P(X_i\in\Pi_j)=\frac{\lambda_j f(x_i|\theta_j)}{\sum_{j=1}^{g}\lambda_j f(x_i|\theta_j)}$$

تلقی نمود، که لازم است برای هر j پارامترهای نامعلوم λ_j و θ_j برآورد شوند.

الگوریتم امید ریاضی و ماکسیمم‌سازی[1] (EM) روشی کاربردی در محاسبات تکراری، برای به دست آوردن برآورد ماکسیمم درستنمایی پارامترها در توزیع‌های آمیخته است. فرض کنید $\psi^{(0)}$ مقدار اولیه ψ باشد، در این صورت مراحل الگوریتم EM به صورت زیر خواهد بود.(مک‌لن و کریشنان، ۱۹۹۷):

مرحله E: محاسبه امید ریاضی لگاریتم تابع درستنمایی در نقطه $\psi^{(0)}$ به شرط مشاهده داده‌های کامل،

$$Q(\psi,\psi^{(0)})=E_{\psi^{(0)}}\{\log L_c(\psi|x)\}$$

مرحلهٔ M: بدست آوردن مقداری مانند ψ^* برای ψ به طوری که

$$Q(\psi^*,\psi^{(0)})=\max_{\psi\in\Omega}(Q(\psi,\psi^{(0)}))$$

مراحل E و M تا زمانی تکرار می‌شوندکه شرط همگرایی $\forall\epsilon>0:|L(\psi^{(k+1)})-L(\psi^{(k)})|<\epsilon$ برقرار شود. (جی اف وو، ۱۹۸۳).

فرض کنید X_1،...،X_n داده‌های ناکامل و $Y_i=(X_i,Z_i)$ داده‌های کامل در الگوریتم EM باشند. در این صورت برآورد پارامترهای (۴) با استفاده از این الگوریتم به صورت زیر خواهند بود.

$$\lambda_j^{(k+1)}=\frac{1}{n}\sum_{j=1}^{n} z_{ji}^{(k)}$$

1) Expectation-Maximization Algorithm

$$\mu_i^{(k+1)} = \sum_{j=1}^{n} \frac{z_{ji}^{(k)} X_j}{\sum_{j=1}^{n} z_{ij}^{(k)}}$$

$$(\Sigma)^{(k+1)} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n} z_{ij}^{(k)} (X_j - \mu_i^{(k+1)})(X_j - \mu_i^{(k+1)})'$$

$$z_{ij}^{(k)} = \frac{\pi_i^{(k)} f(x_j|\theta_i^{(k)})}{\sum_{i=1}^{k} \pi_f^{(k)}(x_j|\theta_i^{(k)})}$$

با برآورد z_{ij} به وسیله الگوریتم EM می توان تحلیل خوشه‌ای را متناسب با بیشترین مقدار $z_{ij}^{(k)}$ انجام داد (وحیدی اصل، محمدزاده و قربانی، ۱۳۸۱ و مک‌لن و کریشنان، ۱۹۹۷). ولی یکی از مسائل اساسی خوشه‌بندی، تعیین تعداد خوشه‌هاست این موضوع با استفاده از معیارهای مختلف توسط اغلب محققان (وحیدی و محمدزاده و قربانی، ۱۳۸۱، فرالی و رافتری، ۱۹۹۸) بررسی شده ولی همچنان به عنوان یک مشکل خوشه‌بندی مطرح است.

۳ تعیین تعداد خوشه‌ها

برای بیان معیار آنتروپی نرمال شده در تعیین تعداد خوشه‌ها نخست به تعریف آنتروپی می‌پردازیم. متغیر تصادفی X با تابع جرم احتمالی $P(x) = P(X = x)$ و مقادیر $\{x_i,\ \ i = 1, \ldots, n\}$ در نظر بگیرید، عدم قطعیت مرتبط با برآمد یک مشاهده از متغیر X را آنتروپی متغیر تصادفی X نامیده و با نماد $H(X)$ نشان می‌دهند که عبارت است از

$$H(X) = -\sum P(x) \log P(x)$$

حال لگاریتم تابع درستنمایی (۲) را در نظر بگیرید

$$L(g) = \sum_{i=1}^{n} \ln \sum_{j=1}^{g} \hat{\lambda}_j f_j(x|\hat{\theta}_j)$$

که در آن $\hat{\lambda}_j$ و $\hat{\theta}_j$ برآوردهای ماکزیمم درستنمایی λ_j و θ_j می‌باشند. با استفاده از محاسبات مستقیم می‌توان نشان داد

$$L(g) = C(g) + E(g)$$

که در آن

$$C(g) = \sum_{j=1}^{g} \sum_{i=1}^{n} \hat{z}_{ij} \log \left\{ \hat{\lambda}_j f(x_i|\hat{\theta}_j) \right\}$$

و

$$E(g) = -\sum_{j=1}^{g}\sum_{i=1}^{n} \hat{z}_{ij} \ln \hat{z}_{ij} \geq 0$$

بر اساس روابط فوق، لگاریتم تابع درستنمایی $L(g)$، به لگاریتم تابع درستنمایی رده‌بندی $C(g)$ و معیار آنتروپی $E(g)$، تفکیک شده است. اگر چه آنتروپی $E(g)$ یک معیار اطلاع است ولی نمی‌توان مستقیماً از آن برای تعیین تعداد خوشه‌ها استفاده کرد زیرا $L(g)$ یک تابع افزایشی از g است و باید نرمالیزه شود. داریم

$$1 = \frac{C(g) - C(1)}{L(g) - L(1)} + \frac{E(g) - E(1)}{L(g) - L(1)} \qquad g > 1 \qquad (5)$$

و معیار آنتروپی نرمال شده که باید برای تعیین تعداد خوشه‌ها می‌نیمم شود عبارت است از

$$NEC(g) = \frac{E(g)}{L(g) - L(1)}, \qquad E(1) = 0$$

اما $NEC(1)$ خوش تعریف نیست زیرا $NEC(1) = \infty$ (بایرناکی، سیلوکس و گوورت ۱۹۹۹). لذا به طور مستقیم قادر به مقایسه حالت $g = 1$ در برابر $g > 1$ در استفاده از $NEC(g)$ نخواهیم بود. لذا می‌بایست $NEC(g)$ را برای رفع این مشکل خاص بسط دهیم. برای تصمیم‌گیری بین $g = 1$ و $g > 1$ می‌توان از این نقطه نظر که $C(g)$ معیاری برای اندازه‌گیری دقت افراز داده‌ها در g خوشه و $C(1) = L(1)$ معیاری برای اندازه‌گیری دقت برازش تک خوشه به داده‌هاست، استفاده کرد. هنگام مقایسه دو تعداد از خوشه‌ها، g و g' ممکن است یک مدل آمیخته با تعداد پارامترهای زیاد مناسب به نظر رسد ولی از دیدگاه ساده، منطقی است اگر برای $g > g'$، $C(g) < C(g')$ باشد g' بر g ترجیح داده شود. بنابراین برای انتخاب $g > 1$ در برابر $g = 1$ لازم است $C(g) > C(1) = L(1)$. در صورتی که $C(g) > L(1)$، تمام جزءهای معادله (۵) نامنفی خواهند بود که در این صورت $0 \leq NEC(g) \leq 1$، که تنها حالت $g > 1$ این شرط را فراهم می‌کند پس اگر $NEC(g) \leq 1$ نباشد دلیلی بر انتخاب بیش از یک خوشه وجود ندارد.

۴ مقایسه معیارها

در این بخش معیارهای NEC، BIC و AIC بر اساس تکنیک شبیه‌سازی مونت کارلو مورد مقایسه قرار می‌گیرند. برای این منظور نمونه‌هایی به حجم n از توزیع نرمال تک متغیره با واریانس‌های برابر یک و همچنین نرمال آمیخته دو متغیره با ماتریس‌های واریانس برابر I تولید

می‌شود.
برای توزیع نرمال تک متغیره چهار نوع توزیع با پارامترهای مختلف زیر در نظر گرفته شده است:
الف) توزیع نرمال استاندارد.
ب) توزیع نرمال آمیخته دو مولفه‌ای بامیانگین‌های $\mu_1 = ۰$ و $\mu_2 = ۲$ و نسبت‌های آمیختگی برابر.
ج) توزیع آمیخته نرمال دو مولفه‌ای با میانگین‌های $\mu_1 = ۰$ و $\mu_2 = ۲$ و نسبت‌های آمیختگی $\lambda_1 = ۰/۷$ و $\lambda_2 = ۰/۳$.
د) توزیع آمیخته نرمال سه مولفه‌ای با میانگین‌های $\mu_1 = ۰$، $\mu_2 = ۲$ و $\mu_3 = ۴$ و نسبت‌های آمیختگی مساوی.
برای حالت دو متغیره سه نوع توزیع زیر در نظر گرفته شده است.
الف) توزیع نرمال دو متغیره استاندارد.
ب) توزیع آمیخته نرمال دو مولفه‌ای با بردار میانگین $\mu'_1 = (۰, ۰)$ و $\mu'_2 = (۲, ۲)$ و نسبت‌های مساوی.
ج) توزیع آمیخته نرمال سه مولفه‌ای با بردار میانگین $\mu'_1 = (۰, ۰)$، $\mu'_2 = (۲, ۲)$ و $\mu'_3 = (۲, -۲)$ و نسبت‌های مساوی.
از هر یک از توزیع‌ها دو نمونه با حجم‌های $n = ۵۰$ و $n = ۲۰۰$ هرکدام به تعداد ۱۰۰ بار شبیه‌سازی شده است. بر اساس نمونه‌های تولید شده، پارامترهای مدلهای آمیخته بر اساس الگوریتم EM برآورد شده‌اند. در جدول ۱ نتایج مربوط به میانگین و انحراف معیار آنتروپی NEC بیان شده است که در آن d نشان دهنده بعد فضای نمونه‌ای و g نشان دهنده تعداد خوشه‌هاست. بر اساس نتایج به دست آمده NEC برای $n = ۲۰۰$ بهتر از $n = ۵۰$ عمل می‌کند. به عنوان مثال در جدول ۱ برای حالت (الف) توزیع آمیخته نرمال تک متغیره میانگین آنتروپی برای $g = ۱$ کمتر از بقیه حالت‌هاست لذا فرض تک خوشه‌بودن پذیرفته می‌شود. همچنین برای حالت (ب) توزیع دو متغیره وقتی که $n = ۲۰۰$ است، میانگین معیار آنتروپی برای $g = ۲$ برابر ۶/۲۴ است که کمتر از مقدار آن برای تعداد خوشه‌های $g = ۱$، $g = ۳$ و $g = ۴$ است و فرض دو مولفه‌ای بودن را تایید می‌کند. جدول ۲ درصد فراوانی انتخاب توزیع آمیخته g مولفه‌ای بر اساس استفاده از معیارهای AIC، BIC و NEC برای حالتهای مختلف مذکور و برای تعداد خوشه‌های مختلف بیان می‌کند.

به عنوان مثال برای حالت اول توزیع نرمال تک متغیره در ۸۵ درصد اوقات NEC جواب صحیح داده است در حالیکه در ۷۰ درصد اوقات AIC پاسخ صحیح می‌دهد. در حالت کلی بر اساس جدول ۲ نتیجه می‌شود که کیفیت تصمیم‌گیری NEC بین AIC و BIC قرار دارد. معیار AIC تعداد مولفه‌های توزیع آمیخته را اندکی بیش تخمین می‌کند و بر عکس BIC تعداد مولفه‌ها را اندکی کمتر تخمین می‌زند.

جدول ۱: میانگین و انحراف معیار برای معیار NEC بر اساس توزیع‌های آمیخته نرمال تک متغیره و دو متغیره

			تعداد خوشه‌ها			
d	n	پارامترهای توزیع	$g=1$	$g=2$	$g=3$	$g=4$
		$p_1=1$	۱۶٫۴۱	۱۷٫۹۳	۲۲٫۸۳	۲۹٫۳۲
		$\mu_1=0$	(۶٫۱۵)	(۷٫۹۷)	(۸٫۶۱)	(۱۰٫۳۵)
		$p_1=0.5$	۱۳٫۸۳	۱۰٫۹۸	۱۵٫۷۲۵	۲۲٫۱۰
۱	۲۰۰	$\mu_1=0, \mu_2=2$	(۵٫۸۴)	(۴٫۳۶)	(۵٫۴۹)	(۶٫۸۱)
		$p_1=0.7$	۱۵٫۹۸	۱۵٫۲۸	۲۲٫۶۷	۳۰٫۵۲
		$\mu_1=0, \mu_2=2$	(۶٫۱۲)	(۵٫۹۵)	(۷٫۳۱)	(۱۰٫۱۱)
		$p_1=\frac{1}{3}, p_2=p_3=\frac{2}{3}$	۲۰٫۴۵	۷٫۴۲	۵٫۸۹	۹٫۱۶
		$\mu_1=0, \mu_2=2, \mu_3=4$	(۱۲٫۸۸)	(۲٫۳۵)	(۰٫۹۷)	(۳٫۴۱)
		$p_1=1$	۱۷٫۹۲	۱۸٫۴۳	۲۳٫۳۵	۳۰٫۴۹
		$\mu_1=0$	(۷٫۲۱)	(۸٫۹۱)	(۱۰٫۷۲)	(۱۲٫۷۴)
		$p_1=0.5$	۱۴٫۲۷	۱۲٫۲۳	۱۷٫۱۸	۲۳٫۹۴
۱	۵۰	$\mu_1=0, \mu_2=2$	(۷٫۴۲)	(۵٫۸۲)	(۷٫۷۹)	(۶٫۹۱)
		$p_1=0.7$	۱۸٫۵۲	۱۶٫۷۶	۲۴٫۵۳	۳۲٫۸۶
		$\mu_1=0, \mu_2=2$	(۱۰٫۱۷)	(۷٫۶۴)	(۹٫۸۶)	(۱۳٫۲۲)
		$p_1=0, p_2=p_3=\frac{1}{2}$	۲۸٫۳۸	۵٫۸۲	۵٫۱۵	۷٫۵۵
		$\mu_1=0, \mu_2=2, \mu_3=4$	(۱۱٫۸۲)	(۲٫۱۰)	(۰٫۶۲)	(۲٫۸۴)
		$p_1=1$	۳٫۴۱	۱۶٫۸۴	۱۸٫۶۹	۲۴٫۴۷
		$\mu_1=[0,0]$	(۳٫۰۲)	(۶٫۱۲)	(۶٫۶۴)	(۷٫۷۶)
		$p_1=0.5$	۸٫۸۶	۶٫۲۴	۸٫۱۲	۱۰٫۴۲
۲	۲۰۰	$\mu_1=[0,0], \mu_2=[2,2]$	(۷٫۴۲)	(۵٫۸۲)	(۷٫۷۹)	(۶٫۹۱)
		$p_1=p_2=p_3=\frac{1}{3}$	۱۱٫۱۹	۳٫۵۹	۲٫۰۲	۲٫۹۰
		$\mu_1=[0,0], \mu_2=[2,2], \mu_3=[2,-2]$	(۱۱٫۸۲)	(۲٫۱۰)	(۰٫۶۲)	(۲٫۸۴)
		$p_1=1$	۹٫۶۷	۱۱٫۵۸	۲۳٫۹۶	۳۰٫۹۵
		$\mu_1=[0,0]$	(۴٫۱۵)	(۶٫۹۹)	(۷٫۴۴)	(۸٫۹۴)
		$p_1=0.5$	۱۰٫۸۱	۸٫۹۷	۱۰٫۳۲	۱۲٫۷۰
۲	۵۰	$\mu_1=[0,0], \mu_2=[2,2]$	(۶٫۶۸)	(۴٫۱۲)	(۵٫۸۳)	(۷٫۵۶)
		$p_1=p_2=p_3=\frac{1}{3}$	۲٫۹۲	۲٫۳۵	۴٫۴۳	۱۱٫۹۷
		$\mu_1=[0,0], \mu_2=[2,2], \mu_3=[2,-2]$	(۰٫۸۸)	(۲٫۱۰)	(۰٫۶۲)	(۲٫۸۴)

جدول ۲: درصد فراوانی انتخاب g خوشه بر اساس توزیع‌های آمیخته نرمال تک متغیره و دو متغیره

				معیارها		
d	n	پارامترهای توزیع	تعداد خوشه‌ها	AIC	BIC	NEC
		$p_۱ = ۱$ $\mu_۱ = ۰$	۱ ۲	۷۰ ۳۰	۹۵ ۵	۸۵ ۱۵
۱	۲۰۰	$p_۱ = ۰/۵$ $\mu_۱ = ۰, \mu_۲ = ۲$	۱ ۲ ۳	۱۰ ۸۰ ۱۰	۴۰ ۶۰ ۰	۱۵ ۸۰ ۵
		$p_۱ = ۰/۷$ $\mu_۱ = ۰, \mu_۲ = ۲$	۱ ۲ ۳	۲۵ ۶۵ ۱۰	۵۰ ۵۰ ۰	۲۵ ۷۰ ۵
		$p_۱ = p_۲ = p_۳ = \frac{۱}{۳}$ $\mu_۱ = ۰, \mu_۲ = ۲, \mu_۳ = ۴$	۱ 2 ۳ ۴	۰ ۵ ۸۰ ۱۵	۱۰ ۱۵ ۷۵ ۰	۰ ۱۰ ۸۰ ۱۰
		$p_۱ = ۱$ $\mu_۱ = ۰$	۱ ۲ ۳	۶۰ ۳۰ ۱۰	۸۵ ۱۵ ۰	۸۰ ۲۰ ۰
۱	۵۰	$p_۱ = ۰/۵$ $\mu_۱ = ۰, \mu_۲ = ۲$	۱ ۲ ۳	۲۵ ۶۰ ۱۵	۵۵ ۴۵ ۰	۳۵ ۵۵ ۱۰
		$p_۱ = ۰/۷$ $\mu_۱ = ۰, \mu_۲ = ۲$	۱ ۲ ۳ ۴	۰ ۱۵ ۶۵ ۲۰	۲۰ ۲۰ ۵۰ ۰	۰ ۲۰ ۷۰ ۱۰
		$p_۱ = p_۲ = p_۳ = \frac{۱}{۳}$ $\mu_۱ = ۰, \mu_۲ = ۲, \mu_۳ = ۴$	۱ ۲ ۳ ۴	۰ ۱۵ ۶۵ ۲۰	۲۰ ۲۰ ۵۰ ۰	۰ ۲۰ ۷۰ ۱۰
		$p_۱ = ۱$ $\mu_۱ = [۰, ۰]$	۱ ۲	۸۰ ۲۰	۱۰۰ ۰	۹۵ ۵
۲	۲۰۰	$p_۱ = ۰/۵$ $\mu_۱ = [۰, ۰], \mu_۲ = [۲, ۲]$	۱ ۲ ۳	۵ ۸۰ ۱۵	۳۵ ۶۰ ۵	۱۰ ۸۰ ۱۰
		$p_۱ = p_۲ = p_۳ = \frac{۱}{۳}$ $\mu_۱ = [۰, ۰], \mu_۲ = [۲, ۲], \mu_۳ = [۲, -۲]$	۱ ۲ ۳ ۴	۰ ۵ ۷۵ ۲۰	۵ ۲۵ ۷۰ ۰	۰ ۵ ۷۵ ۲۰
		$p_۱ = ۱$ $\mu_۱ = [۰, ۰]$	۱ ۲	۶۰ ۴۰	۸۰ ۲۰	۸۰ ۲۰
۲	۵۰	$p_۱ = ۰/۵$ $\mu_۱ = [۰, ۰], \mu_۲ = [۲, ۲]$	۱ ۲ ۳	۱۰ ۶۵ ۲۵	۶۰ ۴۰ ۰	۱۰ ۷۰ ۲۰
		$p_۱ = p_۲ = p_۳ = \frac{۱}{۳}$ $\mu_۱ = [۰, ۰], \mu_۲ = [۲, ۲], \mu_۳ = [۲, -۲]$	۱ ۲ ۳ ۳	۰ ۲۰ ۵۵ ۲۵	۲۵ ۳۰ ۴۵ ۰	۰ ۱۰ ۶۰ ۳۰

مراجع

[1] Biernacki, C., Celeux, G. and Govert, G. (1999), An Improvement of the NEC Criterion for Assessing the Number of Clusters in a Mixture model, Pattern Recognition Letters, 20, 267-272.

[2] Celeux, G. and Soromenho, G. (1996), An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model. Journal of Classification , 13, 195-212.

[3] Fraly, C. and Raftery, A. E. (1998), How Many Clusters? Which Clustering Method? Answer via Model-Based Cluster Analysis. Techincal Report, No, 329. Seattle: Department of Statistics, University of Washington.

[4] Fraley, C. and Raftery, A. E. (1999), MCLUST: Software for Model-Based Cluster Analysis, J. Classification, 16, 297-306.

[5] Hartigan, J. A. (1975), Clustering Algorithms. Wiley, New York.

[6] Jeff C. F., (1983), On the Convergence of the EM Algorithm, Annals of Statistcs, 11, 95-103.

[7] McLachlan, G. J. (1982), The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis. In Krishnan, P. R. and Kanal, L. N. (eds), Handbook of Statistics, 2, 199-208. North-Holland, Amesterdam.

[8] McLachlan, G. J. and Krishnan, T. (1997), The EM Algorithm and Extensions. Wiley, New York.

[۹] محمدقاسم وحیدی اصل، محسن محمدزاده و محمد قربانی (۱۳۸۱)، تعیین مدل خوشه‌بندی احتمالاتی بر اساس معیار اطلاع بیزی، مجموعه مقالات ششمین کنفرانس بین‌المللی آمار ایران، ص ۴۶۵-۴۵۵.

شناخت اشیاء (شکل و ساختار) با استفاده از آمار

هادی گنجی[۱]، محمود صفارزاده[۲]

[۱] فوق لیسانس آمار اقتصادی – اجتماعی، کارشناس ارشد پژوهشکده حمل و نقل
[۲] دکتری برنامه‌ریزی حمل و نقل، معاون پژوهشی پژوهشکده حمل و نقل

چکیده:

۱ مقدمه

در فضای وسیعی از اشیاء منتظم، اندازه‌گیری، تحلیل و مقایسه شکل اشیاء با یکدیگر از اهمیت بالایی برخوردار است [1]. لیدز[1] از جمله اولین افرادی است که از آنالیز شکل برای مشخص کردن قضیه محل مرکزی در جغرافیا آن را به کار برد [1]. ماردیا[2] در سال (۱۹۷۷) توزیع مربوط به تولید مثلث‌ها که به وسیله تعداد مشخصی نقطه تولید می‌شوند را استخراج کرد. در یک نگاه تخصصی‌تر پاسخ برای این سؤال که آیا روستاها در (شهرها) یک منطقه به صورت یک قاعدهٔ مشخص و با فاصله مساوی از روستاهای (شهرهای) اطراف ایجاد شده‌اند [2]، از موضوعات مطرح شده در آنالیز شکل می‌باشد.

تلفیق آنالیز شکل با آمار در سال ۱۹۸۶ به وسیله اولینگس[3] و جانسون[4] بوده است که در مقاله مربوطه اسکلت موش‌ها از نظر شکل ساختاری، اندازه مورد تجزیه و تحلیل قرار گرفته است [3]. در این مقاله در آن هستیم که ضمن معرفی آنالیز و کاربردهای آن، کاربرد آمار را در آنالیز شیء مورد بررسی قرار داده و تکنیک‌های مختلف آماری که در این موضوع کاربرد دارد، معرفی شود.

۲ آنالیز شکل

اشیاء در همه جا هستند چه انسان آن را به وجود آورده باشد و یا اینکه محصول طبیعت باشد. امروزه دانشمندان به دنبال این موضوع هستند که با جمع‌آوری اطلاعات هندسی این اشیاء به مطالعه خصوصیت شکل آنها بپردازند. در حقیقت این موضوع با جمع‌آوری اطلاعات مربوط به نقاط موجود در شیء انجام می‌گیرد. با این توضیح اولیه مشخص شد که جامعه‌ای از نقاط که شیء را تعریف کند، وجود دارد.

با روش‌های مختلف اطلاعات مربوط به برخی از نقاط راکه دارای اهمیت می‌باشند جمع‌آوری کرده و بعد از آن به تجزیه و تحلیل پرداخته می‌شود. در حقیقت یک کار آماری است که شامل

1) Leeds 2) Mardia 3) O'ltiggins 4) Jonson

روش نمونه‌گیری، استنباط آماری راجع به پارامتر آن و نتیجه گیری است. جهت بررسی این موضوع به تعاریف اولیه مورد نیاز، پرداخته می‌شود.

شکل: به کلیه اطلاعات هندسی که از یک شیء بعد از حذف اثرات اندازه، دوران و محل قرارگیری باقی می‌ماند اطلاق می‌گردد [4]. از این تعریف اینگونه استنتاج می‌گردد که شکل اشیاء نسبت به تغییرات هندسی از قبیل تغییر اندازه یال ها یا دوران ثابت باقی می‌ماند. در حقیقت در آنالیز آماری شکل با داشتن اطلاعات مربوط به تعدادی از زوایا و یال‌ها، کل شکل برآورد می‌گردد. البته امروزه با پیشرفت صنایع عکس‌برداری و ثبت اطلاعات لازمه، موضوع از این نیز فراتر رفته و دیگر خصوصیت اشیاء، از قبیل جنس تشکیل دهنده نیز مورد بررسی قرار می‌گیرد.

از این تعریف اینگونه استنباط می‌شود که دو شکل یکسان هستند، اگر با تغییر اندازه و یا دوران یکی را به دیگری بتوان تبدیل کرد. در عمل نیز می‌توان با این ویژگی اشکال گوناگون اشیاء را با یکدیگر مقایسه کرد.

اشکال به وسیله تعدادی محدودی از نقاط مشخص می‌شود که به این نقاط مشخص نقطه علامت[5] گفته می‌شود و ویژگی آن این است که مرز بین نقاط بیرونی و درونی جامعه می‌باشد. به طور مثال در آناتومی جمجمه گوشه چشم را می‌توان به عنوان یک نقطه علامت در نظر گرفت. این نقاط معمولاً به وسیله متخصصین رشته مربوطه مشخص می‌گردد. با استفاده از اطلاعات این نقاط است که استنباط بر روی شکل صورت می گیرد. دو روش جهت تجزیه و تحلیل این اطلاعات وجود دارد: ۱ـ تجربی، ۲ـ هندسی.

۱.۲ روش تجربی

در روش تجربی می‌توان نسبت بین فاصله نقاط علامت یا زوایا را مشخص کرده و این داده‌ها را با استفاده از آنالیز چند متغیره تحلیل کرد که به طور خاص در بیولوژی به آن چند متغیره morphometries گفته می‌شود که ریمنت[6] (۱۹۸۴) به آن پرداخته است. درکل در این روش به تفاوت در فاصله بین طول‌ها یا پهناها در میان نقاط علامت پرداخته می‌شود. به طور مثال در بررسی شباهت بین جمجمه‌ها پاره خط‌های ایجاد شده میان نقاط علامت به وسیله آنالیز چند متغیره طبقه‌بندی می‌شود. مثلاً طبقه‌بندی قسمت‌های مختلف جمجمه با توجه به جنسیت که از اطلاعات نقاط علامت بدست آمده شامل می‌شود.

یک ابزار بسیار مفید در این مورد تحلیل عاملی است که جنبه‌های مختلف اندازه یا خصوصیات شکل را بیان می‌کند. معمولاً عامل اول در برگیرنده اطلاعات مربوط به کلیه متغیرها (اطلاعات مربوط به نقاط علامت) بوده و اندازهٔ کلی شکل از آن استخراج می‌گردد [4].

دستهٔ دیگری از مطالعات تحلیل تفاوت بین اشکال با اندازه‌های یکسان است. این کار معمولاً با برازش یک معادله غیر خطی بر روی طول نقاط علامت همراه است. البته از اطلاعات مربوط به زوایا، نسبت بین طول‌ها و یا استفاده از مقادیر واقعی بر اساس محورهای مختصات نیز استفاده می‌گردد. هنوز هم از روش‌های چند متغیره در بیولوژی و شکل‌شناسی استفاده بسیاری

5) Lamdmark 6) Reyment

می‌گردد که می‌توان به کار برد در طبقه‌بندی شکل اشاره کرد، همچنین باید ذکر کرد که برقراری رابطه خطی بین نسبت اندازه طول‌ها و زوایا کار مشکلی است.

۲.۲ روش‌های هندسی

در دو دهه اخیر تعریف دقیق‌تری از نقاط علامت ارائه شده است که این خود کمک کننده در توسعه آنالیز شکل می‌باشد. البته با گسترش تکنولوژی این کار به دیجیتالی شدن اشیاء منجر شده است. اگر هیچ پیش شرطی بر روی این نقاط علامت وجود نداشته باشد روش کار همان استفاده از آنالیز چند متغیره می‌باشد که با توجه به اینکه مختصات این نقاط اقلیدسی نبوده، این کار مشکل می‌باشد [5].

ایده اصلی در روش هندسی بدین صورت می‌باشد که بجای استفاده از مقادیر عددی (جبر) به اطلاعات هندسی اشیاء پرداخته می‌شود که این کار معمولاً با تبدیل قسمتی از شکل همراه است. در این روش نیز هدف اصلی آنالیز شکل همان استنباط در رابطه با تفاوت اشکال اشیاء است. به طور مثال بیولوژیک با استفاده از روش‌های هندسی در ترمیم و یا تشریح تغییرات شکل، یک راهبرد اساسی است. همچنین این روش در باستان شناسی، ستاره شناسی و جغرافی کاربرد زیادی دارد.

۳ کاربردهای اختصاصی آنالیز شکل [15,11]

جهت مشخص شدن کاربردهای وسیع این موضوع به تعدادی از موارد انجام شده اشاره می‌شود. به طور مثال در پاسخ به این سؤال که چه تغییراتی در شکل به هنگام رشد رخ می‌دهد و یا اینکه شکل اشیاء چگونه به اندازه آنها مرتبط می‌شود یا چه تأثیری بیماری بر روی شکل شیء مورد نظر دارد. ارتباط شکل با دیگر متغیرها از قبیل سن، جنسیت و یا شرایط محیطی چگونه می‌باشد. چگونه می‌توان شکل مورد نظر را طبقه‌بندی و شاخص‌گذاری کرد و در آخر این که چگونه می‌توان تغییرپذیری شکل را توضیح داد. در تمام موارد اشاره شده، روش‌های آنالیز چند متغیره پاسخگو می‌باشند اما یکی از مشکلات اساسی در مورد مطالعه اینگونه موارد حجم کوچک نمونه با تعداد متغیر (نقاط علامت) زیاد می‌باشد. در ذیل به پاره‌ای از این کاربردها به صورت طرح مسأله اشاره می‌شود.

۱.۳ بیولوژی

در یک آزمایش برای مشخص کردن تأثیر وزن بر روی اسکلت و پیدا کردن توضیح برای این تفاوت، موش‌ها را به سه گروه مجزا تقسیم می‌کنند. گروه کنترل، کوچک و بزرگ. گروه کنترل هیچ مشخصه خاصی ندارند ولی در گروه بزرگ موش‌ها با اندام‌های بزرگ و وزن بالا و در گروه کوچک

بالعکس انتخاب شدند. نتیجه آزمایش مشخص کننده این موضوع می‌باشد که این تأثیر چگونه است.

۲.۳ داروسازی

شناسایی تفاوت شکلی در مغز انسان‌ها در دو گروه که اول شامل افراد سالم و گروه دوم شامل افراد با بیماری اسکیزوفرنی تفاوتی در شکل وجود دارد.

۳.۳ کشاورزی [6]

تعیین و برآورد نحوه توزیع ماهیان در زیر آب و دریافت اطلاعات فیزیکی مربوط به آنها.

۴.۳ جغرافیا [6]

قضیهٔ جغرافیای مرکزی به این موضوع اشاره دارد که شهرها در یک شش ضلعی منتظم در یک منطقه هموار توزیع می‌شوند. ایده اولیه این موضوع از ۴۴ نقشه مربوط به ۶ کشور به دست آمده است. اگر درستی این قضیه فرض گردد، می‌توان این آزمون فرض را انجام داد که مثلث‌های ساخته شده تحت شهرهای همسایه متساوی الاضلاع هستند یا خیر؟

تمام موارد اشاره شده با استفاده از عکس‌برداری، ثبت اطلاعات نقاط علامت و استفاده از آنالیز چند متغیره مورد بررسی قرار گرفته‌اند. با توجه به اشاره به برخی از کاربردهای آنالیز شکل در علوم مختلف به یک مورد از کاربردهای آن با آنالیز تصویر پرداخته می‌شود.

۴ آنالیز تصویر

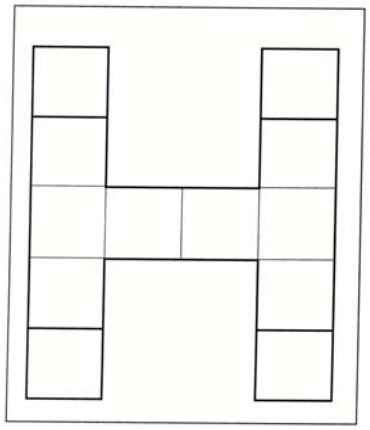
آنالیز تصویر معمولاً با سنجش از راه دور همراه است. سنجش از دور به عنوان تکنولوژی و علمی تعریف می‌شود که به وسیله آن می‌توان با گرفتن تصویر از اشیاء یا پدیده‌ها آن‌ها را تعیین، اندازه‌گیری یا تجزیه و تحلیل نمود. منبع داده‌های سنجش از دور معمولاً تابش الکترومغناطیسی است که از یک پدیده یا شیء، بازتاب یا گسیل می‌شود. کاربردهای سنجش از دور شامل مواردی همچون [7]:
۱ـ طبقه‌بندی پوشش زمین ۲ـ آشکارسازی تغییرات پوشش زمین ۳ـ پویش و بررسی تصویری کیفیت آبها ۴ـ اندازه‌گیری دمای سطح دریاها ۵ـ بررسی و مطالعه برف ۶ـ پویش و بررسی تصویری اجزای جوی ۷ـ آشکارسازی خط واره‌ها ۸ـ تعبیر زمین‌شناسی ۹ـ اندازه‌گیری ارتفاعی (تولید DEM)

جهت مطالعه یک تصویر آن را به گروهی از سلول‌ها (المان‌های) مساوی تقسیم کرده و به هر کدام از این سلول‌ها (المان‌ها) یک پیکسل گفته می‌شود. در حقیقت با انجام این کار یک

ماتریس $r \times c$ تولید می‌شود. مقادیر مربوط به عناصر این ماتریس، روشنایی رنگ خاکستری بازتاب شده از هر سلول می‌باشد که با اشعهٔ X به آن تابش شده است، این مقدار به موقعیت این پیکسل ها بستگی دارد. دامنهٔ تغییرات این رنگ خاکستری از سطح صفر (مشکی) تا ۲۵۵ (سفید) است. بنابراین یک پیکسلی که با عدد ۵۰ مشخص می‌شود خاکستری تیره و ۱۲۸ خاکستری متوسط و ۲۰۰ خاکستری روشن را در عمل از خود نشان می‌دهد.

سایز معمول تصاویر $(r = ۲۵۶) \times (c = ۲۵۶)$ یا $(r = ۵۱۲) \times (c = ۵۱۲)$ است. دلیل آن هم $۲^۹ = ۵۱۲$ و $۲^۸ = ۲۵۶$ است (یعنی در نظرگرفتن ۸ سطح از رنگ خاکستری یا ۹ سطح از آن) که با بیشتر شدن این سطوح کار با آنها بسیار مشکل است. برای تصاویر رنگی نیز به همین منوال می‌باشد، بدین ترتیب که تصاویر رنگی در سه گروه از خاکستری نمایش داده می‌شود. برای هر باند آبی، قرمز و سبز جداگانه که یک تصویر ۲۵۶×۲۵۶ با $۲^{۱۶} = ۶۵۵۳۶$ پیکسل نمایش داده می‌شود. چند متغیره بسیار بزرگ که بعضی از آنها نیز غیراقلیدسی می‌باشد.

به طور مثال تصویر حرف H و نمایش ماتریس مقادیری آن به صورت زیر است:



$$\Downarrow$$

$$\begin{pmatrix} . & . & . & . & . & . \\ . & ۲۱۱ & ۲ & ۱ & ۲۴۳ & . \\ . & ۲۰۰ & ۵ & ۸ & ۲۵۱ & . \\ . & ۲۱۰ & ۲۴۱ & ۲۵۱ & ۲۵۴ & . \\ . & ۲۳۶ & ۲ & ۱۲ & ۲۰۴ & . \\ . & ۲۵۱ & ۷ & ۱۲ & ۲۱۸ & . \\ . & . & . & . & . & . \end{pmatrix}$$

بعد از معرفی اولیه مربوط به آنالیز تصویر نحوهٔ نمونه‌گیری مشخص کردن جامعه، بحث بر روی برآورد مقادیر مربوطه می‌باشد که از آن طریق بتوان طبقهٔ مورد نظر شکل و یا ظاهر آن را بتوان

تخمین زد. در این برآورد دو روش کلاسیک و بیز در استنباط آماری وجود دارد. در حقیقت با این روش می‌توان پیش‌بینی کرد که خصوصیت شکل و یا شکل ظاهری آن به چه ترتیبی می‌باشد.

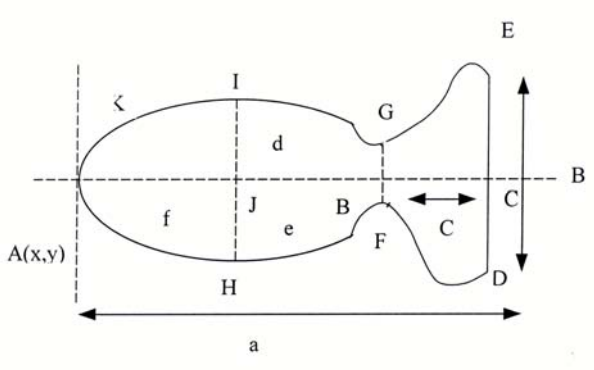
تجزیه و تحلیل تصاویر از روش آماری بیز تقریباً از سال ۱۹۸۰ آغاز گردیده است. برای استفاده از این روش احتیاج به یک مدل پیشین می‌باشد که نشان دهنده اطلاعات اولیه از اشیاء در موقعیت خاص است. همچنین احتیاج به مدل توأم از سطوح مختلف رنگ خاکستری در تصویر است که به محل اشیاء نیز بستگی دارد. به طور مثال در یک تصویر پزشکی می‌توانیم این اطلاعات اولیه را داشته باشیم که تصویر گرفته شده از مغز یا قلب می‌باشد. اطلاعات اولیه می‌تواند در قالب مدل اولیه S_0 ارائه شود. توجه شود که S_0 می‌تواند شامل اطلاعات مربوط به یک شی یا کلیه اشیاء موجود در تصویر باشد. تابع توزیعی اولیه‌ای برای پارامترهای جامعه (شکل) در نظر گرفته می‌شود و با توزیع $\pi(s)$ که کلیه تغییرات S_0 را نشان می‌دهد، مشخص می‌گردد.

بنابراین S تابعی از پارامترهای $\theta_1 \dots \theta_p$ است. مدل کل تصویر نیز مورد نیاز می‌باشد. تصویر I را با ماتریس x_i $i = (i_1, i_2) e \{1 \dots r\} \times \{1 \dots c\}$ فرض کنید [1] تابع توزیعی توأم از سطوح مختلف خاکستری که از شیء S مشخص می‌شود و به صورت $L(I/S)$ نمایش داده می‌شود. طبق تعریف بیز توزیع پسین شیء موجود در تصویر با اطلاعات حاصل از تصویر به شکل زیر خواهد بود: $\pi(S/I) \alpha L(I/S) \pi(S)$ که در اینجا از روش‌های خاص برآورد از قبیل ماکزیمم کردن و یا شبیه‌سازی استفاده می‌گردد. دو روش نیز جهت برآورد مدل پیشین ۱ـ هندسی ۲ـ آماری وجود دارد.

۱.۴ روش هندسی

در روش هندسی مشخصات هندسی شیء از قبیل خطوط، دایره‌ها، بیضی‌ها و قوس‌ها و . . . مورد توجه قرار می‌گیرد. مدل‌های به کار رفته در این روش معمولاً دارای دو پارامتر اساسی است. ۱ـ محل ۲ـ کل شیء مورد نظر. پارامتر کل، بزرگی شیء را نشان می‌دهد و پارامتر محل اطلاعات ریزتری را بیان می‌دارد. برای مشخص شدن موضوع به ارائه یک مثال می‌پردازیم.

به طور مثال شکل یک ماهی را در نظر بگیرید. روش هندسی جهت تدوین توزیع پیشین را می‌توان به صورت زیر در نظر گرفت. شکل را به دو سهمی و یک مثلث به گونه‌ای که طول ماهی a و طول دم آن b است، تقسیم می‌کنیم.



مقادیر a, b, c, d, e و f مشخص کننده اندازه و شکل ماهی است. فرض کنید که مختصات $A(۰, ۰)$ باشد. بنابراین مختصات نقطه $c(a, ۰)$ مشخص می‌گردد. شرط دیگر اینکه $GF < ED$ و نقاط B و C وسط CF و ED قرار دارند و $GFDE$ ذوزنقه است. بنابراین E و D دارای مختصات $\left(a, \frac{b}{۲}\right), \left(a, -\frac{b}{۲}\right)$ می‌باشد.

معادله سهمی AIG به وسیله $I(f, d)$ مشخص می‌گردد به طوری که معادله AHF از نقطه $H(f, -e), (۰, ۰)$ می‌گذرد. در نتیجه مختصات $B(a-c, ۰)$ می‌باشد که مختصات عرض نقاط F و G از این سهمی‌ها مشخص می‌گردد. به علاوه مختصات J نقطه تلاقی IH و AC است. در نتیجه:

$$A = (۰, ۰), B = (a-c, ۰), C = (a, ۰), D = (a, -\frac{b}{۲}), E = (a, \frac{b}{۲})$$
$$F = \left(a-c, -e + k_۲(a-c-f)^۲\right), G = \left(a-c, d-k_۱, (a-c-f)^۲\right)$$
$$H = (f, -e), I = (f, d)$$

و مقادیر $k_۱ = \frac{d}{f}$ و $k_۲ = \frac{e}{f^۲}$ است. همچنین هر نقطه‌ای مثل $x = k$ بر روی سهمی AIG دارای مختصات $\left(a-k, d-k_۱ (a-k-f)^۲\right)$ است. برای مثال برای یک تصویر ۱۲۰ × ۱۲۰ مقادیر قابل قبول برای میانگین و میانگین شرطی به صورت زیر است:

$$E(a) = ۱۱۰, \quad E(d/a) = \frac{a}{۹}, \quad E(b/d) = ۲d$$
$$E(e/d) = \frac{۲d}{۳}, \quad E(c/a) = \frac{a}{۶}, \quad E(f/a) = \frac{۲a}{۱۶}$$

تغییر شکل ماهی می‌تواند با تغییر مختصات نقطه A و طول a تغییر کند که دیگر نقاط نیز با توجه به این دو مقدار تغییراتی خواهند داشت. برای مثال می توان توابع توزیع زیر را در نظر گرفت:

$$a \sim N(۱۱۰, ۲۰), \quad d/a \sim N\left(\frac{a}{۹}, ۳\right), \quad b/d \sim N(۲d, ۱)$$

$$e/d \sim N\left(\frac{۲d}{۳}, ۳\right), \qquad \frac{c}{a} \sim N\left(\frac{a}{۶}, ۱\right)$$

به این ترتیب $\frac{f}{a} \sim N\left(\frac{۷a}{۱۶}, ۰\right)$ است که f دارای مقدار ثابت $\frac{۷a}{۱۶}$ است اگر a مشخص باشد.

۲.۴ روش غیرهندسی

فرض کنید که مختصات نقطه $A(x,y)$ یک متغیر پیوسته یکنواخت بر روی تصویر $N \times N$ باشد. تابع توزیع توأم (x,y) و (a,b,c,d,e) از معادله ۱ به صورت زیر در نظرگرفته می‌شود.

$$\pi(x,y,a,b,c,d,e) = \phi_۱(a)\,\phi_۲(d/a)\,\phi_۳(b/d)\,\phi_۴(e/d)\,\phi_۵(c/a).\frac{۱}{N^۲}$$

اگر $۰ < x,y < N$ و $J = ۱,\ldots,۵$ ، ϕ_j تابع توزیع نرمال با میانگین و واریانس معادلات ۱ باشد. روش ساده دیگر این است که برای $\theta_۱ = x$, $\theta_۲ = y$, $\theta_۳ = a$ توزیع پیشین زیر در نظرگرفته شود.

$$n(\theta_۱,\theta_۲,\theta_۳) = \begin{cases} \exp\left\{-\frac{۱}{۴۰}(\theta_۳ - ۱۱۰)^۲\right\} & ۰ < \theta_۱,\theta_۲ < N \\ ۰ & o.w \end{cases}$$

هدف از دو روش فوق نحوهٔ استفاده از پارامترهای هندسی در برآورد مدل پیشین می‌باشد که می‌توان برای تمام اشیاء تعمیم داد.

مثال: در انتها بعد از ارائه تعاریف اولیه و مفهوم کلی آنالیز شکل و تعیین نقش اساسی آمار در این موضوع مثال واقعی که مورد بررسی قرارگرفته است ارائه می‌شود. البته لازم به ذکر است که به دست آوردن یک چنین داده‌هایی از داخل ایران بسیار مشکل می‌باشد. هدف این مقاله نیز ارائه یکی دیگر از کاربردهای وسیع آمار می‌باشد.

در طبقه‌بندی سطح زمین در ناحیه نیوجرسی که یک منطقهٔ مشخص خاص آن مد نظر می‌باشد، عکسی از آنجا تهیه شده و مورد بررسی قرارگرفته است. این ناحیه شامل بخش مزرعه، جنگل، آب و منطقهٔ شهرنشینی می‌باشد [9,11]. در ابتدا باید به این موضوع اشاره داشت که در نقاط مرزی بین این چهار طبقه است که معمولاً تشخیص و برآورد آنها مفهوم پیدا می‌کند و از اهمیت بیشتری برخوردار است. می‌توان مدل‌های پیش‌بینی تجربی را برای این نقاط مشخص کرد که معمولاً مدل پیشین مناسب در این موارد توزیع Gibbs می‌باشد که شامل دو پارامتر است.

$$\overline{n}(S) = \frac{۱}{۲(\beta)}\exp\{-\beta V(x)\} \qquad \beta > ۰$$

به طوری که $Z(\beta)$ ثابت پایدار[۷] است.

$$Z(\beta) = \int \exp\{-\beta V(x)\}\, dx$$

$V(x)$ همان سطوح مختلف رنگ خاکستری انعکاس شده از پیکسل‌ها است. با توجه به بحث همسایگی در پیکسل‌های در یک شکل و در انواع مختلف تابع توزیع‌های پیشین دیگری نیز مطرح می‌باشد به طور مثال تابع log-Cauchy است. به صورت $\pi(s) = \left(\frac{u^{۲}}{۱+u^{۲}}\right)$ و یا پیشنهاد دیگری مبنی بر استفاده از توزیع log-Cosh که دارای فرمت کلی $\overline{n}(S) = c_{۱} \log\cosh(c_{۲}, \mu)$ می‌باشد.

با مشخص کردن تابع توزیع‌های پیشین و استفاده از الگوریتم EM در برآورد پارامترهای مدل، نقاط نامشخص تصویر و یا به عبارت بهتر با استفاده از قانون ماکزیمم احتمال پیکسل‌ها را به یکی از چهار طبقه اختصاص داده می‌شود.

البته در تشخیص این طبقه‌بندی می‌توان به یک سری از اطلاعات اولیه موجود در خصوص منطقه نیز اشاره داشت و آنها را کلاسه‌بندی کرد:

۱ــ پیکسل رودخانه از نظر روشنایی رنگ خاکستری، کمترین شباهت را به پیشکل مزرعه دارد. همین مورد برای پیکسل مربوط به جاده نیز صادق است.

۲ــ در صورت وجود شیب در یک جاده، پیکسل مربوطه شبیه به مزرعه نمی‌باشد.

۳ــ پیکسل‌های جنوب بیشتر شبیه به جنگل می‌باشند.

۴ــ پیکسل‌های جاده و آب بسیار شبیه به هم هستند.

۵ــ پیکسل‌های جاده و مناطق شهرنشینی کمترین شباهت را به هم دارا می‌باشند.

۶ــ پیکسل‌های جاده و جنگل بسیار شبیه به هم می‌باشند.

(منظور از شبیه به هم این است که سطوح خاکستری منعکس شده از این مناطق تقریباً نزدیک به هم می‌باشند). مورد بالای اشاره شده در ۶ بند در حقیقت ترکیب تجزیه و انتخاب مدل‌های پیشین است.

توزیع‌ها و موارد اشاره شده در حقیقت به صورت شرطی است که هرکدام برآورد پارامترها را در آن سطح طبقه در بردارد. علاوه برآن فرض می‌شود که اشعه‌های بازتاب از شکل از فرآیند پواسن پیروی می‌کند.

$$Y_t = P\left(\sum_{S} a_{ts} x_s\right)$$

که a_{ts} اثر پیکسل‌های همسایه بر مقدار بازتاب‌ها در S طبقه مشخص می‌باشد. با توجه به اطلاعات داده شده و در نظر گرفتن اطلاعات توزیع پیشین، جداول زیر حاصل می‌گردد که:

همانطور که قبلاً نیز بدان اشاره شد، هدف این مقاله انشعاب و یا ارائه روش آماری جهت برآورد نبوده، بلکه مشخص کردن یکی از روش‌های استنباط آماری در یک کار عملی می‌باشد.

7) Normalizing Constant

جدول ۱: خلاصه اندازه نیکویی برازش

			RMSE			MAD		
			Min	Mean	Max	Min	Mean	Max
مزرعه	C_0	Log-Cauchy	۷٫۳	۷٫۶	۷٫۸	۲٫۷	۲٫۸	۲٫۹
		Log Cosh	۷٫۸	۸	۸٫۲	۲٫۹	۲٫۹	۳
جنگل	C_1	Log-Cauchy	۶٫۵	۷٫۱	۸٫۵	۲٫۴	۲٫۶	۲٫۷
		Log Cosh	۶٫۷	۷	۸	۲٫۵	۲٫۷	۲٫۹
آب	C_3	Log-Cauchy	۵٫۸	۶	۶٫۳	$\frac{1}{2}$	۲٫۲	۲٫۳
		Log Cosh	۶٫۳	۶٫۷	۷٫۲	۲٫۲	۲٫۴	۲٫۴
شهرنشینی	C_2	Log-Cauchy	۶٫۱	۶٫۳	۶٫۵	۲٫۱	۲٫۲	۲٫۳
		Log Cosh	۶٫۲	۶٫۴	۶٫۶	۲٫۳	۲٫۴	۲٫۵

Rsme = میانگین توان دوم خطاها

MAD = قدر مطلق توان دوم خط

همانطورکه از جدول می‌توان استنباط کرد، با توجه به استفاده از مدل‌های مختلف توزیع پیشین و برآورد منطقهٔ مورد نظر به طور مثال مشخص کردن منطقهٔ مزرعه با توجه به دو تابع توزیع پیشین به ترتیب مقادیر خطاها ۶، ۷ و یا ۸ می‌باشد و به همین ترتیب برای سایر مناطق، که این می‌تواند روشی جهت پیش‌بینی نقاط غیرمشخص در تصویر باشد.

مراجع

[1] LANL. DRYDEN and K.V (1998), "STATISTICAL SHAPE ANALYSIS", WILEY

[2] Goodall, C.R and Lange, N, (1982), "Growth Cure models for Correlated triangular shapes". I_n proceeding of the zist INTERFACE Symposium.

[3] Kendell D.G. (1983), "The shape of poisson-Delaunay triangles" studies in probability and Related topics.

[4] Mardia, K.V. (1989b) , "Markor models and Bayesian methods in image Amalysis", Journal of Appled statistics, 10: 125-130.

[5] H.-L. (1991a) "On geodesics in Eoclideam shap space". Journal of the London Mathematical Society, 44:360-372.

[6] Mardia, K.V. and Dryden, I.L. (1989b) "The statistical analysis of shap data", Biometrika, 76: 71-282.

[7] Mardia, K.V. and Kanji, G.K. editors (1993), "Statisties and Images", Vol 1, oxford. Carfax.

[8] Mardia, K.V. , Kent, J.T., and Walder, A.N. (1991). "Statistical shape models in Image analysis" Computer science and statistics.

[9] Green, P. (1990), "Bayesian Reconstructions from Emission Tomography Data using a Modified EM Algorithm". IEEE Transactions on Medical Imaging.

[10] H/tchcock, D., Glasbey, A. (1997). "Binary Image Restriction at subpixel Resolution" Biometrics, 1040-1053.

[11] Frigessi, A. and stamder, J. (1994), "Informative Priors for the Buysian classification of satellite images". JASA, Vol. 89. No. 426

مقایسه روش‌های مختلف نمونه‌گیری با استفاده از الگوریتم‌های مونت کارلو و بوت استرپ

عباس محمدخانی[۱]، نصراله ایران‌پناه[۲]

[۱] گروه آمار موسسه آموزش عالی جهاد دانشگاهی
[۲] گروه آمار دانشگاه تربیت مدرس

چکیده: در نظریه نمونه‌گیری از جامعه محدود، هدف برآورد پارامترهای مورد نظر مانند میانگین، درصد و نسبت (میانگین دو جامعه) با استفاده از یک روش نمونه‌گیری مناسب و همچنین مقایسه بین روش‌های مختلف نمونه‌گیری از نظر اندازه‌های دقت برآوردگرها مانند اریبی و خطای معیار است.

در این مقاله مقایسه بین روش‌های نمونه‌گیری مانند تصادفی ساده، طبقه‌بندی و خوشه‌ای به همراه برآوردگرهای نسبتی، رگرسیونی، هارتلی ـ راس و نسبت را با استفاده از دو الگوریتم مونت کارلو و بوت استرپ بر روی یک مثال واقعی (سرشماری دهستانهای کشور در سال ۱۳۷۵) ارائه می‌شود.

در روش مونت کارلو با نمونه‌گیری‌های مکرر از جامعه موجود و در روش بوت استرپ با نمونه‌گیری‌های مکرر از تنها یک نمونه موجود و محاسبه برآورد مورد نظر اندازه‌های دقت برآوردگرها را در روش‌های مختلف نمونه‌گیری برآورد و مقایسه می‌شود. الگوریتم‌های دو روش مونت کارلو و بوت استرپ با استفاده از نرم‌افزار SPLUS برنامه‌نویسی شده است.

همچنین نرم‌افزاری با Visual Basic طراحی گردیده است که روش‌های مختلف نمونه‌گیری را برای برآورد پارامترهای مختلف جامعه به همراه فاصله اطمینان انجام می‌دهد و این روش‌ها را از نظر اندازه دقت با هم مقایسه می‌کند.

واژه‌های کلیدی: مونت کارلو، بوت استرپ، فاصله اطمینان صدکی

۱ مقدمه

نظریه آمار همواره تلاش می‌کند به سه پرسش اساسی زیر پاسخ دهد:

۱) داده‌ها را چگونه جمع‌آوری شود؟

۲) داده‌های جمع‌آوری شده را چگونه خلاصه و تحلیل شود؟

۳) دقت خلاصه داده‌ها چقدر است؟

پرسش ۱ هدف روش‌های نمونه‌گیری و پرسش ۳ از جمله مباحث مهم استنباط آماری است. در نمونه‌گیری از جامعه محدود به حجم N بر اساس اطلاعات موجود از ساختار جامعه، هزینه نمونه‌گیری، سادگی اجرا، هدف مورد بررسی و به خصوص دقت نمونه‌گیری ممکن است روش‌های مختلف نمونه‌گیری قابل اجرا باشد. از جمله روش‌های نمونه‌گیری می‌توان به تصادفی ساده، طبقه‌بندی، خوشه‌ای (یک مرحله‌ای، دو مرحله‌ای و ...)، سیستماتیک و با احتمال متغیر اشاره کرد. میانگین، درصد و نسبت (دو میانگین) از جمله پارامترهای مهم جامعه است که بر اساس روش‌های مختلف نمونه‌گیری به دنبال برآورد مناسب آنها هستیم. اریبی و خطای معیار از مهمترین اندازه‌های دقت برآوردگرها در نظریه نمونه‌گیری هستند.

برآوردگرهای میانگین و درصد نمونه در اکثر روش‌های نمونه‌گیری ناریب و خطای معیار آنها دارای شکل بسته دقیق است. اما برآوردگرهایی مانند نسبت (دو میانگین) نمونه در ساده‌ترین روش نمونه‌گیری یعنی نمونه‌گیری تصادفی ساده، اریب و خطای معیار آن دارای شکل بسته دقیق نمی‌باشد. از شبیه‌سازی مونت کارلو و الگوریتم بوت استرپ می‌توان برای برآورد اریبی و خطای معیار برآوردگرها استفاده کرد.

افرون [۵] و [۶] روش بوت‌استرپ را برای برآورد توزیع و اندازه‌های دقت مانند اریبی و خطای معیار آماره/برآورد ارائه کرد. این روش بر اساس نمونه‌گیری تصادفی ساده باجایگذاری از جامعه است و در روشهای دیگر نمونه‌گیری کاربرد ندارد. بیکل و فریدمن [۴] کاربرد روش بوت‌استرپ را در نمونه‌گیری طبقه‌بندی ارائه کردند.

در بخش ۲ روش‌های مختلف نمونه‌گیری به همراه برآوردگرهای مختلف ارائه می‌شود. در بخش‌های ۳ و ۴ الگوریتم‌های مونت کارلو و بوت استرپ ارائه می‌گردد. در بخش ۵ دو الگوریتم مونت کارلو و بوت استرپ در روش‌های مختلف نمونه‌گیری و در بخش ۶ یک مثال کاربردی برای اجرای دو الگوریتم ارائه می‌گردد و سرانجام در بخش ۷ نرم‌افزار طراحی شده برای انجام و مقایسه روش‌های نمونه‌گیری ارائه می‌شود.

۲ روش‌های نمونه‌گیری

جامعه‌ای محدود به حجم N را با مقادیر صفت $(x_1, \ldots, x_N)$ در نظر می‌گیریم. از جمله اهداف مهم نظریه نمونه‌گیری برآورد پارامترهای میانگین جامعه $\bar{x}_N$ به روش‌های مختلف نمونه‌گیری و برآورد اندازه دقت آن است. در زیر مروری بر برآوردگرها و دقت آنها در روش‌های مختلف نمونه‌گیری خواهیم داشت.

۱) نمونه‌گیری تصادفی ساده: فرض می‌کنیم $(X_1, \ldots, X_n)$ یک نمونه تصادفی ساده با

جایگذاری از جامعه محدود $(x_1, \ldots, x_N)$ باشد در این صورت:

$$\bar{x}_n = \hat{\bar{X}}_N = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad var(\hat{\bar{x}}_N) = \frac{\sigma^2}{n}, \quad \hat{var}(\hat{\bar{x}}_N) = \frac{s_n^2}{n}.$$

برآوردگر میانگین جامعه و برآورد واریانس آن نااریب هستند.

۲) نمونه‌گیری طبقه‌بندی: در نمونه‌گیری طبقه‌بندی فرض می‌شود، واحدهای جامعه در k طبقه مجزا قرار دارند (N_i حجم طبقه iام است، به طوریکه $N = \sum_{i=1}^{k} N_i$). یکی از روش‌های تعیین اندازه نمونه در طبقات روش تخصیص متناسب است. در این روش از هر طبقه به طور مجزا نمونه‌گیری تصادفی ساده به حجم n_i، $(n_i = \frac{N_i}{N} n)$ استخراج و با وزن مناسب $(W_i = \frac{N_i}{N})$ نتایج طبقات بصورت

$$\hat{\bar{x}}_N = \bar{X}_{st} = \sum_{i=1}^{k} W_i \bar{X}_{n_i},$$
$$var(\hat{\bar{x}}_N) = \sum_{i=1}^{k} W_i^2 \frac{\sigma_i^2}{n_i},$$
$$\hat{var}(\hat{\bar{x}}_N) = \sum_{i=1}^{k} W_i^2 \frac{s^2{}_{n_i}}{n_i}.$$

ادغام می‌شوند. برآوردگرهای میانگین جامعه و برآورد واریانس آن نااریب هستند. دقت نمونه‌گیری طبقه‌بندی همواره بیشتر از تصادفی ساده است.

۳) نمونه‌گیری خوشه‌ای دو مرحله‌ای: در نمونه‌گیری خوشه‌ای فرض می‌شود، واحدهای جامعه در N خوشه مجزا قرار دارند (M_i حجم خوشه iام جامعه است به طوریکه $M_0 = \sum_{i=1}^{N} M_i$ حجم جامعه و $\bar{M} = \frac{M_0}{N}$ متوسط حجم هر خوشه در جامعه است). در نمونه‌گیری خوشه‌ای دو مرحله‌ای، در مرحله اول از بین N خوشه جامعه n خوشه نمونه‌گیری و سپس از داخل n خوشه نمونه، نمونه‌گیری مجددی انجام می‌شود (نمونه‌گیری

در هر دو مرحله تصادفی ساده است). نتایج هر خوشه به طور متناسب به صورت

$$\hat{\bar{x}}_N = \bar{X}_{cl} = \frac{1}{n}\sum_{i=1}^{n}\frac{Mi}{\bar{M}}\bar{X}_{i.},$$

$$Var(\hat{\bar{x}}_N) = \frac{\sigma^2{}_b}{n} + \frac{1}{nN}\sum_{i=1}^{N}(\frac{M_i}{\bar{M}})^2\frac{\sigma_i{}^2}{m_i},$$

$$\hat{Var}(\hat{\bar{x}}_N) = \frac{s^2{}_b}{n} + \frac{1}{nN}\sum_{i=1}^{n}(\frac{M_i}{\bar{M}})^2\frac{s_i{}^2}{m_i}.$$

ادغام می‌شود. که در آن $\sigma^2{}_b$ و $s^2{}_b$ به ترتیب واریانس بین میانگین‌های خوشه‌های جامعه و نمونه هستند. برآوردگرهای میانگین جامعه و برآورد واریانس آن نااریب هستند. نمونه‌گیری خوشه‌ای درکاربرد ساده ولی معمولاً دقت بالایی ندارد.

در زیر با استفاده از یک صفت کمکی (Y) که همبستگی خطی بالایی با صفت اصلی (X) دارد و در مورد جامعه آن اطلاع کافی وجود دارد می‌توان با روش نمونه‌گیری تصادفی ساده برآوردهای دقیقتری برای میانگین جامعه ارائه کرد. فرض کنید $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ یک نمونه تصادفی ساده باجایگذاری از جامعه محدود زوجی $\{(x_1,y_1),\ldots,(x_N,y_N)\}$ است و در مورد صفت کمکی Y اطلاع کامل وجود دارد.

۴) برآورد نسبتی (نوع۱):

$$\hat{\bar{x}}_N = \bar{X}_{r1} = R_n.\bar{y}_N, \qquad (R_n = \frac{\bar{X}_n}{\bar{Y}_n}).$$

۵) برآورد رگرسیونی:

$$\hat{\bar{x}}_N = \bar{X}_{re} = \bar{X}_n + \hat{\beta}(\bar{y}_N - \bar{Y}_n), \qquad (\hat{\beta} = \frac{S(X,Y)}{S_y^2}).$$

۶) برآورد نسبتی (نوع ۲):

$$\hat{\bar{x}}_N = \bar{X}_{r2} = \bar{R}_n.\bar{y}_N, \qquad (\bar{R}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{Y_i}).$$

۷) برآورد هارتلی ـ راس:

$$\hat{\bar{x}}_N = \bar{X}_{HR} = \bar{R}_n.\bar{y}_N + \frac{n(N-1)}{N(n-1)}(\bar{X}_n - \bar{R}_n.\bar{Y}_n).$$

۸) برآورد نسبت (میانگین دو جامعه):

$$\hat{r_N} = r_n = \frac{\bar{X}_n}{\bar{Y}_n}.$$

از بین برآوردگرهای ۴ تا ۸ تنها برآوردگر هارتلی ـ راس ناااریب است و سایر برآوردگرها مجانباً ناااریب هستند. همچنین واریانس برآوردگرهای ۴ تا ۸ جواب دقیق ندارد و تنها جواب مجانبی دارند. برای آشنایی بیشتر با روش‌های نمونه‌گیری به مرجع [۲] رجوع کنید.

۳ شبیه‌سازی مونت کارلو

فرض کنید $X_1, \ldots, X_n$ متغیرهای تصادفی مستقل و هم توزیع $(i.i.d.)$ با تابع توزیع تجمعی معلوم F و T آماره (برآوردگر) مورد نظر باشد. با استفاده از شبیه‌سازی مونت کارلو می‌توان توزیع، اریبی، خطای معیار و ... T را بصورت مراحل زیر تقریب کرد.

۱) نمونه تصادفی مستقل و همتوزیع $(X_1, \ldots, X_n)$ از توزیع F بدست می‌آید.

۲) آماره T بر روی نمونه مرحله ۱ بصورت $T \equiv T(X_1, \ldots, X_n)$ محاسبه می‌گردد.

۳) مراحل ۱ و ۲ B بار تکرار می‌شود و $T_1, \ldots, T_B$ بدست می‌آید.

۴) برای Bهای بزرگ تابع توزیع، اریبی و خطای معیار T بصورت

$$F_T(t) \simeq \frac{1}{B}\sum_{i=1}^{B} I(T_i \le t)$$

$$bias(T) \simeq \frac{1}{B}\sum_{i=1}^{B} T_i - \theta$$

$$SE(T) \simeq \{\frac{1}{B}\sum_{i=1}^{B}[T_i - \frac{1}{B}\sum_{j=1}^{B} T_j]^2\}^{\frac{1}{2}}$$

با هر میزان دقتی تقریب می‌شود. در نمونه‌گیری از جامعه محدود به حجم N نیز فرض می‌شود $(X_1, \ldots, X_n)$ یک نمونه تصادفی ساده (باجایگذاری یا بدون جایگذاری) از جامعه$(x_1, \ldots, x_N)$ است، می‌توان توزیع و مشخصات T را مشابه الگوریتم بالا بدست آورد. فقط به جای نمونه‌گیری به حجم n از F در مرحله ۱، نمونه‌گیری از $(x_1, \ldots, x_N)$ انجام می‌شود.

۴ الگوریتم بوت استرپ

فرض کنید $X_1, \ldots, X_n$ متغیرهای تصادفی مستقل و همتوزیع $(i.i.d.)$ با تابع توزیع تجمعی نامعلوم F و یا $(X_1, \ldots, X_n)$ یک نمونه تصادفی ساده باجایگذاری از جامعه محدود و نامعلوم $(x_1, \ldots, x_N)$ باشد. با استفاده از روش بوت‌استرپ می‌توان توزیع، اریبی، خطای معیار و ...آماره $T \equiv T(X_1, \ldots, X_n)$ را با داشتن تنها یک نمونه مشاهده شده $(x_1, \ldots, x_n)$ بصورت مراحل زیر تقریب کرد.

۱) نمونه بوت‌استرپ به حجم n، $(X_1^*, \ldots, X_n^*)$ بصورت تصادفی ساده باجایگذاری از $(x_1, \ldots, x_n)$ بدست می‌آید.

۲) آماره T بر روی نمونه بوت‌استرپ مرحله ۱ بصورت $T^* \equiv T(X_1^*, \ldots, X_n^*)$ محاسبه می‌گردد.

۳) مراحل ۱ و ۲ B بار تکرار می‌شود و $T_B^*, \ldots, T_1^*$ بدست می‌آید.

۴) برای Bهای بزرگ تابع توزیع، اریبی و خطای معیار T بصورت

$$\hat{F}_T(t) = \frac{1}{B} \sum_{i=1}^{B} I(T_i^* \leq t),$$

$$\hat{bias}(T) = \frac{1}{B} \sum_{i=1}^{B} T_i^* - T,$$

$$\hat{SE}(T) = \{\frac{1}{B} \sum_{i=1}^{B} [T_i^* - \frac{1}{B} \sum_{j=1}^{B} T_j^*]^2\}^{\frac{1}{2}},$$

برآورد می‌گردد. برای آشنایی بیشتر با الگوریتم بوت‌استرپ کلاسیک به افرون و تیبشیرانی [۷] و در یک مرجع فارسی به [۱] رجوع کنید.

۵ الگوریتم مونت کارلو و بوت‌استرپ در روشهای نمونه‌گیری

این دو الگوریتم برای ۳ روش نمونه‌گیری تصادفی ساده، طبقه‌بندی و خوشه‌ای دومرحله‌ای بصورت زیر خلاصه می‌شود. قابل ذکراست که در الگوریتم مونت کارلو باید کل جامعه (به حجم N) موجود ولی در الگوریتم بوت‌استرپ تنها یک نمونه (به حجم n) کافی است.

۱) نمونه‌گیری تصادفی ساده: دو الگوریتم برا ی برآورد نسبت (میانگین دو جامعه)(حالت ۸) ارائه می‌شود. برای سایر برآوردهای ۱ و ۴ تا ۷ بطور مشابه قابل تکرار است.

الف)الگوریتم مونت کارلو:
از جامعه $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ به روش نمونه‌گیری تصادفی ساده باجایگذاری نمونه $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ بدست می‌آید و سپس برآورد نسبت $r_n = \frac{\bar{x}_n}{\bar{y}_n}$ محاسبه می‌گردد. اگر این عمل B بار تکرار شود، آنگاه $(r_{n,1}, \ldots, r_{n,B})$ بدست می‌آید. توزیع، اریبی و خطای معیار برآوردگر R_n به ازای Bهای بزرگ با جایگذاری برآوردگر R_n به جای T در بخش ۳ قابل تقریب است.

ب) الگوریتم بوت‌استرپ:
از جامعه زوجی به حجم N تنها یک نمونه تصادفی ساده باجایگذاری به حجم n، $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ موجود است. به روش نمونه گیری تصادفی ساده باجایگذاری نمونه بوت‌استرپ به حجم n، $\{(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)\}$ بدست می‌آید. برآورد نسبت $r_n^* = \frac{\bar{x}_n^*}{\bar{y}_n^*}$ در نمونه بوت‌استرپ محاسبه می‌شود. اگر این عمل B بار تکرار شود، آنگاه $(r_{n,1}^*, \ldots, r_{n,B}^*)$ بدست می‌آید. توزیع، اریبی و خطای معیار برآوردگر R_n به ازای Bهای بزرگ با جایگذاری برآوردگر R_n^* به جای T^* در بخش ۴ قابل تقریب است.

۲) نمونه گیری طبقه‌بندی:
الف)الگوریتم مونت کارلو: با داشتن واحدهای جامعه به حجم N که در k طبقه مجزا قرار دارند، اگر B بار نمونه‌گیری به حجم n مشابه نمونه‌گیری طبقه‌بندی بخش ۲ تکرار شود، آنگاه $(\bar{x}_{st,1}, \ldots \bar{x}_{st,B})$ بدست می‌آید. توزیع، اریبی و خطای معیار برآوردگر $\bar{X}_{st}$ به ازای Bهای بزرگ با جایگذاری برآوردگر $\bar{X}_{st}$ به جای T در بخش ۳ قابل تقریب است.
ب)الگوریتم بوت‌استرپ: فرض کنید نمونه‌ای به حجم n (از هر طبقه بطور مجزا نمونه‌ای به حجم n_i به روش تصادفی ساده باجایگذاری بدست آمده است و $n = \sum_{i=1}^{k} n_i$) بدست آمده است. از هر طبقه بطور مجزا نمونه‌ای به حجم n_i به روش تصادفی ساده باجایگذاری استخراج و میانگین کل که حاصل یک نمونه بوت‌استرپ است بصورت $\bar{x}_{st}^* = \sum_{i=1}^{k} W_i \bar{x}_{n_i}^*$ محاسبه می‌شود. اگر این عمل B بار تکرار شود، آنگاه $(\bar{x}_{st,1}^*, \ldots, \bar{x}_{st,B}^*)$ بدست می آید. توزیع، اریبی و خطای معیار برآوردگر $\bar{X}_{st}$ به ازای Bهای بزرگ با جایگذاری برآوردگر $\bar{X}_{st}^*$ به جای T^* در بخش ۴ قابل تقریب است.

۳) نمونه‌گیری خوشه‌ای دومرحله‌ای:
الف)الگوریتم مونت کارلو: با داشتن واحدهای جامعه به حجم M_0 که در N خوشه مجزا قرار دارند، اگر B بار نمونه‌گیری به حجم m_0، مشابه نمونه‌گیری خوشه‌ای دومرحله‌ای بخش ۲ تکرار شود، آنگاه $(\bar{x}_{cl,1}, \ldots \bar{x}_{cl,B})$ بدست می‌آید. توزیع، اریبی و خطای معیار برآوردگر $\bar{X}_{cl}$ به ازای Bهای بزرگ با جایگذاری برآوردگر $\bar{X}_{cl}$ به جای T در بخش ۳ قابل تقریب است.

ب) الگوریتم بوت‌استرپ: فرض کنید ازبین N خوشه جامعه n خوشه و در داخل هر خوشه نمونه، نمونه‌گیری مجددی به حجم m_i (هر دو مرحله نمونه‌گیری تصادفی ساده باجایگذاری است) بدست آمده‌است. از بین n خوشه نمونه به روش تصادفی ساده باجایگذاری n خوشه انتخاب و در داخل هر خوشه نمونه، نمونه‌گیری مجدد با حجم m_i به روش تصادفی ساده باجایگذاری انتخاب می‌شود. میانگین کل که حاصل یک نمونه بوت‌استرپ است بصورت $\bar{x}^*_{cl} = \frac{1}{n}\sum_{i=1}^{n} \frac{M_i}{M}\bar{x}^*_{i.}$ محاسبه می‌شود. اگر این عمل B بار تکرار شود، آنگاه $(\bar{x}^*_{cl,1}, \ldots, \bar{x}^*_{cl,B})$ بدست می آید. توزیع، اریبی و خطای معیار برآوردگر $\bar{X}_{cl}$ به ازای B های بزرگ با جایگذاری برآوردگر $\bar{X}^*_{cl}$ به جای T^* در بخش ۴ قابل تقریب است.

۶ یک مثال کاربردی

دو الگوریتم مونت کارلو و بوت‌استرپ در روشهای مختلف نمونه‌گیری و برآوردهای مختلف در مثال کاربردی زیر اجرا می‌شود.

در سر شماری عمومی نفوس و مسکن سال ۱۳۷۵ از مناطق روستایی کشور [۳] جامعه را $N = ۲۲۱۳$ دهستان در نظر گرفته و هر دهستان بصورت یک واحد نمونه‌گیری در نظر گرفته می‌شود. دهستانهای کشور در $k = ۲۶$ استان تقسیم گردیده‌اند. دو صفت اصلی X تعداد جمعیت با سواد و صفت کمکی Y تعداد خانوار هر دهستان تعریف می‌شود. واحد اندازه‌گیری هزار نفر برای دو صفت تعریف می‌شود. با توجه به جامعه تعریف شده

$$\bar{x}_N = ۶/۱۷۳, \quad \bar{y}_N = ۱/۹۹۳, \quad \sigma_x = ۵/۶۵۱, \quad \sigma_y = ۱/۶۵۷, \quad \rho_{xy} = ۰/۹۸,$$

است. در جدول ۱، مقدار برآورد اریبی و خطای معیار برآوردگر میانگین جامعه و همچنین یک فاصله اطمینان صدکی ۰/۹۵ برای میانگین جامعه به دو روش مونت کارلو و بوت‌استرپ بر اساس نمونه‌ای تصادفی به حجم $n = ۵۰۰$ نشان داده شده است. نمونه‌گیری به سه روش تصادفی ساده باجایگذاری، طبقه بندی تخصیص متناسب و خوشه‌ای دومرحله‌ای و همچنین برآورد نسبتی و رگرسیونی در نمونه‌گیری تصادفی ساده باجایگذاری است.

در نمونه‌گیری طبقه‌بندی تعداد طبقات $k = ۲۶$ استان و در نمونه‌گیری خوشه‌ای تعداد خوشه‌ها $N = ۲۶$ استان در نظر گرفته می‌شود. در نمونه‌گیری خوشه‌ای از بین $N = ۲۶$ خوشه تعداد $n = ۱۰$ خوشه انتخاب و حجم هر خوشه در نمونه (m_i) نیز متناسب با حجم هر خوشه انتخاب شده است. قابل ذکر است، برای انجام دو روش نمونه‌گیری طبقه‌بندی و خوشه‌ای و برای مقایسه بین روشها تعداد نمونه بزرگ $n = ۵۰۰$ از جامعه انتخاب شده است.نتایج حاصل از شبیه‌سازی مونت کارلو با روابط دقیق نظری (و یا مجانبی) برای مقادیر اریبی و خطای معیار برابر است. نتایج حاصل از الگوریتم بوت‌استرپ نیز حاکی از دقت بالای روش در مقایسه با الگوریتم مونت کارلو است.

جدول ۱: مقایسه اریبی، خطای معیار و فاصله اطمینان برآورد میانگین جامعه در روش‌های مختلف نمونه‌گیری ($n = ۵۰۰$) با دو الگوریتم مونت کارلو و بوت استرپ.

نمونه‌گیری	اریبی		خطای معیار		فاصله اطمینان	
	مونت‌کارلو	بوت‌استرپ	مونت‌کارلو	بوت‌استرپ	مونت‌کارلو	بوت‌استرپ
تصادفی ساده	۰٫۰۰۰	۰٫۰۰۰	۰٫۲۵۲	۰٫۲۵۰	(۵٫۷۰،۶٫۶۹)	(۶٫۱۹،۶٫۷۳)
طبقه بندی	۰٫۰۰۰	۰٫۰۰۲	۰٫۲۳۶	۰٫۲۳۹	(۵٫۷۲،۶٫۶۵)	(۵٫۶۲،۶٫۵۸)
خوشه‌ای	۰٫۰۰۰	۰٫۰۰۲	۱٫۵۰۴	۱٫۳۸۸	(۳٫۵۷،۹٫۳۷)	(۲٫۷۳،۸٫۱۸)
برآورد نسبتی	۰٫۰۰۰	۰٫۰۰۰	۰٫۰۵۷	۰٫۰۵۰	(۶٫۰۶،۶٫۲۹)	(۶٫۱۲،۶٫۲۳)
برآورد رگرسیونی	۰٫۰۰۰	۰٫۰۰۰	۰٫۰۵۰	۰٫۰۴۷	(۶٫۰۸،۶٫۲۷)	(۶٫۰۹،۶٫۲۹)

جدول ۲: مقایسه اریبی، خطای معیار و فاصله اطمینان برآوردهای مختلف میانگین و نسبت جامعه در یک نمونه تصادفی ساده ($n = ۱۰$) با دو الگوریتم مونت کارلو و بوت استرپ.

برآورد	اریبی		خطای معیار		فاصله اطمینان	
	مونت‌کارلو	بوت‌استرپ	مونت‌کارلو	بوت‌استرپ	مونت‌کارلو	بوت‌استرپ
معمولی	۰٫۰۰۱	۰٫۰۰۷	۱٫۷۹۶	۰٫۹۸۲	(۳٫۶۰، ۱۰٫۶۹)	(۴٫۷۰، ۸٫۵۳)
نسبتی ۱	۰٫۰۲۷	۰٫۰۳۱	۰٫۳۷۴	۰٫۳۷۶	(۵٫۴۲،۶٫۸۷)	(۵٫۶۴،۶٫۴۲)
نسبتی ۲	۰٫۲۵۳	۰٫۱۹۷	۴٫۶۵۲	۲٫۴۲۹	(۵٫۲۲،۷٫۰۴)	(۵٫۶۳،۶٫۵۸)
هارتلی رائو	۰٫۰۰۶	۰٫۰۱۰	۱٫۳۶۶	۰٫۹۲۱	(۵٫۳۸،۷٫۰۹)	(۵٫۶۴،۶٫۴۲)
رگرسیونی	۰٫۰۳۱	۰٫۰۲۳	۰٫۳۸۹	۰٫۲۲۷	(۵٫۳۹،۶٫۹۰)	(۵٫۵۵،۶٫۴۴)
نسبت	۰٫۰۱۴	۰٫۰۰۵	۰٫۱۸۸	۰٫۱۹۸	(۲٫۷۲،۳٫۴۵)	(۲٫۸۳،۳٫۲۲)

در جدول ۲، با استفاده از تنها یک نمونه کوچک به حجم $n = ۱۰$ به روش تصادفی ساده باجایگذاری برآورد مقدار اریبی و خطای معیار برآوردگر میانگین جامعه و یک فاصله اطمینان صدکی ۰٫۹۵ به دو روش مونت کارلو و بوت‌استرپ نشان داده شده‌است. برآورد میانگین جامعه به روشهای معمولی، نسبتی نوع ۱ و ۲، هارتلی ـ راس، رگرسیونی و برآورد نسبت (میانگین دو جامعه) است.

قابل ذکر است که به جز برآوردگر معمولی میانگین جامعه برای برآوردگرهای دیگر مقدار اریبی و خطای معیار جواب دقیقی نظری ندارد و به صورت شبیه سازی مونت کارلو این مقادیر دقیق محاسبه گردیده است. نتایج حاصل از الگوریتم بوت‌استرپ حاکی از دقت بالای روش در مقایسه با الگوریتم مونت کارلو است. قابل توجه است که در الگوریتم مونت کارلو بازنمونه‌گیری به حجم $n = ۱۰$ از جامعه $N = ۲۲۱۳$ و در الگوریتم بوت‌استرپ بازنمونه‌گیری به حجم $n = ۱۰$ از تنها یک نمونه به حجم $n = ۱۰$ بدست آمده است.

۷ طراحی نرم‌افزار روش‌های نمونه‌گیری

با استفاده از Visual Basic نرم‌افزاری تهیه گردیده است که برآورد میانگین، درصد، مقدار کل، تعداد کل و نسبت (میانگین دو جامعه) را به روشهای مختلف نمونه‌گیری تصادفی ساده (با و بدون جایگذاری)، طبقه‌بندی، خوشه‌ای، سیستماتیک و با احتمال متغیر محاسبه می‌کند.

در بانک اطلاعاتی نرم‌افزار مقادیر دو صفت اصلی وکمکی X و Y در جامعه‌ای به حجم N قرار می‌گیرد. در نرم‌افزار به روشهای مختلف نمونه‌گیری، نمونه‌های به حجم n بصورت تصادفی استخراج می‌شود. در نرم‌افزار علاوه بر برآورد میانگین، درصد، مقدار کل، تعداد کل و نسبت (میانگین دو جامعه)، برآورد مقدار واریانس برآوردگر و همچنین فاصله اطمینان با ضریب اطمینان دلخواه نیز محاسبه می‌شود.

هدف از طراحی این نرم‌افزار مقایسه روشهای مختلف نمونه‌گیری با استفاده مثالهای واقعی از جامعه است. این نرم افزار در نسخه‌های مختلف Windows بصورت خودکار (Auto Run) قابل نصب و اجرا است. این نرم‌افزار از نویسنده اول قابل دریافت است.

مراجع

[۱] ایران پناه، نصراله و پاشا، عین‌اله (۱۳۷۶). آشنایی با الگوریتم بوت استرپ. اندیشه آماری، سال دوم، شماره ۱، ۴۶-۳۳.

[۲] عمیدی، علی (۱۳۷۸). نظریه نمونه‌گیری و کاربردهای آن. جلد اول و دوم، مرکز نشر دانشگاهی.

[۳] شناسنامه دهستانهای کشور، سرشماری عمومی نفوس و مسکن، ۱۳۷۵. جلدهای ۲۶-۱، جداول ۴ و ۵.

[4] Bickel, P. J. and Freedman, D. A. (1984), Asymptotic Normality and the Bootstrap in Stratified Sampling, Annals of Statistics, **12**, 470-482.

[5] Efron, B. and Tibshirani, R. (1993), An Introductionto the Bootstrap, Chapman and Hall, New York.

[6] Efron, B. (1979), The Jackknife, the Bootstrap and Other Resampling Plans, Society for Industrial and Applied Mathematics, Philadelphia, PA.

[7] Efron, B. and Tibshirani, R. (1993), An Introductionto the Bootstrap, Chapman and Hall, New York.

پیشگویی فضایی بیزی با استفاده از روشهای مونت کارلو

محسن محمّدزاده، مجید جعفری خالدی

گروه آمار دانشگاه تربیت مدرس

چکیده: اغلب، تجزیه و تحلیل داده‌های فضایی با فرض گوسی بودن میدان تصادفی صورت می‌پذیرد. امّا در بعضی مسائل با داده‌های مثبتی مواجه‌ایم که لگاریتم آنها گوسی یا تقریباً گوسی است. در این مقاله با فرض آنکه میدان تصادفی تبدیل یافته، گوسی با ساختار میانگین و کوواریانس پارامتری است و پارامترها متغیرهایی تصادفی با توزیع پیشین مناسب هستند پیشگوی فضایی با استفاده از رهیافت بیزی تعیین می‌شود. چون در مسائل کاربردی معمولاً محاسبه این پیشگو دشوار است بکمک روشهای مونت کارلو، یک تقریب برای توزیع پیشگو ارائه و براساس آن پیشگوی فضایی بیزی تقریبی معرفی خواهد شد.

واژه‌های کلیدی: داده‌های فضایی، میدان تصادفی لگ ـ گوسی، پیشگویی فضایی بیزی، روشهای مونت کارلو

۱ مقدمه

برای مدل‌بندی داده‌های فضایی معمولاً فرض می‌شود داده‌ها تحققی از یک میدان تصادفی $\{Z(t)\ ;\ t \in D \subseteq R^d\}$، $d \geq 1$، هستند، که در آن D یک مجموعه اندیس گذار است. پیشگویی فضایی مقدار نامعلوم میدان تصادفی در یک یا چند موقعیت مشخص و بر اساس مشاهدات بدست آمده در سایر موقعیتها، در علوم مختلف از جمله هواشناسی، محیط زیست، همه‌گیر شناسی، زمین شناسی، جغرافیا، زمین شناسی، جغرافی، کشاورزی و . . . که بنوعی با داده‌های فضایی سر و کار دارند، یکی از مسائل مهم بشمار می‌رود. اغلب با فرض گوسی بودن میدان تصادفی، پیشگوی فضایی تعیین می‌شود. اما در بعضی مسائل با داده‌های مثبتی مواجه‌ایم که از مدل گوسی تبعیت نمی‌کنند، بلکه تبدیل لگاریتمی آنها گوسی یا تقریباً گوسی است. در این صورت میدان تصادفی مورد نظر لگ ـ گوسی[1] نامیده می‌شود. بعنوان مثال در هواشناسی مشاهدات مربوط به میزان بارندگی عموماً از این نوع بشمار می‌روند. در این مقاله با فرض آنکه میدان تصادفی تبدیل یافته، گوسی با ساختار میانگین و کوواریانس پارامتری است و پارامترها متغیرهایی تصادفی با توزیع پیشین مناسب هستند، پیشگوی فضایی با استفاده از رهیافت بیزی تعیین می‌شود. اما در عمل محاسبه این پیشگو دشوار است، لذا در این مقاله با

1) Log-Gaussian

استفاده از روشهای مونت کارلو، توزیع تقریبی پیشگو و پیشگوی فضایی بیزی تقریبی ارائه خواهند شد.

در سالهای اخیر بدنبال توسعه روشهای مونت کارلو در آمار بیزی، استفاده از روشهای بیزی برای تجزیه وتحلیل داده‌های فضایی رشد و توسعه چشمگیری یافته است. برگر و همکاران (۲۰۰۱) توزیع‌های پیشینی که پسین‌های نامناسب[۲] نتیجه می‌دهند، را مطالعه کردند. برای حالتی که با داده‌های گمشده مواجه‌ایم، روش بیزی ارائه شده توسط اوح و همکاران (۲۰۰۲) می‌تواند رهگشا باشد. آلدورث و کرسی (۲۰۰۳) پیشگویی بیز تجربی توابع غیر خطی از میدان تصادفی با مشاهدات نویز دار را بررسی کردند. هنگامی که مشاهدات فاقد نویز هستند، جعفری و محمدزاده (۲۰۰۳) پیشگوی فضایی بیزی را برای یک میدان تصادفی گوسی ارائه کردند. جعفری و محمدزاده (۲۰۰۴a) نیز پیشگویی برای میدانهای تصادفی گوسی با مشاهدات نویز دار را ارائه نمودند. سپس جعفری و محمدزاده (۲۰۰۴b) پیشگوی فضایی بیزی برای یک میدان تصادفی که تبدیلی نامعلوم از آنها گوسی باشد را با استفاده از توزیع‌های پیشین نامناسب تعیین کردند که بدنبال آن در این مقاله با استفاده از توزیع‌های پیشین مناسب و بکمک تکنیک مونت کارلو پیشگوی فضایی بیزی برای میدان تصادفی لگ ـ گوسی ارائه می‌شود.

۲ پیشگویی فضایی بیزی

فرض کنید $\{Z(t)\ ;\ t \in D\}$ یک میدان تصادفی از متغیرهای تصادفی مثبت و بردار $Z = (Z(t_1), \cdots, Z(t_n))$ میدان تصادفی را در موقعیتهای $t_1, \cdots, t_n$ نمایش دهد و پیشگویی مقدار نامعلوم میدان تصادفی در نقطه t_0، یعنی $Z(t_0)$، براساس بردار مشاهدات $z = (z(t_1), \cdots, z(t_n))$ مورد نظر باشد. با فرض آنکه میدان تصادفی

$$\{Y(t) = \ln Z(t)\ ;\ t \in D\}$$

گوسی حقیقی مقدار بترتیب با توابع میانگین و کوواریانس

$$\begin{aligned} \mathrm{E}(Y(t)) &= f'(s)\beta \\ \mathrm{Cov}(Y(s), Y(t)) &= \sigma^2 \rho(s,t;\theta) \end{aligned}$$

باشد، که در آن $f(t) = (f_1(t), \cdots, f_q(t))'$ یک بردار $p \times 1$ از مولفه‌های غیرتصادفی معلوم، $\beta = (\beta_1, \cdots, \beta_p) \in R^p$ پارامترهای رگرسیونی، $\sigma^2 = Var(Y(t))$ واریانس و $\rho(\cdot,\cdot;\theta)$ تابع همبستگی میدان $Y(\cdot)$ با بردار پارامترهای $\theta = (\theta_1 \cdots, \theta_q) \in \Theta \subseteq R^q$ است. بدین ترتیب

$$Y = \ln Z = (\ln Z(t_1), \cdots, \ln Z(t_n)) \sim N_n(X\beta, \sigma^2 \Sigma_\theta)$$

2) Improper

است، که در آن ماتریس طرح $X_{n\times q}$ و ماتریس همبستگی فضایی نمونه Σ_θ بصورت

$$X=\begin{pmatrix} f'(t_1)\\ \vdots\\ f'(t_n)\end{pmatrix},\quad \Sigma_\theta=(\rho(t_i,t_j;\theta))$$

تعریف می‌شوند. با فرض اینکه X ماتریسی رتبه کامل و Σ_θ ماتریسی معین مثبت باشد، تابع درستنمایی بعد از حذف ضرایب ثابت بصورت

$$L(\phi;z)=(\frac{1}{2\sigma^2\pi})^{\frac{n}{2}}|\Sigma_\theta|^{-\frac{1}{2}}\exp\{-\frac{1}{2\sigma^2}(\ln z-X\beta)'\Sigma_\theta^{-1}(\ln z-X\beta)\}J$$

خواهد بود، که در آن $J=(\Pi_{i=1}^n|z_i|)^{-1}$ ژاکوبین تبدیل و $\phi=(\beta,\sigma^2,\theta)$ بردار $p+q+1$ بعدی پارامترهای مدل است. با توجه به‌گوسی بودن میدان تصادفی $Y(\cdot)$، توزیع توام $(Y(t_0),Y)$ نرمال $n+1$ متغیره است و می‌توان نشان داد توزیع $Y(t_0)$ بشرط z و ϕ، نرمال بترتیب با میانگین و واریانس

$$\begin{aligned} E(Y(t_0)|z,\phi) &= (f'(t_0)\beta+r'_\theta\Sigma_\theta^{-1}(\ln z-X\beta)\\ Var(Y(t_0)|z,\phi) &= \sigma^2(1-\mathbf{r}'_\theta\Sigma_\theta^{-1}\mathbf{r}_\theta))\end{aligned}$$

خواهد بود، که در آن $r_\theta=(\rho(t_0,t_i;\theta))_{n\times 1}$ است. بنابراین بشرط z و ϕ، توزیع پیشگو برای $Z(t_0)=\exp(Y(t_0))$ بصورت

$$\begin{aligned} f(z_0|z,\phi) &= (\frac{1}{2\pi\sigma^2 z_0^2(1-\mathbf{r}'_\theta\Sigma_\theta^{-1}\mathbf{r}_\theta)})^{\frac{1}{2}}\\ &\times \exp\{-\frac{(\ln z_0-E(Y(t_0)|z,\phi))^2}{2\sigma^2(1-\mathbf{r}'_\theta\Sigma_\theta^{-1}\mathbf{r}_\theta)}\} \qquad z_0>0 \quad (1)\end{aligned}$$

خواهد بود. با فرض معلوم بودن پارامتر ϕ، پیشگوی بهینه برای $Z(t_0)$ با تابع زیان قدر مطلق خطا، میانه توزیع (۱) یعنی

$$Median\ of\ [Z(t_0)|z,\phi]=\exp\{f'(t_0)\beta+r'_\theta\Sigma_\theta^{-1}(\ln z-X\beta)\}$$

می‌باشد. چون پیشگوی بهینه به پارامتر $p+q+1$ بعدی ϕ بستگی دارد، هنگامی که ϕ نامعلوم است، پیشگوی فضایی به روش بیزی تعیین می‌گردد. بدین منظور توزیعهای پیشین مناسب برای پارامترها اختیار می‌شود و به جهت ساده‌تر شدن محاسبات، پارامترهای (β,σ^2) از پارامتر θ مستقل در نظرگرفته می‌شوند. بنابراین توزیع پیشین توام پارامترهای مدل را می‌توان بفرم

$$\pi(\phi)=\pi(\beta,\sigma^2,\theta)=\pi(\beta|\sigma^2)\pi(\sigma^2)\pi(\theta)$$

نوشت. اکنون توزیعهای پیشین حاشیه‌ای

$$(\beta|\sigma^۲) \sim N(\beta_۰, \sigma^۲ V_۰)\, \sigma^۲ \sim \chi^۲_{SCI}(a, b)$$

را در نظر می‌گیریم، که مقادیر $\beta_۰$، $V_۰$، a و b معلوم هستند و $\chi^۲_{SCI}$ نشاندهنده توزیع کای دوی معکوس می‌باشد. توزیع پیشین $\pi(\theta)$ نیز بطور دلخواه اختیار می‌شود. چون تعیین توزیع پسین پارامترهای مدل از رابطه

$$\pi(\beta, \sigma^۲, \theta|z) \propto f(z|\phi)\pi(\beta|\sigma^۲)\pi(\sigma^۲)\pi(\theta) \qquad (۲)$$

دشوار می‌باشد، ابتدا توزیع پسین حاشیه‌ای β را با شرطی کردن بر پارامترهای $\sigma^۲$ و θ تعیین می‌کنیم. بسادگی می‌توان نشان داد این توزیع بفرم

$$(\beta|z, \sigma^۲, \theta) \sim N(\hat{\beta}_۱, \sigma^۲ \hat{V}) \qquad (۳)$$

است، که در آن

$$\begin{aligned} \hat{\beta}_۱ &= (V_۰^{-۱} + X'\Sigma_\theta^{-۱}X)^{-۱}(V_۰^{-۱}\beta_۰ + X'\Sigma_\theta^{-۱}y) \\ \hat{V} &= (V_۰^{-۱} + X'\Sigma_\theta^{-۱}X)^{-۱} \end{aligned}$$

می‌باشند. سپس توزیع پسین حاشیه‌ای $\sigma^۲$ با شرطی کردن بر پارامتر θ بصورت

$$(\sigma^۲|z, \theta) \sim \chi^۲_{SCI}(a + n, b') \qquad (۴)$$

تعیین می‌شود، که در آن

$$b' = \frac{ab + n\hat{\sigma}^۲ + \hat{\beta}' V_{\hat{\beta}}^{-۱}\hat{\beta} + \beta_۰ V_۰^{-۱}\beta_۰ - \mathbf{m}'\hat{V}\mathbf{m}}{a + n}$$

و $\hat{\sigma}^۲$، $\hat{\beta}$، $V_{\hat{\beta}}$ و $\mathbf{m}$ بترتیب برابر

$$\begin{aligned} \hat{\sigma}^۲ &= \frac{۱}{n}(y - X\beta)'\Sigma_\theta^{-۱}(y - X\beta) \\ \hat{\beta} &= (X'\Sigma_\theta^{-۱}X)^{-۱}X'\Sigma_\theta^{-۱}y \\ V_{\hat{\beta}} &= (X'\Sigma_\theta^{-۱}X)^{-۱} \\ \mathbf{m} &= (V_{\hat{\beta}}^{-۱}\hat{\beta} + V_۰^{-۱}\beta_۰) \end{aligned}$$

می‌باشند. در مرحله بعد توزیع پسین حاشیه‌ای θ بصورت

$$\pi(\theta|z) \propto \frac{f(z|\phi)\pi(\beta|\sigma^2)\pi(\sigma^2)\pi(\theta)}{\pi(\beta|z,\sigma^2,\theta)\pi(\sigma^2|z,\theta)} \qquad (۵)$$

تعیین می‌شود. اکنون بر اساس قانون بیز و با جایگذاری توزیعهای پسین حاشیه‌ای (۳)، (۴) و (۵) در رابطه

$$\pi(\phi|z) = \pi(\beta,\sigma^2,\theta|z) = \pi(\beta|z,\sigma^2,\theta)\pi(\sigma^2|z,\theta)\pi(\theta|z) \qquad (۶)$$

توزیع پسین پارامترهای مدل قابل تعیین می‌باشد. از آنجاکه عموما تعیین $\pi(\theta|z)$ و به تبع آن توزیع پسین (۶) بسیار دشوار است، با استفاده از روشهای مونت کارلوی زنجیر مارکف (MCMC)[3]، یک نمونه تصادفی از توزیعهای (۵) و (۶) بدست آورده می‌شود و استنباط بیزی پارامترهای مدل براساس این نمونه صورت می‌پذیرد (جعفری و محمدزاده، ۲۰۰۳).
اکنون می‌توان توزیع پیشگوی بیزی $Z(t_0)$ را از رابطه

$$f(z_0|z) = \int f(z_0|z,\phi)\pi(\phi|z)d\phi \qquad (۷)$$

بدست آورد. با توجه به آنکه محاسبه تحلیلی توزیع پیشگو از رابطه (۷) مشکل می‌باشد، ابتدا با انتگرال‌گیری از $\pi(\beta|z,\sigma^2,\theta)f(z|\phi)$ نسبت به β، می‌توان نشان داد توزیع $Z(t_0)$ بشرط z,σ^2,θ لگ ـ نرمال بصورت

$$(Z(t_0)|z,\sigma^2,\theta) \sim LN(\mu_1,\Sigma_1)$$

است، که در آن

$$\begin{aligned}
\mu_1 &= (f'(t_0) - r'_\theta\Sigma_\theta^{-1}X)(V_0^{-1} + X'\Sigma_\theta^{-1}X)^{-1}V_0^{-1}\beta_0 \\
&+ [r'_\theta\Sigma_\theta^{-1} + (f'(t_0) - r'_\theta\Sigma_\theta^{-1}X)(V_0^{-1} + X'\Sigma_\theta^{-1}X)^{-1}X'\Sigma_\theta^{-1}]y \\
\Sigma_1 &= \sigma^2(1 - r'_\theta\Sigma_\theta^{-1}r) \\
&+ [(f'(t_0) - r'_\theta\Sigma_\theta^{-1}X)(V_0^{-1} + X'\Sigma_\theta^{-1}X)^{-1}(f'(t_0) - r'_\theta\Sigma_\theta^{-1}X)']
\end{aligned}$$

سپس با انتگرال‌گیری از $\pi(\sigma^2|z,\theta)f(z(t_0)|z,\sigma^2,\theta)$ نسبت به σ^2، توزیع $Z(t_0)$ بشرط z,θ لگ ـ تی استیودنت بصورت

$$(Z(t_0)|z,\theta) \sim LT_{n+a}(\mu_1, b'\Sigma_1)$$

3) Monte Carlo Markov Chain

بدست می‌آید. اکنون می‌توان توزیع پیشگوی بیزی را بصورت

$$f(z_\circ|z) = \int \pi(\theta|z) f(z_\circ|z,\theta) d\theta \qquad (۸)$$

نوشت، که یک انتگرال q گانه و تعیین آن ساده‌تر از رابطه (۷) می‌باشد. با این وجود محاسبه تحلیلی آن امکان‌پذیر نیست، لذا روش مونت کارلو جهت تقریب آن مورد استفاده قرار می‌گیرد. برای این منظور، از توزیع پیشین $\pi(\theta)$ یک نمونه تصادفی $i.i.d$ به حجم M بصورت $\{\theta_i\}_{i=1}^{M}$ انتخاب می‌شود. بدین ترتیب، برای مقادیر به اندازه کافی بزرگ M می‌توان

$$f(z_\circ|z) \approx \frac{\sum_{i=1}^{M} f(z(t_\circ)|z,\theta_i) f(z|\theta_i)}{\sum_{i=1}^{M} f(z|\theta_i)}$$

را بعنوان توزیع تقریبی پیشگو مورد استفاده قرار داد و بر اساس تابع زیان قدر مطلق خطا پیشگوی فضایی بیزی تقریبی برای $Z(t_\circ)$ میانه توزیع تقریبی پیشگو، یعنی

$$\hat{Z}(t_\circ) = Median\ of\ [Z(t_\circ)|z]$$

تعیین می‌گردد.

۳ بحث و نتیجه‌گیری

برای یک میدان تصادفی لگ ـ گوسی بر اساس رهیافت بیزی و بکمک روشهای مونت کارلو پیشگوی فضایی بیزی تعیین گردید. لازم بذکر است که روابط بیان شده برای میدانهای تصادفی که تبدیلی معلوم از آنها گوسی است، قابل توسعه می‌باشد. در این مقاله پیشگویی فضایی بیزی با فرض اینکه داده‌ها فاقد نویز هستند، مورد مطالعه قرارگرفت. اما در اغلب مسائل آمار فضایی، بدلایل مختلف از جمله خطای اندازه‌گیری، مشاهدات همراه با نویز می‌باشند. انجام این نوع پیشگویی بر اساس مشاهدات نویزدار بعنوان یک مسئله جدید قابل بررسی است.

مراجع

[1] Aldworth, J. and Cressie, N.(2003). Prediction of Nonlinear Spatial Functionals. Statistical Planning and Inference (Article in Press).

[2] Berger, J. O., Deoliveira, V. and Sanso, B. (2001). Objective Bayesian Analysis of Spatially Correlated Data. JASA 96: 1351-1370.

[3] Jafari, K. M. and Mohammadzadeh, M. (2003). Bayesian Spatial Prediction for a Gaussian Random Field. Procedings of the Fourth Seminar of the Probability and Stochastic Processes.

[4] Jafari, K. M. and Mohammadzadeh, M. (2004a). Bayesian Spatial Prediction for a Gaussian Random Field with Noisy Observations. To appear in Journal of Science, Isfahan University, Iran.

[5] Jafari, K. M. and Mohammadzadeh, M. (2004b). Bayesian Spatial Prediction for a Transformed Random Field. To appear in Journal of Science, Tehran University, Iran.

[6] Oh, M.S., Shin, W. D. and Kim, J. H. (2002). Bayesian Analysis of Regression Models with Spatially Corelated Errors and Missing Observation. Computational Statistics and Data Analysis, 39: 387-400.

پیش‌بینی بیزی طول عمر برای مدل پارتو با حجم نمونه تصادفی

محسن محمدزاده[۱]، منصور زرگر[۲]

[۱] دانشگاه تربیت مدرس، گروه آمار
[۲] مؤسسه آموزش عالی نجف آباد، گروه آمار

چکیده: یکی از مسائل مهم کاربردی، پیش‌بینی طول عمر یا زمانهای شکست $n-r$ مؤلفه در یک نمونه nتایی بر اساس زمانهای شکست r مؤلفه ترتیبی اول نمونه می‌باشد. چون در اینگونه مسائل اغلب با مواردی مواجه می‌شویم که بدلایل مختلف، از جمله محدودیتهای شرایط آزمایش، حجم نمونه ثابت باقی نمی‌ماند، لازم است پیش‌بینی طول عمر بر اساس حجم نمونه متغیر صورت پذیرد. در این مقاله برای یک مدل طول عمر پارتو پیش‌بینی بیزی در حالتی که حجم نمونه یک متغیر تصادفی با توزیع پواسن یا دوجمله‌ای باشد، مورد بررسی قرار می‌گیرد. سپس فواصل پیش‌بینی بر اساس دو حالت حجم نمونه ثابت و تصادفی با تکنیک شبیه‌سازی مورد مقایسه عددی قرار گرفته، نشان داده می‌شود فواصل پیش‌بینی که بر اساس حجم نمونه تصادفی بدست می‌آیند بطور متوسط دارای طول کوتاهتری هستند و می‌توان با این روش پیش‌بینی‌های دقیقتری را برای زمان شکست هر یک از مؤلفه‌های $r+1$ به بعد نمونه بدست آورد.

واژه‌های کلیدی: مدل طول عمر پارتو، پیش‌بینی بیزی ،اندازه نمونه تصادفی

۱ مقدمه

توزیع پارتو بوسیله پارتو برای توزیع درآمد معرّفی شد. اهمیّت این توزیع به کاربردهای آن در بسیاری از مطالعات اقتصادی اجتماعی به قرن نوزدهم به بعد مربوط می‌شود. توزیع پارتو نقش مهمّی در بررسی اندازهٔ جمعیّت شهرها، حوادث با منابع طبیعی، تغییرات قیمت سهام، توزیعهای درآمد، ریسک بیمه، شکستهای شغلی، خطای خوشه‌بندی کردن در مدارهای ارتباطی و غیره بازی کرده است. آرنولد (۱۹۸۲) یک بررسی تاریخی گسترده‌ای از کاربرد آن در زمینهٔ توزیع درآمد را می‌دهد.

در این مقاله به پیش‌بینی در مدل طول عمر پارتو، که یکی از مدلهای مهم طول عمر است، می‌پردازیم و پیش‌بینی را برای مشاهدات آینده در یک نمونه انجام می‌دهیم (پیش‌بینی تک نمونه‌ای). فرض کنید n شئ را مورد آزمایش عمر قرار داده‌ایم و پس از بدست آمدن r زمان شکست اول آزمون را خاتمه داده مقادیر طول عمر آنها بصورت مقادیر ترتیبی در دسترس هستند (داده‌های سانسور شده نوع II). اکنون قصد داریم بر اساس این r مشاهده ترتیبی برای طول عمر مشاهدات آینده در همین نمونه یعنی $n-r$ حالت باقیمانده پیش‌بینی داشته باشیم. برای

مثال فرض کنید در یک شرکت بیمه n شغل بیمه شده است و پس از گذشت ۶ ماه تعداد r تا از این شغلها ورشکسته شده‌اند و برای گرفتن خسارت به شرکت بیمه مراجعه کرده‌اند. پس مدت زمان فعالیت این r شغل تا زمان ورشکستگی در دست می‌باشد و همچنین فرض کنید مدت زمان فعالیت شغلها تا ورشکستگی از یک مدل طول عمر مانند مدل طول عمر پارتو پیروی می‌کند. حال می‌خواهیم بر اساس این r زمان شکست که بصورت آماره‌های ترتیبی هستند، زمان شکست خوردن $n-r$ حالت باقیمانده را پیش‌بینی کنیم و فاصله اطمینانی برای پیش‌بینی مشاهدات بعدی بدست آوریم و همچنین مهمتر از همه اینکه، زمان شکست خوردن آخرین شغل یعنی $X_{(n)}$ را پیش‌بینی کنیم.

مسئله مهمی که باید مورد توجه قرار گیرد آنست که در اکثر کاربردها، اندازه نمونه ثابت فرض می‌شود. امّا مسأله استنباط آماری در مواقعی که اندازهٔ نمونه یک متغیّر تصادفی است، خیلی مهّم است ، چراکه در عمل نمونه‌های با اندازه تصادفی، در یک حالت طبیعی، بطور فراوان اتفاق می‌افتد. حتی در بعضی کاربردها داشتن اندازهٔ نمونه ثابت غیر ممکن است. زیرا همیشه بعضی از مشاهدات، به دلایل مختلف از بین می‌روند. برای مثال در بعضی از آزمایشات کشاورزی و زیست شناسی، تعدادی از عضوهای نمونهٔ تحت بررسی به دلایل نامعلومی می‌میرند. پس مجبوریم اندازه نمونه را تصادفی فرض کنیم که در اینجا اندازه نمونهٔ یک متغیر تصادفی و دارای یکی از توزیعهای دو جمله‌ای یا پواسن می‌باشد. که در همین بخش بطور توأم، بعد از بدست آوردن فواصل پیش‌بینی با اندازه نمونه ثابت به پیش‌بینی با اندازه نمونه تصادفی می‌پردازیم.

فرض کنید n شئ را مورد آزمون عمر قرار داده‌ایم، که طول عمر آنها از یک توزیع پارتو با تابع توزیع زیر پیروی می‌کند.

$$F(x|\alpha,\sigma) = ۱ - (\frac{\sigma}{x})^{\alpha} \qquad x \geq \sigma > ۰, \alpha > ۰ \qquad (۱)$$

این تابع توزیع دارای دو پارامتر شکل α و مقیاس σ می‌باشد. و تابع بقاء بصورت:

$$\bar{F}(x) = (\frac{\sigma}{x})^{\alpha} \qquad , x \geq \sigma$$

تعریف می‌شود. و تابع چگالی احتمال آن برابر است با:

$$f(x|\alpha,\sigma) = \frac{\alpha\sigma^{\alpha}}{x^{\alpha+۱}} \qquad (۲)$$

اکنون اگر r مشاهده ترتیبی اول در دسترس باشند، بر اساس این r آماره ترتیبی اول از یک نمونه تصادفی از توزیع پارتو (۲)، یعنی $X_{(۱)} \leq X_{(۲)} \leq \ldots \leq X_{(r)}$ ، تابع درستنمایی برای $X = (X_{(۱)}, X_{(۲)}, \ldots, X_{(r)})$ برابر است با:

$$L(\alpha,\sigma;\mathbf{x}) \propto \prod_{i=۱}^{r} f(x_{(i)})\{۱ - F(x_{(r)})\}^{n-r}$$

که برای توزیع پارتو داریم:

$$L(\alpha,\sigma;x) \propto \alpha^r \exp\{-\alpha\{\ln(x_{(r)}^{n-r}\prod_{i=1}^{r} x_{(i)}) - n\ln\sigma\}\} \qquad (3)$$

اگر $Y = X_{r+s}$ مولفهٔ مرتبهٔ $(r+s)$ام باشد و بعنوان sامین مشاهده بعدی تلقی گردد، آنگاه چگالی شرطی Y بشرط x عبارتست از:

$$\begin{aligned} f(y|x) &= \frac{f(y,x)}{f(x)} \\ &= \frac{s\binom{n-r}{s}\{F(y)-F(x_{(r)})\}^{s-1}\{1-F(y)\}^{n-r-s}f(y)}{\{1-F(x_{(r)})\}^{n-r}} \end{aligned} \qquad (4)$$

$y > x_{(r)}$.

و این تابع چگالی شرطی برای توزیع پارتو عبارتست از:

$$f(y|x) = \frac{\alpha s\binom{n-r}{s}}{y}\sum_{j=0}^{s-1}(-1)^j\binom{s-1}{j}(\frac{x_{(r)}}{y})^{\alpha(n-r-s+j+1)} \quad y > x_{(r)} \qquad (5)$$

۲ پیش‌بینی بیزی در مدل طول عمر پارتو

از آنجا که توزیع پارتو دارای دو پارامتر می‌باشد، پیش‌بینی در ۳ حالت زیر انجام می‌گیرد.
الف) حالت پارامتر شکل نامعلوم
ب) حالت پارامتر شکل و مقیاس نامعلوم
ج) حالت بی‌اطلاعی

۱.۲ حالت پارامتر شکل نامعلوم

در حالتی که σ معلوم و α نامعلوم است، اگر در تابع درستنمایی (۳)، $r = n$ قرار دهیم آنگاه آماره $Z = \ln\prod_{i=1}^{n}(\frac{X_i}{\sigma^n})$ برای α بسنده و دارای توزیع $\Gamma(n,\alpha)$ خواهد بود. بنابراین یک خانواده پیشین مزدوج طبیعی برای α را خانواده گاما در نظر می‌گیریم. این مطلب را آرنولد و پرس (۱۹۸۳) مورد بررسی قرار دادند. بنابراین چگالی پیشین بصورت

$$\pi(\alpha|a,b) = \frac{\alpha^{a-1}}{\Gamma(a)}b^a\exp(-b\alpha) \qquad \alpha>0, a>0, b>0 \qquad (6)$$

و چگالی پسین α بصورت

$$\begin{aligned}\pi(\alpha|x,b) &\propto \pi(\alpha|a,b)\cdot f(x|\alpha,\sigma)\\ &= \frac{\alpha^{r+a-1}}{\Gamma(r+a)}H^{r+a}\exp(-\alpha H) \qquad \alpha>0 \end{aligned} \qquad (۷)$$

خواهد بود، که در آن $H=b+\ln(\sigma^{-n}x_{(r)}^{n-r}\prod_{i=1}^{r}x_i)$ است.

با ضرب معادلات (۵) و (۷) و انتگرالگیری روی α، چگالی پیش‌بینی Y بشرط x موقعی که اندازه نمونه ثابت است بصورت

$$\begin{aligned} f(y|x,n) &= \int_\alpha f(y|x,\alpha)\pi(\alpha|x,b)d\alpha\\ &= \int_\alpha f(y,\alpha|x)d\alpha\\ &= \frac{s(r+a)H^{r+a}}{y}\sum_{j=0}^{s-1}\frac{(-1)^j\binom{n-r}{s}\binom{s-1}{j}}{\{H+a_j\ln(\frac{y}{x_{(r)}})\}^{r+a+1}} \end{aligned} \qquad (۸)$$

بدست می‌آید، که در آن $a_j=n-r-s+j+1$ می باشد. تابع بقاء پیش‌بینی Y به شرط x,n عبارتست از:

$$P_r(Y\geq t|x,n)=sH^{r+a}\cdot\sum_{j=0}^{s-1}\frac{(-1)^j\binom{n-r}{s}\binom{s-1}{j}}{a_j\{H+a_j\ln(\frac{t}{x_{(r)}})\}^{r+a}}=\tau \qquad (۹)$$

حال اگر $\tau=(1+\gamma)/2$ و $\tau=(1-\gamma)/2$، آنگاه مرزهای بالایی و پایینی برای $y=x_{(r+s)}$، یعنی sامین مشاهده بعدی موقعی که اندازه نمونه ثابت است، بدست خواهد آمد. حال فرض کنید که اندازه نمونهٔ N، یک متغیّر تصادفی پواسن با تابع جرم احتمال زیر باشد:

$$p(n)=\frac{e^{-\lambda}.\lambda^n}{n!} \qquad , \qquad n=0,1,2,\ldots \qquad (۱۰)$$

با استفاده ازکنسول (۱۹۸۴) وگوپتا وگوپتا (۱۹۸۴) چگالی پیش‌بینی Y بشرط x برای موقعی که اندازه نمونه تصادفی است، بشکل

$$\begin{aligned} f(y|x) &= \frac{1}{P_r(N\geq r+s)}\cdot\sum_{n=r+s}^{\infty}f(y|x,n)P(n)\\ &= \frac{cs(r+a)}{y}\cdot\sum_{n=r+s}^{\infty}\Omega(s,j)\frac{\lambda^nH^{r+a}}{n!\{H+a_j\ln(\frac{y}{x_{(r)}})\}^{r+a+1}} \end{aligned} \qquad (۱۱)$$

است، که در آن $c = \frac{e^{-\lambda}}{1-\sum_{m=0}^{r+s+1}\frac{e^{-\lambda}\lambda^m}{m!}}$

$\Omega(s,j) = \binom{n-r}{s}\sum_{j=0}^{s-1}(-1)^j\binom{s-1}{j}$ و $a_j = n-r-s+j+1$

و تابع بقاء پیش‌بینی برای Y بشرط x عبارتست از:

$$P_r(Y \geq t|x) = cs \cdot \sum_{n=r+s}^{\infty} \frac{\lambda^n H^{r+a}\Omega(s,j)}{a_j n!\{H + a_j \ln(\frac{t}{x_{(r)}})\}^{r+a}} \qquad (۱۲)$$

برای بدست آوردن فاصله پیش‌بینی ۱۰۰γ درصد برای y، روشهای عددی تکراری لازم هستند، که با پیدا کردن t از معادله (۱۲) برای یک مقدار $\tau = P_r(Y \geq t)$ داده شده با $\tau = (1+\gamma)/2$ و $\tau = (1-\gamma)/2$، قابل محاسبه هستند. برای پیش‌بینی زمان شکست اولین مشاهده بعدی یعنی $y = x_{(r+1)}$ داریم:

$$\begin{aligned} P_r(Y \geq t|x) &= \frac{e^{-\lambda}}{1-\sum_{m=0}^{r}\frac{e^{-\lambda}\lambda^m}{m!}} \\ &\cdot \sum_{n=r+1}^{\infty}\frac{\lambda^n}{n!}\{1+\frac{(n-r)\ln(\frac{t}{x_{(r)}})}{H}\}^{-(r+a)} \end{aligned} \qquad (۱۳)$$

و اگر اندازهٔ نمونه N یک متغیّر تصادفی از توزیع دوجمله‌ای با پارامترهای M و p باشد، آنگاه

$$f(y|\sigma,x) = \frac{ks(r+a)}{y}\cdot\sum_{n=r+s}^{M}\Omega(s,j)\frac{\binom{n}{M}p^n q^{M-n}H^{r+a}}{\{H+a_j\ln(\frac{y}{x_{(r)}})\}^{r+a+1}}$$

که در آن $k^{-1} = 1-\sum_{W=0}^{r+s-1}\binom{M}{W}p^W q^{M-W}$. و تابع بقاء پیش‌بینی

$$P_r(Y \geq t|\sigma,x) = ks\sum_{n=r+s}^{M}\Omega(s,j)\frac{\binom{M}{n}p^n q^{M-n}H^{r+a}}{a_j\{H+a_j\ln(\frac{t}{x_{(r)}})\}^{r+a}} \qquad (۱۴)$$

می‌باشد. با قرار دادن $s = 1$، در معادلهٔ (۱۴)، تابع بقاء پیش‌بینی $y = x_{(r+1)}$ عبارتست از:

$$\begin{aligned} P_r(x_{(r+1)} \geq t|x) &= \frac{1}{1-\sum_{W=0}^{r}\binom{M}{W}p^W q^{n-W}} \\ &\cdot \sum_{n=r+1}^{M}\frac{\binom{M}{n}p^n q^{M-n}H^{r+a}}{\{H+(n-r)\ln(\frac{t}{x_{(r)}})\}^{r+a}} \end{aligned} \qquad (۱۵)$$

همچنین برای بدست آوردن فاصله پیش‌بینی y، روشهای عددی تکراری لازم هستند.

۲.۲ حالت پارامترهای مقیاس و شکل نامعلوم

در این بخش، در حالتی که هر دو پارامتر مقیاس و شکل مجهولند، فاصله پیش‌بینی محاسبه می‌شود. در این حالت $X_{(1)} = Min(X_1, X_2, \ldots, X_n)$ و $\sum_{i=1}^{r} \ln X_i = \ln \prod_{i=1}^{r} X_i$ برای σ و α، آماره‌های بسنده توأم مینیمال هستند. در نتیجه توزیع پیشین مزدوج توأم طبیعی تعمیم یافته توسط آرنولد وپرس (۱۹۸۳) را بعنوان پیشین برای α و σ استفاده می‌کنیم، که در آن فرض می‌شود α دارای توزیع گاما است و توزیع شرطی σ بشرط α از شکل تابع توانی می‌باشد. این پیشین بفرم

$$\pi(\sigma, \alpha) = g_1(\sigma|\alpha) \cdot g_2(\alpha)$$

است، که در آن

$$g_1(\sigma|\alpha) = \zeta\alpha\sigma^{\zeta\alpha-1}\sigma_0^{-\zeta\alpha} \qquad , 0 < \sigma < \sigma_0$$

و

$$g_2(\alpha) = \frac{\alpha^{\delta-1}}{\Gamma(\delta)}\{\ln\mu - \zeta \ln\sigma_0\}^{\delta}.\exp\{-\alpha(\ln\mu - \zeta\ln\sigma_0)\}$$

یعنی

$$g_2(\alpha) \sim \Gamma(\delta, \ln\mu - \zeta\ln\sigma_0)$$

بنابراین

$$\pi(\sigma, \alpha) = A^{-1}.\alpha^{\delta}.\sigma^{\zeta\alpha-1}.\mu^{-\alpha} \qquad , 0 < \sigma < \sigma_0 \quad , \alpha > 0 \qquad (۱۶)$$

و پارامترهای $\delta, \zeta, \mu, \sigma_0$ مثبتند و A تعدیل کنندهٔ ثابتی است که بوسیلهٔ فرمول زیر تعیین می‌شود

$$A = \zeta^{-1}.\Gamma(\delta).\{\ln\mu - \zeta\ln\sigma_0\}^{-\delta}$$

پیشین (۱۶) توزیع گاما- توانی نامیده می‌شود و با نماد $P\Gamma(\delta, \zeta, \mu, \sigma_0)$ نشان داده می‌شود و نتایج بدست آمده در بخش قبل را می‌توان با در نظر گرفتن $\delta = \zeta = 0, \mu = 1, \sigma_0 = \infty$ بدست آورد.

با بکارگیری قضیّه بیزی و با استفاده از تابع درستنمائی (۳)، ملاحظه می‌شود که چگالی پسین σ, α بشرط x در همان خانوادهٔ گاما - توانی با پارامترهای $(r + \delta, n + \zeta, \mu(x_{(r)}^{n-r}\prod_{i=1}^{r} x_i), Min(\sigma_0, x_1))$ قرار دارد. بنابراین:

$$\pi(\sigma, \alpha|x) = A'^{-1}\alpha^{r+\delta} \cdot \left(\frac{\sigma^{n+\zeta-\frac{1}{\alpha}}}{\mu.x_{(r)}^{n-r}.\prod_{i=1}^{r} x_i}\right)^{\alpha} = \pi(\sigma, \alpha).f(x|\sigma, \alpha)$$

$$.\alpha > 0 \quad , 0 < \sigma < L^*$$

بعد از مقداری محاسبات ریاضی، چگالی پسین می‌تواند بشکل

$$\pi(\sigma,\alpha|x)=\frac{n+\zeta}{\Gamma(r+\delta)}.(\alpha u)^{r+\delta}.\exp(-\alpha v)\qquad \alpha>0,\ \ 0<\sigma<L^{*}\quad (۱۷)$$

نوشته می‌شود که در آن

$$u=\ln\{(\mu.x_{(r)}^{n-r}.\prod_{i=1}^{r}x_i)/L^{*n+\zeta}\}$$

و

$$v=\ln\{(\mu.x_{(r)}^{n-r}.\prod_{i=1}^{r}x_i)/\sigma^{n+\zeta-\frac{1}{\alpha}}\}$$

با ضرب معادلات (۵) و (۱۷) و انتگرالگیری روی σ,α، تابع چگالی پیش‌بینی بیزی Y بشرط x با اندازه نمونه ثابت بشکل زیر می‌باشد:

$$f(y|x)=\frac{s(r+\delta)\binom{n-r}{s}u^{r+\delta}}{y}\sum_{j=0}^{s-1}\frac{(-1)^j\binom{s-1}{j}}{\{u+a_j\ln(\frac{y}{x_{(r)}})\}^{r+\delta+1}}\quad y>x_{(r)}\quad (۱۸)$$

چگالی پیش‌بینی Y موقعی که اندازه نمونه N یک متغیّر تصادفی با تابع جرم احتمال (۱۰) می‌باشد، برابر است با:

$$f(y|x)=\frac{cs(r+\delta)}{y}\cdot\sum_{n=r+s}^{\infty}\Omega(s,j).\frac{\lambda^n.u^{r+\delta}}{n!\{u+a_j\ln(\frac{y}{x_{(r)}})\}^{r+\delta+1}}\qquad (۱۹)$$

و تابع بقاء پیش‌بینی عبارتست از:

$$P_r(Y\geq t|x)=cs\sum_{n=r+s}^{\infty}\Omega(s,j)\frac{1}{n!a_j}\{\frac{u}{u+a_j\ln(\frac{t}{x_{(r)}})}\}^{r+\delta}\qquad (۲۰)$$

برای $s=1$، تابع بقاً پیش‌بینی $Y=X_{(r+1)}$ عبارتست از:

$$\begin{aligned}P_r(x_{(r+1)}\geq t|x) &= \frac{e^{-\lambda}}{1-\sum_{m=0}^{r}\frac{e^{-\lambda}\lambda^m}{m!}}\\ &\cdot\sum_{n=r+1}^{\infty}\frac{\lambda^n}{n!}\cdot\{\frac{u}{u+a_j\ln(\frac{t}{x_{(r)}})}\}^{r+\delta}\end{aligned}\qquad (۲۱)$$

دوباره با این فرض که اندازهٔ نمونهٔ N، یک متغیّر تصادفی با توزیع دوجمله‌ای است، تابع چگالی پیش‌بینی Y بصورت

$$f(y|x) = \frac{ks.(r+\delta)}{y} \cdot \sum_{n=r+s}^{M} \Omega(s,j) \frac{\binom{M}{n} p^n q^{M-n} u^{r+\delta}}{\{u + a_j \ln(\frac{y}{x_{(r)}})\}^{r+\delta+1}} \qquad (۲۲)$$

است و تابع بقأ پیش‌بینی عبارتست از:

$$P_r(Y \geq t|x) = ks \sum_{n=r+s}^{M} \Omega(s,j) \frac{\binom{M}{n} p^n q^{M-n}}{a_j} \cdot \{\frac{u}{u + a_j \ln(\frac{t}{x_{(r)}})}\}^{r+\delta} \qquad (۲۳)$$

با قراردادن $s = 1$ در معادله (۲۳)، فاصله پیش‌بینی 100γ درصد برای $x_{(r+1)}$، جوابهای t از معادلات

$$\frac{1}{1 - \sum_{W=0}^{r} \binom{W}{n} p^W q^{M-W}} \sum_{n=r+1}^{M} \binom{M}{n} p^M q^{M-n} \{\frac{u}{u + (n-r)\ln(\frac{t}{x_{(r)}})}\}^{r+\delta} = \tau$$

بدست می‌آید که در آن $\tau = \frac{1-\gamma}{2}$ و $\tau = \frac{1+\gamma}{2}$ می‌باشند.

۳.۲ بدون اطلاعات قبلی از α, σ

یک پیشین بدون اطلاع (NIP) پیشینی است که اثراتش بر همهٔ مقادیر پارامترها بی‌اهمیّت است. و هیچ اطلاعی به آنچه بوسیله داده‌های تجربی بدست آمده است، نمی‌افزاید. بنابراین، عموماً، استنباط بیزی بر اساس یک NIP فقط جنبهٔ نظری دارد و اینچنین استنباط بیزی، در حل مسائل پیش‌بینی، موقعی که پیدا کردن جوابهای کلاسیک فوق العاده مشکل می‌باشد، مهّم است. زیرا تقریب بیزی بر اساس یک NIP می‌تواند شبیه یک روش ریاضی برای بدست آوردن فاصله پیش‌بینی کلاسیک عمل کند. با استفاده از استدلال باکس و تیو (۱۹۷۳) و نیو و همدی (۱۹۸۷) می‌توان نشان داد که NIP برای پارامترهای α, σ بشکل زیر است:

$$\pi(\sigma, \alpha) \propto \frac{1}{\sigma.\alpha} \quad , \sigma > 0 \qquad (۲۴)$$

چگالی پسین توأم α, σ بشرط x عبارتست از:

$$\begin{aligned} \pi(\sigma, \alpha|x) &= f(x|\alpha, \sigma).\pi(\sigma, \alpha) \\ &= \frac{n(\alpha z)^{r-1}.\sigma^{n\alpha-1}}{\Gamma(r-1).x_{(r)}^{(n-r).\alpha}.\prod_{i=1}^{r} x_i^{\alpha}} \qquad , \sigma < x_{(1)} \end{aligned} \qquad (۲۵)$$

که در آن $z = \sum_{i=1}^{r} \ln x_i + \ln(x_{(r)}^{n-r}/x_{(1)}^{n})$.

همانند بخش قبلی می‌توان با ضرب معادلات (۵) و (۲۵) و انتگرالگیری روی σ و α، چگالی پیش‌بینی $y = x_{(r+s)}$ را بصورت

$$f(y|x) = \frac{cs(r-1)}{y} \cdot \sum_{n=r+s}^{\infty} \Omega(s,j) \cdot \frac{z^{r-1}}{n!\{z + a_j \ln(y/x_{(r)})\}^{r}} \qquad (۲۶)$$

بدست آورد، که متناظراً تابع بقاء پیش‌بینی بفرم

$$P_r(Y \geq t|x) = cs \sum_{n=r+s}^{\infty} \Omega(s,j) \cdot \frac{\lambda^n z^{r-1}}{a_j.n!\{z + a_j \ln(t/x_{(r)})\}^{r-1}} \qquad (۲۷)$$

خواهد بود. برای $s = 1$، تابع بقاً پیش‌بینی $x_{(r+1)}$ عبارتست از:

$$\begin{aligned} P_r(x_{(r+1)} \geq t|x) &= \frac{e^{-\lambda}}{1 - \sum_{m=0}^{r} \frac{e^{-\lambda}\lambda^m}{m!}} \\ &\cdot \sum_{n=r+1}^{\infty} \frac{\lambda^n}{n!}\{1 + \frac{(n-r)\ln(\frac{t}{x_{(r)}})}{z}\}^{r-1} \end{aligned} \qquad (۲۸)$$

دوباره اگر اندازه نمونهٔ N، دارای توزیع دوجمله‌ای باشد، تابع چگالی پیش‌بینی Y بشرط x بصورت

$$f(y|x) = \frac{ks(r-1)}{y} \sum_{n=r+s}^{M} \Omega(s,j) \frac{\binom{M}{n} p^n q^{M-n} z^{r-1}}{\{z + a_j \ln(\frac{y}{x_{(r)}})\}^{r}} \qquad (۲۹)$$

خواهد شد و تابع بقاء پیش‌بینی عبارتست از:

$$P_r(Y \geq t|x) = ks \sum_{n=r+s}^{M} \Omega(s,j) \frac{\binom{M}{n} p^n q^{M-n} z^{r-1}}{a_j\{z + a_j \ln(\frac{t}{x_{(r)}})\}^{r-1}} \qquad (۳۰)$$

برای $s = 1$، تابع بقاء پیش‌بینی $x_{(r+1)}$ برابر است با:

$$\begin{aligned} P_r(x_{(r+1)} \geq t|x) &= \frac{1}{1 - \sum_{W=0}^{r} \binom{M}{W} p^W q^{M-W}} \\ &\cdot \sum_{n=r+1}^{\infty} \frac{\binom{M}{n} p^n q^{M-n} z^{r-1}}{\{z + (n-r)\ln(\frac{t}{x_{(r)}})\}^{r-1}} \end{aligned} \qquad (۳۱)$$

مثال ۱: حالت پارامتر شکل نامعلوم
فرض کنید مدت زمانی که یک شغل فعّالیّت می‌کند تا شکست بخورد بر حسب سال توزیع پارتو با $\sigma = ۱$ دارد و ۱٫۰۱، ۱٫۰۵، ۱٫۰۸، ۱٫۱۴، ۱٫۲۸، ۱٫۳۰، ۱٫۳۳، ۱٫۴۳، ۱٫۵۹ و ۱٫۶۲ طول عمر مفید ۱۰ نمونهٔ تصادفی از اینچنین شغلی باشند. اگر فقط شش زمان شکست مرتب شدهٔ اوّل در اختیار باشند، علاقمند به محاسبه فاصله پیش‌بینی برای طول عمر اولین شغل از $n - r$ تای باقیمانده هستیم (یعنی $x_{(۷)}$). جدول ۱، فواصل پیش‌بینی ۹۵ درصد برای $x_{(۷)}$ را نشان می‌دهد که با حل کردن معادلات (۹)، (۱۳) و (۱۵) بطور تکراری بدست آمده‌اند. با این فرض که توزیع پیشین α، گاما با مقادیر مختلف $a = ۱, ۵$ و $b = ۱, ۲, ۵$ است. برای یک اندازه نمونهٔ تصادفی، با توزیعهای پواسن و دوجمله‌ای با پارامترهای مختلف $\lambda = ۱, ۲۰, ۵۰$ و $(M, P) = (۲۰, ۰/۸), (۵۰, ۰/۵), (۵۰۰, ۰/۱)$، محاسبات را انجام می‌دهیم. مقادیر مختلف P, M, λ و پارامترهای پیشین برای بررسی حساسیت فواصل به این پارامترها، انتخاب شده‌اند.

مثال ۲: حالت پارامتر مقیاس و شکل نامعلوم
برای تشریح نتایج بخش ۲.۲، از همان داده‌های مثال ۱ استفاده می‌کنیم (که α, σ نامعلومند). و با حل معادلات (۱۸)، (۲۱)، (۲۳)، (۲۸) و (۳۱) بطور عددی فاصله پیش‌بینی ۹۵ درصد برای $x_{(۷)}$ را با استفاده از همان مقادیر P, M, λ، بدست می‌آوریم. پارامتر توزیعهای پیشین (شامل NIP) با مقادیر مختلف $L^* = x_{(۱)}$ و $\delta = ۱, ۵, ۱۰$ و $\mu = ۳, ۷, ۳۰$ و $\zeta = ۲, ۶, ۲۰$ بکار گرفته شده‌اند و نتایج در جدول ۲ آمده است.

۳ مقایسهٔ پیش‌بینی بیزی با اندازه‌های نمونه ثابت و تصادفی

۱.۳ پیش‌بینی در توزیع پارتو با پارامترهای α نامعلوم و σ معلوم

فرض کنید طول عمر اشیاء مورد آزمایش از توزیع پارتو با تابع توزیع

$$F(x|\alpha, \sigma) = ۱ - (\frac{\sigma}{x})^{\alpha} \qquad ۰ < \sigma \leq x, \alpha > ۰$$

و تابع چگالی احتمال

$$f(x|\alpha, \sigma) = \frac{\alpha \sigma^{\alpha}}{x^{\alpha+۱}}$$

پیروی می‌کنند، که در آن پارامتر α نامعلوم و σ معلوم می‌باشد. اندازه نمونه اولیه $n = ۱۰$ در نظر گرفته می‌شود و پس از بدست آمدن طول عمر $r = ۶$ شئ اول آزمایش خاتمه می‌یابد (داده‌های سانسور شده نوع II). اکنون قصد داریم با فرض ثابت بودن اندازه نمونه، برای اولین مشاهده بعدی، یعنی $x = x_{(۷)}$، بر اساس r مشاهده اول، پیش‌بینی فاصله‌ای داشته باشیم (پیش‌بینی تک نمونه‌ای). در روش شبیه‌سازی ابتدا باید ۶ عدد از توزیع پارتو، بعنوان مقادیر مشاهده شده

جدول ۱: فاصله پیش‌بینی ۹۵ درصد برای $x_{(۷)}$ موقعی که α نامعلوم و σ معلوم

پیشین		مرزها	اندازه نمونه ثابت ($n = ۱۰$)	نتایج برای یک اندازه نمونه تصادفی (پواسن) با پارامتر λ			نتایج برای یک اندازه نمونه تصادفی (دو جمله‌ای) با پارامترهای زیر M, P مقادیر			
a	b			$\lambda = ۱$	$\lambda = ۲۰$	$\lambda = ۵۰$	$M = ۲۰$ $P = ۰/۸$	$M = ۵۰$ $P = ۰/۵$	$M = ۵۰۰$ $P = ۰/۱$	
۱	۲	LB	۱٫۳۱	۱٫۳۳	۱٫۳۱	۱٫۳۰	۱٫۳۱	۱٫۳۰	۱٫۳۰	
		UB	۲٫۱۷	۶٫۱۳	۱٫۷۰	۱٫۵۵	۱٫۷۵	۱٫۶۲	۱٫۵۵	
۱	۵	LB	۱٫۳۲	۱٫۳۵	۱٫۳۱	۱٫۳۰	۱٫۳۱	۱٫۳۱	۱٫۳۰	
		UB	۳٫۲۳	۲۸٫۲۰	۱٫۹۴	۱٫۶۱	۲٫۰۷	۱٫۷۷	۱٫۶۱	
۵	۱	LB	۱٫۳۰	۱٫۳۱	۱٫۳۰	۱٫۳۰	۱٫۳۰	۱٫۳۰	۱٫۳۰	
		UB	۱٫۶۲	۲٫۴۰	۱٫۴۸	۱٫۴۳	۱٫۵۰	۱٫۴۶	۱٫۴۳	

جدول ۲: فاصله پیش‌بینی ۹۵ درصد برای $x_{(۷)}$ و موقعی که α, σ نامعلومند

پیشینها با $L^* = x_{(۱)}$			مرزها	اندازه نمونه ثابت ($n = ۱۰$)	نتایج برای یک اندازه نمونه تصادفی (پواسن) با پارامتر λ			نتایج برای یک اندازه نمونه تصادفی (دو جمله‌ای) با پارامترهای زیر مقادیر M, P			
δ	ζ	μ			$\lambda = ۱$	$\lambda = ۲۰$	$\lambda = ۵۰$	$M = ۲۰$ $P = ۰/۸$	$M = ۵۰$ $P = ۰/۵$	$M = ۵۰۰$ $P = ۰/۱$	
۱	۲	۳	LB	۱٫۳۱	۱٫۳۲	۱٫۳۰	۱٫۲۷	۱٫۳۰	۱٫۳۰	۱٫۳۰	
			UB	۱٫۹۰	۳٫۷۰	۱٫۶۱	۱٫۴۷	۱٫۶۵	۱٫۵۷	۱٫۵۲	
۵	۶	۷	LB	۱٫۳۱	۱٫۳۲	۱٫۳۰	۱٫۲۸	۱٫۳۰	۱٫۳۰	۱٫۳۰	
			UB	۱٫۷۳	۳٫۰۶	۱٫۵۱	۱٫۴۰	۱٫۵۳	۱٫۴۷	۱٫۴۳	
۱۰	۲۰	۳۰	LB	۱٫۳۱	۱٫۳۲	۱٫۳۰	۱٫۲۸	۱٫۳۰	۱٫۳۰	۱٫۳۰	
			UB	۱٫۶۸	۲٫۹۶	۱٫۴۷	۱٫۳۸	۱٫۴۹	۱٫۴۳	۱٫۴۰	
NIP			LB	۱٫۳۱	۱٫۳۱	۱٫۳۰	۱٫۳۰	۱٫۳۰	۱٫۳۰	۱٫۳۰	
			UB	۱٫۸۵	۲٫۷۹	۱٫۶۸	۱٫۶۲	۱٫۷۰	۱٫۶۵	۱٫۶۲	

تولید کنیم. برای این امر ابتدا از توزیع یکنواخت $U(۰,۱)$، ۶ مقدار $u_۱, u_۲, \ldots, u_۶$ را تولید می‌کنیم، سپس بر اساس قضیه احتمال انتگرال با گرفتن $F_X(X) = U$ و بدست آوردن معکوس این تابع بصورت $F_X^{-۱}(U) = X$، مقادیر توزیع پارتو از

$$x = \sigma \cdot (۱ - u)^{-۱/\alpha}$$

حاصل می‌شود. اکنون برای $\alpha = ۷$، $\sigma = ۱$، ۶ مقدار برای x بدست می‌آید، که از توزیع پارتو پیروی می‌کنند. اکنون با استفاده از معادله (۹) با قرار دادن دو مقدار $\tau = ۰/۹۵$ و $\tau = ۰/۰۵$ و با قرار دادن $s = ۱$ و حل آن با نرم‌افزار MATLAB، دو مقدار برای $x = x_{(۷)}$ بدست می‌آید، که همان کران بالا و پائین فاصله پیش‌بینی برای $x_{(۷)}$، موقعی که اندازه نمونه ثابت می‌باشد، هستند. بهمین طریق با استفاده از داده‌های تولید شده و با فرض اینکه اندازه نمونه متغیر تصادفی با تابع جرم احتمال پواسن یا دوجمله‌ای می‌باشد، می‌توان با استفاده از معادلات (۱۳) و (۱۵) فاصله پیش‌بینی برای اولین مشاهده بعدی $x_{(۷)}$ را وقتی اندازه نمونه تصادفی است، بدست آورد. نتایج بدست آمده در جدول ۳ خلاصه شده‌اند، که نتایج زیر حاصل می‌شوند.

۱ــ با زیاد شدن مقدار λ، فواصل پیش‌بینی کوتاهتر می‌شوند.

۲ــ با زیاد شدن مقدار M، فواصل پیش‌بینی کوتاهتر می‌شوند.

۳ــ برای $\lambda = ۱$، طول فواصل پیش‌بینی با اندازه نمونه ثابت، کوتاهتر و بهتر از از اندازه نمونه تصادفی می‌باشد. اما برای $\lambda = ۲۰$ و $\lambda = ۵۰$ طول فواصل با اندازه نمونه تصادفی کوتاهتر می‌باشد.

۴ــ با زیاد شدن مقدار a، فواصل کوتاهتر می‌شوند.

۵ــ با زیاد شدن مقدار b، فواصل بزرگتر می‌شوند.

۲.۳ پیش‌بینی در توزیع پارتو با پارامترهای α و σ نامعلوم و پیشین بی‌اطلاع

در این بخش همانند بخش قبل ابتدا از توزیع پارتو ۶ عدد تولید می‌کنیم و بعد با فرض آنکه α و σ هر دو نامعلومند، برای اولین مشاهده بعدی، فواصل پیش‌بینی را بدست می‌آوریم. همچنین فرض می‌کنیم که هیچ اطلاع قبلی از پارامترهای توزیع نداریم و از پیشین بدون اطلاع استفاده می‌کنیم. میانگین فواصل پیش‌بینی شبیه‌سازی شده برای مقادیر مختلف λ را با استفاده از معادلات (۱۸)، (۲۱) و (۲۸) بدست می‌آوریم. نتایج این بررسی در جدول ۴ آمده و نتایج زیر حاصل می‌شوند.

۱ــ با زیاد شدن مقدار λ، فواصل کوتاهتر می‌شوند.

۲ــ برای λهای بزرگ، فواصل پیش‌بینی با اندازه نمونه تصادفی بهتر از اندازه نمونه ثابت است.

جدول ۳: میانگین طول فواصل پیش‌بینی برای $x = x_{(۷)}$ در توزیع پارتو با پارامتر نامعلوم α و با اندازه نمونه ثابت و تصادفی

پارامترهای توزیع پیشین		اندازه نمونه ثابت	نتایج برای اندازه نمونه تصادفی با توزیع دوجمله‌ای و M, p با مقادیر λ			نتایج برای اندازه نمونه تصادفی با توزیع پواسن و λ با مقادیر			
a	b	$n = ۱۰$	$\lambda = ۱$	$\lambda = ۲۰$	$\lambda = ۵۰$	$M = ۲۰$ $p = ۰/۸$	$M = ۵۰$ $p = ۰/۵$	$M = ۱۰۰$ $p = ۰/۵$
۱	۱	۰٫۸۵۰	۶٫۳۱۴	۰٫۴۳۹	۰٫۱۸۹	۰٫۷۳۷	۰٫۲۷۱	۰٫۰۹۵
۱	۲	۱٫۱۶۵	۱۴٫۷۳۶	۰٫۴۹۲	۰٫۲۸۵	۰٫۷۷۰	۰٫۵۳۳	۰٫۱۱۰
۱	۵	۲٫۴۸۸	۷۶٫۴۵۳	۰٫۸۲۷	۰٫۳۵۵	۰٫۷۵۸	۰٫۴۸۹	۰٫۲۲۴
۵	۱	۰٫۴۳۷	۳٫۳۲۱	۰٫۳۰۷	۰٫۳۰۳	۰٫۶۹۴	۰٫۲۶۷	۰٫۱۳۱
۵	۲	۰٫۶۰۷	۵٫۳۱۷	۰٫۳۴۰	۰٫۳۱۵	۰٫۶۳۶	۰٫۳۵۴	۰٫۱۶۰
۵	۵	۰٫۹۹۶	۱۳٫۲۱۰	۰٫۵۰۹	۰٫۳۳۰	۰٫۵۱۷	۰٫۴۲۶	۰٫۲۵۲

جدول ۴: میانگین طول فواصل پیش‌بینی برای $x = x_{(۷)}$ در توزیع پارتو با پارامتر نامعلوم α و σ و با اندازه نمونه ثابت و تصادفی موقعی که پیشین بی‌اطلاع داشته باشیم.

پیشین بی‌اطلاع	اندازه نمونه ثابت	اندازه نمونه تصادفی پواسن با مقادیر λ			
	$n = ۱۰$	$\lambda = ۱$	$\lambda = ۱۰$	$\lambda = ۲۰$	$\lambda = ۵۰$
NIP	۰/۷۷۶	۱۰/۵۵۹	۱/۱۸۳	۰/۲۳۷	۰/۰۷۱

۴ بحث و نتیجه‌گیری

در مدل طول عمر پارتو، میانگین طول فواصل پیش‌بینی برای مشاهدات بعدی در یک نمونه که با استفاده از روش شبیه‌سازی بدست می‌آیند، به پارامترهای توزیع پیشین و توزیع اندازه نمونه بستگی دارند. به همین دلیل در حالت کلی نمی‌توان هر یک از دو روش پیش‌بینی بر اساس اندازه نمونه ثابت و تصادفی را بر دیگری ترجیح داد. وقتی اندازه نمونه متغیری با توزیع پواسن $p(\lambda)$ در نظر گرفته می‌شود، برای مقادیر بزرگ λ، طول فواصل پیش‌بینی حاصل از روش اندازه نمونه تصادفی، کوتاهتر از طول فواصل پیش‌بینی حاصل از اندازه نمونه تصادفی می‌باشد. دلیل این تفاوت می‌تواند ناشی از بزرگتر شدن حجم نمونه ثانوی برای مقادیر بزرگ λ باشد. همچنین بر اساس نتایج بدست آمده، فواصل پیش‌بینی به پارامترهای توزیع پیشین حساس هستند. برای توزیع پارتو با پیشین $Gamma(a, b)$، وقتی a زیاد می‌شود، فواصل پیش‌بینی برای اولین مشاهده ثانوی کوتاهتر و وقتی b زیاد می‌شود، فواصل طولانی‌تر می‌شوند.

مراجع

[1] Arnold, B. C. (1982), Pareto Distribution, International Co-Operative Publishing House.

[2] Arnold, B. C. and Press, S. J. (1983), Bayesian Inference for Pareto Populations, J. of Econometrics, 21, 287-306.

[3] Arnold, B. C. and Press, S. J. (1989), Baysian Estimation and Prediction for Pareto Data, J. Amer. Statist. Assoc., 84, 1079-1084.

[4] Consul, P. C. (1984), On the Distribution of Order Statistics for Random Sample Size, Statist. Netherlands, 38, 249-256.

[5] Dunsmore, I. R. (1974), The Baysian Predictive Distribution in Life Testing Models, Journal of Technometrics, 16, 455-460.

[6] Epsten, B. and Sobel, M. (1953), Life Testing, J. Amer. Statist. Assoc, 48, 486-502.

[7] Faulkenberry, D. G. (1973), A Method of Obtaining Prediction Intervals, J. Amer. Statist. Assoc. , 68, 433-435.

[8] Gupta, D. and Gupta, R. C. (1984), On the Distribution of Order Statistics for a Random Sample Size, Statist. Netherlands, 38, 13-19.

[9] Hewett, J. E. (1968), A Note on Prediction Intervals Based on Partial Observations in Certain Life Test Experiments, Technometrics, 10, 850-853.

[10] Kulldorff, G. and Vannman, K. (1973), Estimation of the Location and Scale Parameters of a Pareto Distribution by Linear Functions of Order Statistics, J. Amer. Statist. Assoc., 68, 218-227.

[11] Nigm, A. M. and Hamdy, H. I. (1987), Baysian Prediction Bounds for the Pareto Lifetime Model, Commun. Statist., Theory and Methods, 16, 1761-1772.

[12] Soliman, A. A. (2000), Bayes Prediction in a Pareto Lifetime Model with Random Sample Size, The Statistician, 49, Part 1, 51-62.

شبیه‌سازی احتمالات ورشکستگی در فرآیندهای مخاطره بیمه

نادر مظاهری

کارشناس ارشد آمار شرکت توزیع برق شهرستان اصفهان

چکیده: یک مساله مهم در امور بیمه تعیین احتمالات ورشکستگی است. محاسبه احتمالات ورشگستگی معمولاً از روشهای تحلیلی میّسر نبوده و به همین علّت شبیه سازی این احتمالات دارای اهمّیت می‌باشد. ورشکستگی پیشامدی با احتمال کم بوده، شبیه‌سازی معمولی آن به دلیل اینکه نیازمند تولید مسیرهای نمونه‌ای بسیار زیاد است دارای کارائی لازم نمی‌باشد و شبیه‌سازی باید با روشهای خاصّی انجام شود. در این مقاله به یک نوع از این روشها با نام نمونه‌گیری ویژه خواهیم پرداخت. در فرآیندهای بیمه با اندازه‌های خسارت زیر نمائی یک نمونه‌گیری ویژه تغییر اندازه احتمال بصورت چرخش زیر نمائی می‌باشد. در این مقاله شرط‌هائی برای بهینه بودن مجانبی برآوردکننده‌های ناشی از این نوع شبیه‌سازی برای احتمالات ورشکستگی با افق زمانی متناهی و نامتناهی را شرح می‌دهیم و با معرّفی دو مدل پواسون مرکب و مارکف ـ مدول پواسون مرکب به شبیه‌سازی احتمال ورشکستگی در هر یک از این دو مدل می‌پردازیم.

واژه‌های کلیدی: احتمال ورشکستگی، پیشامد کم محتمل، چرخش نرخ خطر، چرخش نرخ خطر تأخیری وزنی، فرایند مخاطره بیمه، کارائی شبیه‌سازی، نمونه‌گیری ویژه

۱ مقدمه

فرض کنیم X_i اندازه خسارت nام در زمان ورود η_n $(n \geq ۱)$با $۰ = \eta_۰ < \eta_۱ < ...$ باشد. زمان بین ورود nامین و $n-۱$امین ادعا را با $\xi_n = \eta_n - \eta_{n-۱}$ نشان می‌دهیم. اگر $N(t)$ تعداد تصادفی از ادعاها در فاصله زمانی $[۰, T)$ باشد فرایند مخاطره بیمه را به صورت زیر در نظر می‌گیریم:

$$U(t) = u + \Pi(t) - \sum_{i=۱}^{N(t)} X_i$$

که در آن $\Pi(t), u$ به ترتیب نشان دهنده ذخیره اولیه و تابع حق بیمه می‌باشند. در این مقاله تابع حق بیمه را به صورت ct با نرخ c در نظر می‌گیریم. برای اطلاعات بیشتر از خواص تابع حق بیمه می‌توان به اسموزن [۱] (۲۰۰۰) مراجعه کرد. دیده می‌شود که $U(t)$ ذخیره تا زمان t است.

1) asmussen

اگر فرایند مازاد خسارت را به صورت $u - U(t)$ تعریف کنیم و آن را با $S(t)$ نشان دهیم در این صورت احتمال ورشکستگی با افق زمانی متناهی T به صورت زیر داده می‌شود:

$$\psi(u,T) = P\left(U(t) < ۰, t < T\right) = \left(\sup_{t<T} S(t) > u\right) \qquad (۱)$$

همچنین احتمال ورشکستگی با افق نامتناهی را به صورت $\psi(u) = \psi(u,\infty)$ تعریف می‌کنیم.
فرض کنیم با احتمال یک

$$\lim_{t\to\infty} \frac{\sum_{i=۱}^{N(t)} X_i}{t} = w\beta \quad , \quad \lim_{t\to\infty} \frac{N(t)}{t} = w$$

وقتی فرایند مازاد خسارت $S(t)$ با رانش منفی است (یعنی $S(t) \to -\infty$ وقتی $t \to \infty$) احتمال ورشکستگی با افق نامتناهی کمتر از ۱ خواهد بود که در این صورت $c > w\beta$ یا $\rho = \frac{w\beta}{c} < ۱$. اگر $S(t)$ با رانش غیر منفی باشد با توجه به تعریف (۱) ورشکستگی با افق متناهی با احتمال ۱ رخ می‌دهد.

در ساده‌ترین حالات نیز احتمالات ورشکستگی به صورت صریح امکان‌پذیر نمی‌باشد و اغلب در تقریب زدن آن به روشهای گوناگون سعی شده است. بخشی از این تقریب‌ها شامل برآوردهای ناشی از شبیه‌سازی می‌باشد. باید توجه داشت که وقتی $\psi(u) \to ۰$, $u \to \infty$ و $\psi(u,T) \to ۰$ یعنی با زیاد شدن u احتمال ورشکستگی کم می‌شود و بنابراین ورشکستگی یک پیشامد کم محتمل خواهد بود. از این رو شبیه‌سازی معمولی[۲] که برآورد $\psi(u,T)$ و $\psi(u)$ با میانگین نمونه‌ای از تکرارهای مستقل $I\left(\sup_{۰\le t<T} S(t) > u\right)$ و $I\left(\sup_{t\ge ۰} S(t) > u\right)$ ($I(.)$ متغیر تصادفی نشانگر می‌باشد) است دارای کارایی لازم نبوده و روشهای شبیه‌سازی خاصی مورد نیاز است. در شبیه‌سازی معمولی باید مسیرهای نمونه‌ای بسیاری تولید کرد تا به مسیر نمونه‌ای که روی آن ورشکستگی رخ‌می‌دهد دسترسی پیدا کنیم. تولید این مسیرهای نمونه‌ای ضعف این نوع شبیه‌سازی را نمایان می‌سازد.

نمونه‌گیری ویژه[۳] یک روش خاص شبیه‌سازی می‌باشد که تحت آن سرعت وقوع پیشامدهای کم محتمل بالاتر می‌رود [هیدل برگر[۴] (۱۹۹۵) و اسموزن (۲۰۰۰)]. در این روش پس از تغییر اندازه احتمال از نسبت درستنمایی برای تعیین برآورد ناااریب استفاده می‌شود. تغییر اندازه احتمال باید به گونه‌ای باشد که واریانس برآورد کننده در مقایسه با شبیه‌سازی معمولی بزرگ نشود [لهتنن[۵] و نیرهینن[۶] (۱۹۹۲)]. در حالتی که توزیع اندازه خسارت با دم سبک باشد تغییر اندازه احتمال به صورت چرخش نمایی[۷] مفید می‌باشد [اسموزن (۱۹۸۵)]. بدین معنی که این رهیافت به طور

2) naive simulation 3) importance sampling 4) Heidel berger 5) Lehtohnen
6) Nyrhinen 7) exponential twisting

مجانبی بهینه می‌باشد. به طور مجانبی بهینه بودن ملاکی استاندارد برای میزان کارایی شبیه‌سازی پیشامدهای کم محتمل می‌باشد.

هنگامی که اندازه خسارت با دم سنگین باشد به کار بردن چرخش نمایی برای حل مسأله مفید واقع نمی‌شود. توزیع با دم سنگین توزیعی است که تابع مولد گشتاور متناهی نداشته و به همین علت از چرخش نمایی که مبتنی بر تابع مولد گشتاور است نمی‌توان استفاده کرد. اسموزن و بینسوانگر[8] (۱۹۹۷) اولین رهیافت برای شبیه‌سازی در حضور توزیع‌های زیرنمایی[9] (توزیع‌های زیرنمایی کلاسی از توزیع‌های با دم سنگین هستند) را ارائه دادند. آنها در صدد شبیه‌سازی احتمال تجاوز یک مجموع هندسی از متغیرهای تصادفی $i.i.d.$ با توزیع زیرنمایی از یک مقدار مشخص بودند. جونجا[10] و شهاب‌الدین[11] (۱۹۹۹) با معرفی چرخش نرخ خطر[12] یا چرخش زیرنمایی به حل مسأله فوق پرداختند.

آنها همچنین با بهبود بخشیدن این روش به وسیله چرخش نرخ خطر تأخیری[13] نشان دادند که شبیه‌سازی مبتنی بر آن برای چند توزیع زیرنمایی به طور مجانبی بهینه می‌باشد. این تکنیکها روشهایی برای برآورد احتمال آن است که یک مجموع هندسی از متغیرهای تصادفی $i.i.d.$ از یک مقدار داده شده تجاوز کند. حل چنین مسأله‌ای منجر به برآورد احتمال ورشکستگی نیز می‌شود. زیرا از تبدیل Pollaczek-Khinchin احتمال ورشکستگی در فرایندهای مخاطره بیمه با ادعاهای پواسون و اندازه‌های خسارت $i.i.d.$ به مسأله مجموع هندسی ذکر شده تبدیل می‌شود. به هر حال در اکثر موارد حتی به کمک چنین تبدیلی نیز برآورد احتمالات ورشکستگی بسیار مشکل است.

بوتز[14] و شهاب‌الدین (۲۰۰۰) برآورد $\psi(u)$ را با زمانهای بین ورود $i.i.d.$ و بدون استفاده از تبدیل Pollaczek-Khinchin در نظر گرفتند. آنها با شرط اینکه اندازه‌های خسارت $i.i.d.$ بوده و از زمانهای بین ورود مستقل باشند مستقیماً به شبیه‌سازی فرایند مازاد خسارت پرداختند. بدین صورت که قدم تصادفی $S_i = S_{i-1} + (X_i - c\xi_i)$ به دست آمده از فرایند مازاد خسارت را شبیه‌سازی کردند.

آنها تغییر اندازه چرخش زیرنمایی برای اندازهای خسارت X_i رابه کار بردند. سپس نشان دادند که برآوردکننده حاصل برای توزیع‌های اندازه خسارت وایبول به طور مجانبی در یک مجموعه بزرگ بهینه[15] می‌باشد. به طور مجانبی در یک مجموعه بزرگ بهینه بودن ملاک دیگری برای کارایی شبیه‌سازی پیشامدهای کم محتمل می‌باشد.

در این مقاله به تعمیم کارهای انجام شده توسط بوتز و شهاب‌الدین (۲۰۰۰) پرداخته می‌شود و خواهیم دید که باارائه شرطهای کافی روی فرایند خسارت و توزیع‌های اندازه خسارت تغییر

8) Binswanger 9) subexponential 10) Juneja 11) Shahabuddin 12) hazard rate twistinghazard 13) delayed hazard rate twisting 14) Boots 15) larg set asymptotically optimal

اندازه‌های استفاده شده به طور مجانبی در یک مجموعه بزرگ بهینه می‌باشند. همچنین حالتی را در نظر می‌گیریم که اندازه‌های خسارت وابسته بوده و به شبیه‌سازی احتمالات ورشکستگی خواهیم پرداخت.

۲ نمادگذاری و مدلها

اگر دم توزیع F را با $\bar{F} = ۱ - F$ چگالی آن را با f و تابع نرخ خطرش را با $\lambda(x) = \frac{f(x)}{\bar{F}(x)}$ نشان دهیم داریم: $\bar{F}(x) = e^{-\Lambda(x)}$ به طوریکه $\Lambda(x) = \int_۰^x \lambda(s)ds$ تابع خطر است. توزیع دم انتگرالگیری شده[۱۶] F را به صورت $F_I(x) = \frac{\int_۰^x \bar{f}(y)dy}{\int_۰^\infty \bar{F}(y)dy}$ به شرط $(\int_۰^\infty \bar{F}(y)dy < \infty)$ تعریف می‌کنیم و $\Lambda(x), \lambda_I(x)$ را به ترتیب تابع نرخ خطر و تابع خطر متناظر با F_I در نظر می‌گیریم. برای توابع دلخواه $z_۱(x)$ و $z_۲(x)$, $z_۱(x) \sim z_۲(x)$ بدین معنی است که وقتی $x \to \infty, \frac{z_۱(x)}{z_۲(x)}$ به ۱ همگراست.

زمان ورشکستگی را به صورت $\sigma(u) = inf(t : S(t) > u)$ تعریف می‌کنیم و قرار می‌دهیم $\tau(u) = N(\sigma(u))$. اگر ورشکستگی رخ دهد $\tau(u)$ تعداد ادعاها تا قبل از وقوع ورشکستگی می‌باشد. برای $T < \infty$ تعریف می‌کنیم $\tau(u,T) = N(Min(\delta(u),T))$. در این صورت داریم:

$$\psi(u,T) = P(\sigma(u) < T) \quad , \quad \psi(u) = P(\sigma(u) < \infty) = P(\tau(u) < \infty)$$

۱.۲ مدل مخاطره بیمه

مدل مخاطره بیمه را مانند بخش ۱ در نظر می‌گیریم. اندازه‌های خسارت X_i می‌توانند به طور تصادفی یکی از m نوع مختلف از توزیعهای $F_۱, F_۲, ..., F_m$ با میانگینهای متناهی را داشته باشند. فرض کنیم X_iها مستقل بوده و حداقل یکی از F_iها $(۱ \le i \le m)$ دارای توزیع زیرنمایی باشد. توزیعهای زیرنمایی رده‌ای خاص از توزیعهای با دم سنگین هستند. در مقابل توزیعهای با دم سنگین توزیعهای با دم سبک توزیعهایی هستند که دم آنها با نرخ نمایی (سریعتر از توزیعهای با دم سنگین) کاهش می‌یابد.

تعریف: اگر X_n دنباله‌ای از متغیرهای تصادفی $i.i.d.$ و غیر منفی با توزیع F و nامین پیچش F^{*n} باشد F را زیرنمایی گوییم هرگاه برای هر $n \ge ۲$ داشته باشیم:

$$\frac{\bar{F}^{*n}(u)}{n\bar{F}(u)} \sim \frac{P(X_۱ + \cdots + X_n > u)}{nP(X_۱ > u)} \to ۱ \quad , \quad (u \to \infty)$$

16) integrated tail

توزیع‌هایی مانند لگ نرمال، وایبول و پارتو متعلق به این خانواده هستند. تابع نرخ خطر بسیاری از توزیع‌های زیرنمایی نهایتاً نزولی می‌باشد. درامبرچ[17] و کلاپلبرگ[18] (۱۹۹۷) جزئیات در مورد توزیع‌های زیرنمایی آمده است. حال در مدلهای مخاطره بیمه دو مدل زیر را معرفی می کنیم.

۱.۱.۲ مدل پواسون مرکب

در این مدل زمانهای ورود ادعاها یک فرایند تجدید با زمانهای بین ورود نمایی هستند و بنابراین تعداد ادعاها تشکیل یک فرایند پواسون را می‌دهد. همچنین در این مدل X_iها $i.i.d.$ بوده که از ξ_iها نیز مستقل می‌باشند.

۲.۱.۲ مدل مارکوف ـ مدل پواسون مرکب

در این مدل زمانهای بین ورود ادعاها $i.i.d.$ مستقل از X_iها با فرایند پواسون بوده شدت ورود و توزیع اندازه خسارت بر اساس فرایند مارکوف $J(t), t \geq ۰$ با فضای حالت متناهی $\{۱, ۲, ..., m\}$ همراه توزیع مانای یکتای $\pi_۱, \pi_۲, ..., \pi_m$, $\pi_i > ۰$ برای هر $۱ \leq i \leq m$ مشخص می‌شود. وقتی $J(t)$ در حالت i است شدت ورود را w_i نرخ حق بیمه c را برابر ۱ و توزیع اندازه خسارت را F_i قرار می‌دهیم. در این صورت داریم $F_i = P\left(X_i \leq x | J(\eta_j) = i\right)$. در این مدل داریم:

$$\beta = \frac{\sum_{i=۱}^{m} w_i \pi_i \beta_i}{\sum_{i=۱}^{m} w_i \pi_i} \quad , \quad w = \sum_{i=۱}^{m} \pi_i w_i$$

که در آن β_i میانگین F_i می‌باشد. بنابراین داریم: $\rho = \sum_{i=۱}^{m} \pi_i w_i \beta_i$. اگر تعریف کنیم

$$P_i(.) = P\left(. | J(.) = i\right)$$

احتمال ورشکستگی قبل از زمان T با حالت اولیه i را به صورت $\psi_i(u, T)$ نشان می‌دهیم. در ادامه برای سادگی اندیس i را حذف می‌کنیم.

۳ شبیه‌سازی پیشامدهای کم محتمل

همانطورکه در بخش ۱ گفته شد برای u بزرگ $\psi(u, T)$ کوچک بوده و شبیه‌سازی معمولی برای برآورد آن کاربرد چندانی ندارد. در اینجا نوع دیگری از شبیه‌سازی با نام نمونه‌گیری ویژه را معرفی می‌کنیم. فرض کنیم فرایند تصادفی که درصدد شبیه‌سازی آن هستیم در یک فضای احتمال با

17) Embrech 18) Kluppelbrg

اندازه احتمال P تعریف شده باشد. $A(u)$ را پیشامدی با احتمال $a(u) = P(A(u))$ طوری در نظر می‌گیریم که وقتی $u \to \infty$, $a(u) \to 0$. فرض کنیم Q اندازه احتمال دیگری روی همان فضای احتمال بوده به طوریکه P نسبت به آن مطلقاً پیوسته باشد. در این صورت داریم:

$$a(u) = E_Q\left(I(A(u)dP/dQ\right) \qquad (۲)$$

در (۲) dP/dQ نسبت درستنمایی است و اندیس Q نشان‌دهنده آن است که امید تحت اندازه Q گرفته می‌شود. نسبت درستنمایی dP/dQ به صورت زیر تعریف می‌شود:

$$\frac{dP}{dQ} = L\left(X_1, X_2, ...X_n\right) = \Pi_{i=1}^{n} \frac{f(X_i)}{g(X_i)}$$

که در آن f و g توابع چگالی X_i به ترتیب تحت اندازه احتمال Q, P می‌باشند. در نمونه‌گیری ویژه مسیرهای نمونه‌ای تحت اندازه احتمال Q را تولید کرده در هر حالت نسبت درستنمایی را محاسبه می‌کنیم. سپس از میانگین نمونه‌ای $I\left(A(u)(\frac{dP}{dQ})\right)$ها برآورد نااریب برای $a(u)$ تشکیل می‌دهیم. مسأله مهم در نمونه‌گیری ویژه تعیین اندازه احتمال جدید Q می‌باشد به طوریکه واریانس این برآورد نااریب در مقایسه با شبیه‌سازی معمولی مقدار بزرگی نباشد.

یک ملاک استاندارد برای مشخص کردن کارایی شبیه‌سازی پیشامدهای کم محتمل به طور مجانبی بهینه بودن می‌باشد.

تعریف: $\hat{a}(u)$ برآوردکننده به طور مجانبی بهینه برای $a(u)$ می‌باشد هرگاه:

$$\lim_{u\to\infty} SUP \frac{\log\left(Var[\hat{a}(u)]\right)}{\log\left(a^2(u)\right)} \geq 1$$

در بسیاری از موارد رسیدن به چنین ملاکی مشکل می‌باشد از این رو با صرف نظر از اریبی ملاک ضعیف‌تر زیر را به کار می‌بریم:

تعریف: بهینگی مجانبی در مجموعه بزرگ نرمال شده[19] (w.a.o) فرض کنیم $\delta \in (0, 1)$ ثابت و مشخص باشد. تجزیه پارامتری $a(u)$ با پارامتر β را به صورت زیر در نظر می‌گیریم:

$$a(u) = \gamma_\beta(u) + \epsilon_\beta(u)$$

اگر داشته باشیم:

۱) برای هر β داده شده $\lim_{u\to\infty} \sup \frac{\epsilon_\beta(u)}{a(u)} \leq a$

۲) برای هر u مشخص $\lim_{\beta\to\infty} \sup \frac{\epsilon_\beta(u)}{a(u)} = 0$

19) work - normalaized large set asymptiotically optimal

۳) برای هر β داده شده یک برآورد نااریب برای $\gamma_\beta(u)$ وجود داشته به طوریکه:

$$\lim_{u\to\infty} inf \frac{\log\left(work(u)\times Var[\hat{\gamma}_\beta(u)]\right)}{\log\left(\gamma\beta^{۲}(u)\right)} \geq ۱$$

در این صورت $\hat{\gamma}_\beta(u)$ را برآورد به طور مجانبی بهینه در یک مجموعه بزرگ نرمال شده برای $a(u)$ می‌نامیم. $work(u)$ تابعی از u بوده که محاسبات در هر بار تکرار شبیه‌سازی را نشان می‌دهد[بوتز وشهاب‌الدین (۲۰۰۰)]. در تعریف فوق پارامتر β برای کنترل اریبی نسبی برآوردکننده به کار می‌رود به طوریکه برای هر مقدار مشخص u اریبی نسبی کمتر از δ باشد. برای سادگی از $\gamma(u)$ به جای $\gamma_\beta(u)$ و از $\epsilon(u)$ به جای $\epsilon_\beta(u)$ استفاده می‌کنیم.

اندازه احتمال جدید Q که از آن استفاده می‌کنیم چرخش نرخ خطر(HRT) روی هر یک از توزیعهای اندازه خسارت می‌باشد. چرخش نرخ خطر به صورت جایگزین کردن توزیع F_i با توزیع جدید $F_{j,\theta_u}(x) = ۱ - e^{-\Lambda_j(x)(۱-\theta_u)}$, $(۰ \leq \theta_u < ۱)$ می‌باشد و چگالی مطابق با F_{j,θ_u} به صورت $f_{j,\theta_u}(x) = (۱-\theta_u)\lambda_j(x)e^{-(۱-\theta_u)\Lambda_j(x)}$ است. θ_u تابعی مناسب از u است که با ساختن کران بالا برای نسبت درستنمایی و می‌نیمم کردن آن نسبت به θ_u به دست می‌آید. چرخش نرخ خطر تأخیری وزنی (WDHRT)[۲۰] تعمیمی برای HRT می‌باشد به طوریکه برآوردکننده ناشی از آن بهینگی بیشتری نسبت به HRT دارد[بوتز و شهاب‌الدین (۱۹۹۹)]. چگالی چرخش نرخ خطر تأخیری وزنی را با معرفی پارامتر وزنی w_u و پارامتر تأخیری x_u^* (هر دو توابعی از u هستند) به صورت زیر تعریف می‌کنیم:

$$f_{j,\theta_u,x_u^*}(x) = \begin{cases} \frac{f_j(x)}{۱+w_u} & x \leq x_u^* \\ \left(۱ - \frac{F_j(x_u^*)}{۱+w_u}\right)\frac{f_{j,\theta_u}(x)}{F_{j,\theta_u}(x_u^*)} & x > x_u^* \end{cases}$$

تابع توزیع آنرا نیز با F_{j,θ_u,x_u^*} نشان می‌دهیم.

انتخاب هر یک از این پارامترها باید به گونه‌ای باشد که در نمونه‌گیری ویژه خواص اندازه احتمال جدید بسیار نزدیک به خواص اندازه احتمال اصلی باشد.

۴ احتمالات ورشکستگی با افق نامتناهی

همانطورکه گفته شد با استفاده از چرخش نرخ خطر برای برآورد احتمال ورشکستگی از:

$$E_Q\left(I(\tau(u) < \infty)L_Q(u)\right)$$

20) weighted delayed hazard rate twisting

استفاده می‌کنیم. که در آن L تابع نسبت درستنمایی می‌باشد. در اینجا با یک مسأله روبرو هستیم. تغییرات $L_Q(u)$ روی مجموعه مسیرهای نمونه‌ای زیاد است. اگر بتوانیم تابعی از u مانند $k_\circ(u)$ طوری پیدا کنیم که:

$$E\left(I(\tau(u)<\infty)\right) \sim E\left(I(\tau(u)\le K_\circ(u))\right) = E_Q\left(I(\tau(u)\le K_\circ(u)L_Q(u))\right)$$

در این حالت اگر L دارای واریانس کمی روی $(\tau(u)\le k_\circ(u))$ باشد می‌توانیم بر این مسأله فائق آییم.

حال این سؤال پیش می‌آید که چه میزان اریبی را می‌توان تحمل کرد؟ اگر قرار دهیم:

$$\gamma(u) = E\left(I(\tau(u)\le k_\circ(u))\right)$$

در این صورت اگر $\frac{E(I(\tau(u)<k_\circ(u)))}{E(I(\tau(u)\le\infty))} < 1-\delta$ می‌توان از اریبی به وجود آمده صرف نظر کرد[بوتز و شهاب‌الدین(۲۰۰۱)]. مسأله دیگر تعیین $k_\circ(u)$ می‌باشد. برای تعیین این تابع شرطهای زیر را در نظر می‌گیریم. خواهیم دید که این شرطها برای برقراری ملاک w.a.o مورد نیاز می‌باشد.

۱.۴ شرطهای انحرافات بزرگ

۱) $\psi(u)$ به گونه‌ای باشد که: $\lim_{u\to\infty}\frac{-\log(\psi(u))}{\Lambda_I(u)} = 1$

۲) برای هر $\delta > 0$ داده شده تابع $k_\circ(u)$ به گونه‌ای باشد که:

$$\lim_{u\to\infty}\frac{P\left(\tau(u)\le k_\circ(u)\right)}{P\left(\tau(u)<\infty\right)} \ge 1-\delta$$

۳) $\log\left(k_\circ(u)\right) = O\left(\Lambda_I(u)\right)$

۴) ثابت $b > 1$ به گونه‌ای باشد که: $\lim_{u\to\infty}\frac{k_\circ(u)}{(\Lambda(u))^{b-1}} = 0$ و

$$\lim_{u\to\infty}\frac{k_\circ(u)F^{\leftarrow}(1-\Lambda(u)^{-b})}{u} = 0$$

که در آن $F^{\leftarrow} = inf\left(x : F(x) = y\right)$.

بدین با شرطهای فوق محدودیتهایی را برای $k_\circ(u)$ قائل می‌شویم.

رده زیادی از فرایندهای مخاطره بیمه در شرطهای انحرافات بزرگ ذکر شده صدق می‌کنند. به عنوان مثال اگر F زیرنمایی باشد در مدل پواسون مرکب و مارکوف ـ مدل پواسون مرکب شرط (۱) برقرار است. همچنین توزیع وایبول با shape پارامترکمتر از (۱) در شرطهای (۲) الی (۴) صدق می‌کند.

۲.۴ الگوریتم شبیه‌سازی و کارایی آن

برای شبیه‌سازی از چرخش نرخ خطر تأخیری وزنی روی هر یک از توزیعهای اندازه خسارت استفاده می‌کنیم. پارامترها را به صورت زیر در نظر می‌گیریم:

$$\theta_u = ۱ - \frac{۱}{\Lambda(u)} \quad , \quad w_u = \frac{c_۱ \log \delta}{k_۰(u)} \quad , \quad x_u^* = F^{\leftarrow}(۱ - \Lambda(u)^{-b}) \quad (*)$$

به طوریکه $c_۱ > ۱$ و b ثابت ذکر شده در شرط (۴) می‌باشد. باید توجه داشت که از $\Lambda(x_u^*) = b \log(\Lambda(u))$ نتیجه می‌شود که وقتی $u \to \infty$, $x_u^* \to \infty$ زیرا $\Lambda(u) \to \infty$.

الگوریتم ۱ (WDHRT برای برآورد $\psi(u)$): به جای $F_i = ۱ - e^{-\Lambda_i(x)}$ از WDHRT روی F_i با پارامترهای θ_u , w_u , x_u^* انتخاب شده به صورت $(*)$ استفاده می‌کنیم. L را نسبت درستنمایی قرار می‌دهیم. شبیه‌سازی را با تکرار از

$$S = \sum_{i=۱}^{min(\tau(u),k_۰(u))} (X_i - c\xi_i)$$

انجام می‌دهیم. اگر $S > u$ باشد قرار می‌دهیم $V = L$ و در غیر این صورت قرار می‌دهیم $V = ۰$. میانگین K تکرار $i.i.d$ از V را به عنوان یک برآورد ناااریب از $P(\tau(u) \geq k_۰(u))$ در نظر می‌گیریم. از این برآورد کننده برای برآورد $\psi(u)$ استفاده می‌کنیم.

قضیه: تحت شرطهای انحرافات بزرگ اگر توزیع F دارای تابع نرخ خطر نهایتاً نزولی با خاصیت $\lim_{u\to\infty} \frac{\Lambda(u)}{\Lambda_I(u)} = ۱$ باشد الگوریتم ۱ برآورد کننده به طور مجانبی بهینه در یک مجموعه بزرگ نرمال شده برای $\psi(u)$ فراهم می‌کند (برای اثبات می‌توان به بوتز و شهاب‌الدین (۲۰۰۱) مراجعه کرد).

۵ احتمالات ورشکستگی با افق متناهی

$a_۰$ را کمینه روی هر a می‌گیریم به طوریکه $P(N(T) \geq k) \leq a^k$ برای k به قدر کافی بزرگ برقرار باشد. شرط زیر را در نظر می‌گیریم:

$$\lim_{u\to\infty} -\frac{\log(\psi(u,T))}{\Lambda(u)} = ۱$$

قضیه: با فرض نهایتاً نزولی بودن $\lambda(x)$ در مدلهای تجدید و مارکوف ـ مدل پواسون مرکب $\psi(u,T)$ در شرط فوق صدق می‌کند.

۱.۵ الگوریتم شبیه‌سازی و کارایی آن

x_u^*، θ_u را مانند قبل انتخاب می‌کنیم. اما w_u را طوری انتخاب می‌کنیم که $w_u < \frac{۱}{a_۰} - ۱$.
اگر $a_۰$ نامعلوم باشد قرار می‌دهیم: $w_u = ۰$.
الگوریتم ۲ (WDHRT برای برآورد $\psi(u,T)$): به جای $F_i = ۱ - e^{-\Lambda_i(x)}$ از چرخش نرخ خطر تأخیری وزنی استفاده می‌کنیم. تکرارها را از

$$S = \sum_{i=۱}^{\tau(u.T)} (X_i - cmin(\delta(u), T))$$

تولید می‌کنیم. اگر $S > u$ باشد قرار می‌دهیم $V = L$ و اگر $S \le u$ باشد قرارمی‌دهیم $V = ۰$. میانگین K تکرار $i.i.d$ از V را به عنوان برآورد ناااریب $\psi(u,T)$ در نظر می‌گیریم.
قضیه: با فرض نهایتاً نزولی بودن تابع نرخ خطر خسارت الگوریتم ۲ با پارامترهای θ_u ، x_u^* ، w_u به طور مجانبی بهینه می‌باشد.

۶ ورشکستگی با خسارتهای وابسته

در بسیاری از مواقع فرض استقلال اندازه‌های خسارتها غیر واقعی جلوه‌کرده به طوریکه صرفنظر از وابستگی خسارتها می‌تواند برای برآورد احتمال ورشکستگی گمراه کننده باشد. دیده شده است که وقتی خسارتها وابسته هستند نسبت به حالتی که مستقل‌اند زمان ورشکستگی زودتر رخ می‌دهد[امیلیانو[۲۱] و کلوین[۲۲](۲۰۰۲)].
تعریف: تابع زیر را از $[۰,۱]^n$ به $[۰,۱]$ برای $n \ge ۱$ به صورت زیر در نظر می‌گیریم:

$$F_{X_۱,X_۲,...,X_n}(x_۱, x_۲, ..., x_n) = C\left(F_{X_۱}, F_{X_۲}, ..., F_{X_n}\right) \qquad (۳)$$

به طوریکه $F_{X_۱,X_۲,...,X_n}$ تابع توزیع تجمعی بردار تصادفی از متغیرهای $(X_۱, ..., X_n)$ و F_{X_i} ها $(۱ \le i \le n)$ نشان دهنده توزیعهای حاشیه‌ای هستند. تابع C یک کاپولا[۲۳] نامیده می‌شود.
C یک تابع توزیع تجمعی توأم با توزیعهای حاشیه‌ای یک متغیره $U(۰,۱)$ می‌باشد. کاپولاها در واقع توابعی هستند که ارتباط بین توزیع توأم و توزیعهای حاشیه‌ای را با پارامترهای توصیف کننده وابستگی متغیرها نشان می‌دهد. در رابطه (۳) تابع کاپولا منحصربفرد نبوده ولی با فرض مطلقاً پیوسته بودن توزیعهای حاشیه‌ای یک متغیره تابع C تنها تابع با تعریف (۳) می‌باشد.
مثال: بردار $(u_۱, u_۲, ..., u_n)$ را با توزیعهای حاشیه‌ای یک متغیره $U(۰,۱)$ در نظر می‌گیریم.

21) Emiliano 22) Kelvin 23) capula

تابع

$$C\left(u_1, u_2, ..., u_n\right) = -\frac{1}{\log \eta} \log\left(1 + \frac{\Pi_{k=1}^{n}(\eta^{u_k} - 1)}{(\eta - 1)^{n-1}}\right) \qquad (۴)$$

یک کاپولا بوده که موسوم به فرانک کاپولا[۲۴] می‌باشد. تابع کاپولای ارائه شده در (۴) را می‌توان به صورت:

$$C\left(u_1, ...u_n\right) = \phi^{-1}\left(\phi(u_1) + \phi(u_2), ..., \phi(u_n)\ \right) \qquad (۵)$$

نوشت. در (۵) ,ϕ به صورت:

$$\phi(t) = \log\left(\frac{\eta^t - 1}{\eta - 1}\right) \qquad (۶)$$

می‌باشد. در (۴) و (۶)، η به گونه‌ای است که $\eta - 1$ پارامتر یک توزیع لگاریتمی گسسته می‌باشد[امیلیانو و کلوین (۲۰۰۲)].

۱.۶ شبیه‌سازی احتمالات ورشکستگی با خسارتهای وابسته

در بخشهای (۴) و (۵) الگوریتم‌های شبیه‌سازی مطرح شد. در آنجا دیدیم که مسیرهای نمونه‌ای از فرایند مورد نظر را تولیدکرده و از بین آنها خود را روی مسیرهایی که ورشکستگی روی آنها رخ می‌دهد متمرکز کردیم. حال به تولید چنین مسیرهایی با فرض وابسته بودن خسارتها می‌پردازیم. با استفاده از تابع کاپولای (۴) و مولد (۶) مسیرهای نمونه‌ای را به صورت زیر تولید می‌کنیم:

۱) متغیر تصادفی Z را با توزیع لگاریتمی گسسته و با پارامتر $\eta - 1$ تولید می‌کنیم.

۲) n متغیر تصادفی $i.i.d$ با توزیع $U(0, 1)$ تولید می‌کنیم. این متغیرهای تولید شده را به صورت $U^* = (u_1^*, u_2^*, ..., u_n^*)$ نشان می‌دهیم.

۳) قرار می‌دهیم $U = M_Z\left(z^{-1}\log U^*\right)$ به طوریکه $M_Z(.)$ تابع مولد گشتاور Z به صورت $M_Z(t) = \frac{\log\left(1-(1-\eta)e^t\right)}{\log(\eta)}$ می‌باشد و $\log U^* = (\log u_1^*, \log u_2^*, ..., \log u_n^*)$.

۷ نتایج عددی

در این بخش با استفاده از الگوریتمهای ارائه شده به برآورد احتمالات ورشکستگی با افق زمانی متناهی و نامتناهی در دو مدل پواسون مرکب و مارکوف ـ مدل پواسون مرکب می‌پردازیم. نتایج عددی ارائه شده از برنامه نوشته شده به زبان پاسکال برای شبیه‌سازی به دست آمده است. وقتی F دارای توزیع وایبول است برای انتخاب w_u , c_1 را به طور تجربی برای احتمال ورشکستگی

24) Frank's copula

با افق نامتناهی انتخاب می‌کنیم. این انتخاب را با $\tilde{w}_u$ نشان می‌دهیم. وقتی F لگ نرمال است قرار می‌دهیم $w = ۰$.

از قضیه حد مرکزی نصف طول فاصله اطمینان$(\eta - ۱)۱۰۰$ درصد عبارت است از:

$$Z_{۱-\eta/۲}\frac{(Var(\hat{a}(u)))^{\frac{۱}{۲}}}{a(u)}$$

که در آن Z_n، nامین چندک توزیع نرمال استاندارد را نشان می‌دهد. خطای نسبی را به صورت زیر تعریف می‌کنیم:

$$RE[\hat{a}(u)] = Z_{۱-\eta/۲}\frac{(Var(\hat{a}(u)))^{\frac{۱}{۲}}}{a(u)}$$

۱.۷ احتمالات ورشکستگی با افق متناهی در مدل پواسون مرکب

در جدول ۱ برآوردهای $\psi(u,T)$ در مدل پواسون مرکب با اندازه‌های خسارت وایبول و پارامترهای $\beta = ۲$, $\rho = ۰/۵$, $b = ۲/۱$ آمده است. تعداد تکرارها در شبیه‌سازی $n = ۳۰۰۰۰۰$ می‌باشد. با مقایسه خطاهای نسبی می‌بینیم که انتخاب $\tilde{w}_u$ بر $w = ۰$ ترجیح دارد.

جدول ۱: برآورد $\psi(u,T)$ در مدل پواسون مرکب با اندازه‌های خسارت وایبول $(۱, ۰/۵)$

u	پارامتر	برآورد $\psi(u, ۱۰۰)$	خطای نسبی به درصد
200	$w = ۰$	1.3E-5	10.9
200	$w = \tilde{w}_u$	1.49E-5	3.9
800	$w = ۰$	1.04E-11	17.5
800	$w = \tilde{w}_u$	1.01E-11	4.7

جدول ۲: برآورد $\psi(u,T)$ در مدل مارکوف ـ مدل پواسون مرکب با وایبول $(۱, ۰/۵)$ و وایبول $(۱, ۰/۷۵)$

u	برآورد $\psi(u)$	خطای نسبی به درصد
100	2.55E-4	3.4
400	1.69E-8	3.8

۲.۷ مدل مارکوف مدل ـ پواسون مرکب

جدولهای ۲ و ۳ برآوردهای $\psi(u,T)$ را در دو مدل مارکوف ـ مدل پواسون مرکب نشان می‌دهد. در جدول ۲ $n = ۱۰۰۰۰۰۰$, $b = ۲/۱$, $w = \tilde{w}_n$ و در جدول ۳ $\tilde{w} = ۰$ می‌باشد. در هر دو جدول $F_۲$وایبول $(۱, ۰/۷۵)$ با $\beta_۲ = ۱/۱۹۰۶$ بوده و $\pi_۱ = ۰/۵$, $\pi_۲ = ۰/۵$, $w_۱ = ۰/۲$, $w_۲ = ۰/۳$ و $\rho = ۰/۳۸$ می‌باشند. در جدول ۲ $F_۱$ وایبول $(۱, ۰/۵)$ بوده و $\beta_۱ = ۲$ می‌باشد. در جدول ۳ $F_۱$ پارتو $(۱, ۰/۵)$ با $\beta_۱ = ۲$ است.

جدول ۳: برآورد $\psi(u, ۵۰)$ در مدل مارکوف ـ مدل پواسون مرکب با پارتو $(۲, ۱)$ و وایبول $(۱, ۰/۷۵)$

u	برآورد $\psi(u)$	خطای نسبی به درصد
100	3.32E-5	6.8
400	8.53E-6	6.5

مراجع

[1] Asmussen, S. (1985), Conjugate processes and the simulation of ruin problem, Stochastic processes and their applications, 20:213-229.

[2] Asmussen, S. (2000), Ruin probabilities, World scientific, Singapore, New Jersey, London, Hong Kong.

[3] Asmussen, S. and Binswanger, K. (1997), Simulation of ruin probabilities for subexponential claim, ASTIN Bulletin, 27(2):297-318.

[4] Boots, N.K. and Shahabuddin, P. (2000), Simulation GI/GI/1 queues and insurance risk processes with subexponential distributions, Proceeding of the 2000 winter simulation conference 656-665, IEEE press, piscat away, New Jersay.

[5] Boots, N.K. and Shahabuddin, P. (2001), A framework for simulating small ruin probabilities in insurance risk processes with subexponential distributions, Research Report, Dept. of Industrial Engineering and Operation Research, Columbia University, NY 10027.

[6] Embrechts, P. Kluppelberg, C. and Mikosch, T. (1997), Modeling Extremal events, Springer- Verlag, Berlin, Heidelberg.

[7] Emiliano, A. Valdez and Kelvin, Mo. (2002), Ruin probabilities with dependent claims, Faculty of Commerce and Economics, The University of New South Wales, Sydney, Australian 2052.

[8] Heidelberger, P. (1995) Fast simulation of rare events in queueing and reliability models, ACM Transaction on Modeling and Computer Simulation, 6:43-85.

[9] Juneja, S. and Shahabuddin, P. (1999), Simulation heavy-tailed processes using delayed hazard rate twisting, Research Report, Dept. of Industrial Engineering and Operation Research, Columbia University, NY 10027.

[10] Lehtonen, T. and Nyrhinen, H. (1992), Simulating level- crossing probabilities by importance sampling, Advances in applied probability , 24:858-874.

تعیین دبی‌ها اوج[1] با استفاده از داده‌های بالاتر از یک آستانه معین[2]

سید سعید موسوی ندوشنی

دانشکده صنعت آب و برق (شهید عباسپور)

چکیده: یکی از روش‌های برآورد سیلابها، تحلیل فراوانی[3] است. در این روش با استفاده از توزیع آماری مناسب می‌توان چندک‌های[4] سیل را محاسبه نمود. اما برای این کار ما به یک سری از داده نیاز داریم، که ازکل آمار موجود اخذ می‌گردد. برای این گزینش دو روش پیشنهاد می‌گردد:

۱) سری حداکثرهای سالانه (یک حداکثر در هر سال)

۲) سری بالاتر از یک آستانه معین

در این مقاله بر روی نحوه نمونه‌گیری با استفاده از روش (۲) بحث خواهد شد، که معرف بهتری از جامعه آماری سیلابها خواهد بود. نحوه انتخاب آستانه و آزمون‌هایی نظیر همگنی، ایستایی و استقلال (بر روی داده‌های بالاتر از آستانه معین) مورد علاقه قرار می‌گیرد. انجام آزمون‌های فوق به ما مقدار تقریبی آستانه را دیکته می‌کند و در انتها ملاحظه خواهد شد که یافتن مقدار مذکور یک رویکرد چندگانه است.

واژه‌های کلیدی: سری، آستانه، تحلیل فراوانی، ایستایی، همگنی، نمونه‌گیری

۱ مقدمه

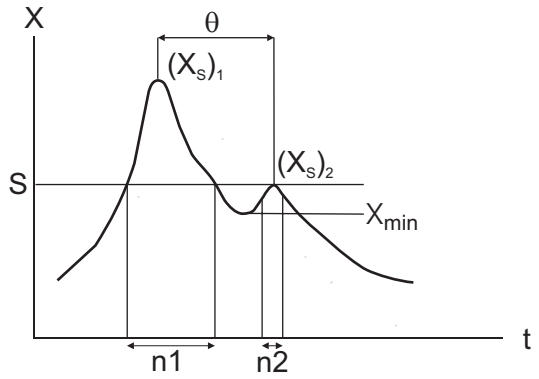
برنامه‌ریزی موثر و بهره‌برداری از طرحهای آبی نیاز به آگاهی از رفتار احتمالی مقادیر حدی دارد. تحلیل فراوانی می تواند عهده‌دار تبیین چنین رفتاری باشد. تحلیل فراوانی با سریهای زیر قابل انجام است:

۱) سری حداکثرهای سالانه (یک حداکثر در هر سال)

۲) سری بالاتر از یک آستانه معین

با توجه به اینکه در روش آستانه معین، برای یکسال به انتخاب یک داده محدود نمی‌شویم، می‌توان مدعی شد که انتخاب عقلایی‌تری از جامعه سیلابها انجام شده است. با توجه به مقدار آستانه ممکن است که پاره‌ای از داده‌های روش حداکثر سالانه انتخاب نشود.

1) peak discharge 2) Peak Over Threshold 3) frequency analysis 4) quantiles



شکل ۱: نمودار تعیین ضابطه بازه زمانی بین سیلابها

روش دوم در مقایسه با روش اول، از اطلاعات بیشتری برخوردار است. استفاده از روش مذکور بالاخص در مواقعی که با داده‌ها کم روبرو هستیم، توصیه می‌شود.

۱.۱ نکات مورد بحث مقاله

در این مقاله ابتدا به نحوه نمونه‌گیری بالاتر از یک آستانه معین اشاره خواهد شد، سپس به مطالعه فرآیند پرداخته می‌شود و در خاتمه یک مطالعه موردی انجام خواهد شد.

۲ روش نمونه‌گیری با استفاده از آستانه معین

۱.۲ ضابطه استقلال دبی‌های اوج

در روش تحلیل فراوانی، فرض استقلال داده‌ها بررسی می‌شود. در ادبیات موضوع مورد مطالعه چندین ضابطه برای بررسی این فرض وجود دارد. مطابق (USWRC, 1976) بازه زمانی دو سیل متوالی پنج روز باضافه لگاریتم طبیعی سطح حوضه آبریز است (بر حسب مایل مربع)، علاوه بر این مطابق شکل ۱ جریانهای میانی بین دو نقطه اوج متوالی باید ۷۵ درصد کوچکترین دو دبی متوالی باشد. بنابراین نقطه اوج ۲(X_s) رد خواهد شد، چنانچه شرایط زیر برقرار شود.

$$\theta < ۵\, days + \ln(A) \quad \text{یا} \quad X_{min} > \frac{۳}{۴} \min[Q_۱, Q_۲]$$

کانن[5] (۱۹۷۹) این ضابطه را ارائه می‌دهد که نقطه اوج ۲(X_s) رد می‌شود، اگر داشته باشیم:

$$\theta < ۳T_p \quad \text{یا} \quad X_{min} > \frac{۲}{۳}(X_s)_۱$$

که در آن T_p متوسط اولین پنج هیدروگراف تیپ است.

انتخاب ضابطه استقلال مساله پیچیده‌ای است، میکل[6] (۱۹۸۴) حداقلی را برای θ بدست نمی‌دهد، اما برای آزمون استقلال ضرایب خودهمبستگی مرتبه اول و دوم برای سری زمانی مقادیر دبی‌های اوج را پیشنهاد می‌کند. اگر فرض صفر با سطح معنی‌دار بودن معینی رد شد، آنگاه مقدار بیشتری را برای بازه زمانی θ در نظر می‌گیریم.

۲.۲ انتخاب مقدار آستانه

برای این کار دو رویکرد متفاوت می‌توان اتخاذ نمود:

۱) استفاده از یک ضابطه فیزیکی مانند تعیین میزان سیلاب برای یک رودخانه مشخص

۲) استفاده از روش آماری، مقدار آستانه باید به گونه‌ای تعیین شود که دبی‌های اوج بالاتر از آن تشکیل یک سری داده‌ی مستقل از هم را بدهند و ضمناً فرآیند پواسون نیز آنرا تبیین کند.

در اینجا آزمون‌های که برای انتخاب میزان آستانه، پیشنهاد شده است، مورد بررسی قرار می‌گیرد.

۱.۲.۲ آزمون ۱: متوسط تعداد داده‌های بالاتر از آستانه

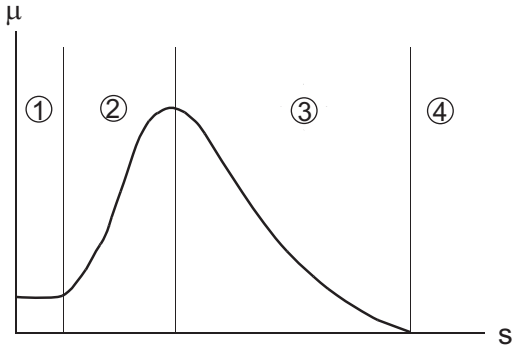
شرط استقلال مشاهدات، یکی از فرضیات فرآیند پواسون است. این نکته قابل توجه است که انتخاب مقدار کوچک آستانه همیشه نمی‌تواند شرط استقلال را محقق سازد. انتخاب آستانه کوچک، نقاط اوج را بدست می‌دهد، که همیشه از هم مستقل نیستند. افزایش مقدار آستانه به تغییر شرایط کمک می‌کند. مطابق شکل ۱ اگر $X_{min} < S < (X_s)$۲ باشد، آنگاه مقادیر وابسته را افزایش می‌دهد. بطور کلی با افزایش مقدار آستانه می‌توان چهار ناحیه را از هم تمیز داد. این نواحی در شکل ۲ نشان داده شده است که در آن μ متوسط تعداد نقاط اوج بالاتر از آستانه است.

اکنون به شرح نواحی پرداخته می‌شود.

۱) ناحیه ۱: برای مقادیر اندک آستانه، کل سری زمانی (با طول NY، یعنی N سال متوالی) بالاتر از آستانه قرار می‌گیرد.

۲) ناحیه ۲: در این ناحیه با افزایش مقدار آستانه، متوسط تعداد نقاط اوج بالاتر از آستانه افزایش می‌یابد.

5) Cunnane 6) Miquel



شکل ۲: نمودار تغییرات متوسط مقادیر بالاتر از آستانه بر حسب مقدار آستانه

۳) ناحیه ۳: افزایش مقدار آستانه از یک اندازه به بعد، باعث کاهش متوسط تعداد نقاط اوج بالاتر از آستانه می‌گردد.

۴) ناحیه ۴: در این ناحیه هیچ نقطه‌ی اوجی بالاتر از آستانه قرار نمی‌گیرد، چون مقدار آستانه از کل سری بزرگتر است.

۲.۲.۲ آزمون ۲: متوسط مقادیر داده‌های بالاتر از آستانه

دیویسن و اسمیت[۷] (۱۹۹۰) و ندن و بیلیس[۸] (۱۹۹۳) پیشنهاد کردند که انتخاب مقدار آستانه برای اینکه در دامنه مقادیر متوسط بالاتر از آستانه $(\overline{X}_s - S)$ قرار گیرد، باید تابع خطی از مقدار آستانه یعنی S باشد. این آزمون می‌تواند در انتخاب مقدار آستانه به گونه‌ای عمل کند که پارامترهای تابع چگالی احتمال داده‌های بالاتر از آستانه را به حداکثر پایداری خود برسد. میکل[۹] (۱۹۸۴) توصیه می‌کند که تحلیل حساسیت چندک[۱۰] برآورده شده را که تابعی از مقدار آستانه است، انجام شود.

۳.۲.۲ آزمون ۳: اندیس پراکنش

مقادیر آستانه بر اساس فرض پواسون بودن تعداد نقاط اوج، می‌تواند انتخاب شود. آشکار و رسل[۱۱] (۱۹۸۷) این آزمون را بر اساس پیشنهاد کانن[۱۲] (۱۹۷۹) انجام دادند. آستانه S طوری انتخاب می‌شود که اندیس پراکنش $\hat{i} = \frac{\hat{\sigma}^2_{m_t}}{\hat{\mu}_{m_t}}$ (که در آن m_t تعداد نقاط اوج در فاصله $[0, t]$)

7) Davison and Smith 8) Naden and Bayliss 9) Miquel 10) quantile
11) Ashkar and Rousselle 12) Cunnane

در فاصله اطمینان $[I_t(۰/۰۵), I_t(۰/۹۵)]$ قرار گیرد. شرح بیشتر این آزمون در ادامه مقاله خواهد آمد.

۴.۲.۲ بحث

رزبرگ و مدسن[۱۳] (۱۹۹۲) توصیه نمودند وقتی که از آزمون ۱ و ۲ استفاده می‌شود که مقدار $\mu > ۲$ باشد. در اینصورت توزیع GP یا توزیع نمایی منفی، دارای برازش مناسبی است. متوسط مقادیر بالاتر از یک آستانه دارای تغییرات خطی نسبت به مقدار آستانه، در توزیع GP است و برای توزیع نمایی منفی تغییرات ثابت است.

۱) توزیع GP

$$G_s(x) = \Pr[X_s < x] = ۱ - [۱ - (k/a)(x - S)]^{۱/k}$$
$$with\ \hat{X}_S - S = \frac{a}{k+۱} \qquad (۱)$$

$$G_{s^*}(x) = \Pr[X_{s^*} < x] = ۱ - [۱ - (k^*/a^*) \times (x - S^*)]^{۱/k^*}$$
$$with\ k^* = k;\ a^* = a - k(S - S^*);$$
$$\overline{X}_{s*} - S^* = \overline{X}_s - S - (S^* - S)k/(k+۱)$$

۲) توزیع نمایی منفی (توزیع GP با $k = ۰$)

$$G_s(x) = ۱ - \exp[(x - S)/a] \text{ با } \overline{X}_s - S = a \qquad (۲)$$

اگر آزمون شماره ۳ رد شود، می‌بایست از توزیعهایی نظیر دوجمله‌ای یا دوجمله‌ای منفی استفاده نمود. در نتیجه می‌توان ادعا نمود که روش کلی و واحد برای انتخاب مقدار آستانه وجود ندارد. بهرحال، انتخاب مقدار آستانه کاملاً به توزیع انتخاب شده و فرض استقلال بستگی دارد. بنابراین رویکردهای متفاوت می‌تواند به انتخاب آن کمک نماید. لذا خطوط کلی زیر که می‌تواند مقدار آستانه را معین سازد، بشرح زیر می‌باشد:

۱) مشخص نمودن محدوده‌ایی از آستانه که آزمون‌های ۲ و ۳ را محقق می‌سازد.

۲) انتخاب بزرگترین مقدار آستانه که در آن μ بزرگتر از ۲ یا ۳ باشد.

13) Rosbjerg and Madsen

۳ مطالعه فرآیند وقوع

فرآیند وقوع حوادث E یا توسط مدت زمان بین دو حادثه متوالی θ تعیین می‌شود، یا تعداد حوادثی (m_t) است که در بازه زمانی $[۰, t]$ رخ می‌دهد. در هر یک از حالات، می‌توان توزیع احتمال و مقدار میانگین را داشت.

۱) فاصله بین دو حادثه

$$F(x) = \Pr[\theta < x],$$
$$f(x)dx = \Pr[x < \theta < x + dx]$$

دوره بازگشت حادثه بصورت زیر تعریف می‌شود:

$$T = E(\theta) = \int_{۰}^{\infty} \theta f(\theta)\, d\theta \qquad (۳)$$

۲) تعداد m_t حادثه در فاصله زمانی $[۰, t]$ رخ داده است که $w_k(t) = \Pr[m_t = k]$ می‌باشد. همچین می‌توان $E(m_t)$ را روی فاصله زمانی $[۰, t]$ تعریف نمود، و از اندیس پراکنش $I_t = \frac{var(m_t)}{E(m_t)}$ نیز می‌توان استفاده کرد.

لانگ[۱۴] (۱۹۹۷) خواص توزیع‌های فرآیند را نشان داد. در این مرحله آزمونها و رویه استفاده از آنها برای انتخاب توزیع و مقدار آستانه تشریح می‌شود.

۱.۳ آزمون اندیس پراکنش

از آزمون $\chi^۲$ می‌توان برای نیکویی برازش استفاده نمود، توصیه می‌شود که m_t را به حداقل پنج دسته با دست کم هفت و یا هشت عضو در هر دسته تقسیم نمود. معمولاً آماری‌های دیگری برای رفع محدودیت آزمون $\chi^۲$ پیشنهاد می‌شود. کانن[۱۵] (۱۹۷۹) برای توزیع پواسون اندیس پراکنش (I_t) را پیشنهاد نمود. فرض می‌کنیم که تعداد سیلابهای سالانه دارای توزیع نرمال است، مشروط بر آنکه پارامتر توزیع پواسون بیشتر از ۵ باشد و با توجه به اینکه میانگین و واریانس توزیع پواسون با هم برابر است، چنین بدست می‌آید:

۱) عبارت $h = \sum_{i=۱}^{NY}\{[m_۱(i) - E(m_۱)]/[var(m_۱)^{۱/۲}]\}^۲$ از توزیع $\chi^۲$ پیروی می‌کند که دارای درجه آزادی NY است.

۲) عبارت $\hat{h} = \sum_{i=۱}^{NY}[m_۱(i) - \overline{m}_۱]^۲/\overline{m}_۱ = (NY - ۱)\hat{i}_۱$ از توزیع $\chi^۲$ پیروی می‌کند که دارای درجه آزادی $NY - ۱$ است (یک درجه آزادی برای برآورد پارامتر توزیع

14) Lang et al. 15) Cunnane

توزیع پواسون کم می‌شود) و $m_۱(۱), m_۱(۲), \cdots, m_۱(NY)$ تعداد حوادث در NY سال متوالی است. در حالت نظری مقدار اندیس پراکنش برابر واحد است.

لذا داریم:

۱) اگر $\hat{i}_۱ < I_۱(\omega/۲)$ باشد، فرض پواسون بودن با سطح معنی‌دار $\omega\%$ حذف می‌شود و از توزیع دوجمله‌ای $\hat{i}_۱ < ۱$ استفاده می‌گردد.

۲) اگر $\hat{i}_۱ \in [I_۱(\omega/۲), I_۱(۱-\omega/۲)[$ باشد، فرض پواسون بودن پذیرفته می‌شود.

۳) اگر $\hat{i}_۱ > I_۱(۱-\omega/۲)$ از توزیع دوجمله‌ای منفی می‌توان استفاده نمود $(\hat{i}_۱ > ۱)$

با توجه به $I_۱(p) = \chi^۲(p)/(NY-۱)(p = \omega/۲ \; or \; p = ۱-\omega/۲)$ است.

۲.۳ آزمون ایستایی

لانگ[16] (۱۹۹۵) فاصله تولرانس m_t را در فاصله $[۰, t]$ پیشنهاد کرد که در شکل شماره ۳ نشان داده شده است. اگر NF تعداد نقاط اوج مشاهده شده در فاصله $[۰, t_{end}[$ باشد و $۰ < t < t_{end}$ و $۰ \leq m_t \leq NF$، آنگاه داریم:

$$-\mu t \leq \epsilon_t \leq NF - \mu t \quad , \; \epsilon_t = m_t - \mu t \qquad (۴)$$

ما پیشنهاد می‌کنیم که فرآیند سیلاب توسط یک فرآیند پواسون همگن تبیین شود:

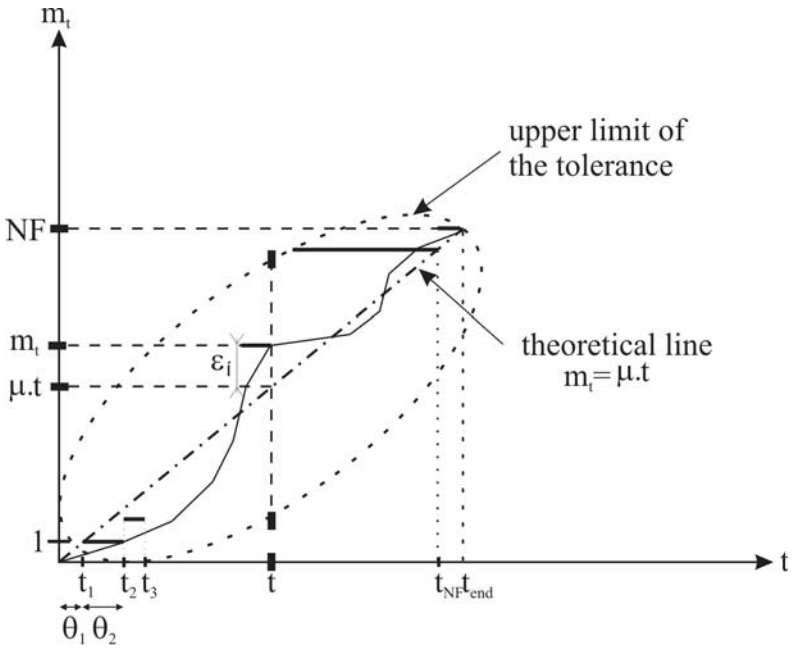
$$w_k(t) = \Pr(m_t = k) = e^{-\mu t}\frac{(\mu t)^k}{k!} \; , \; E(m_t) = \mu t$$

اگر ما احتمال شرطی داشتن k نقطه اوج در فاصله $[۰, t]$ را با $w^*_k(t)$ نشان دهیم، آنگاه NF نقطه اوج در فاصله $[۰, t_{end}]$ خواهیم داشت و می‌توان نوشت:

$$\begin{aligned} w^*_k(t) &= \Pr[m_t = k | m_{t_{end}} = NF] \\ &= [w_k(t) w_{NF-k}(t_{end} - t)]/[w_{NF}(t_{end})] \\ &= \binom{NF}{k}(t/t_{end})^k (۱ - t/t_{end})^{NF-k} \end{aligned} \qquad (۵)$$

با توجه به عبارت $w^*(t) = w^*_{NF-k}(t_{end} - t)$، فاصله تولرانس متقارن و به نقطه‌ای به مختصات $(t_{end}/۲, NF/۲)$ است.

16) Lang



شکل ۳: نمودار فاصله تولرانس تعداد مقادیر بالاتر از آستانه

۳.۳ آزمون فصلی

اردا و دیگران[17] (۱۹۹۳) نشان دادند که روند فصلی فرآیند جریان رودخانه می‌تواند تاثیر بسزایی روی توزیع مقادیر حدی بگذارد. در رویکرد فصلی دو عامل اهمیت دارد، مطالعه یک عامل که توام با ریسک است و دوم عدم ایستایی است که رخ می‌دهد.

۴ مدل نمودن نقاط اوج بالاتر از یک آستانه

اگر X یک متغیر تصادفی باشد، X_s مقدار حداکثر در هر حالت سیلابی را نشان می‌دهد، که بالاتر ار مقدار آستانه S است. اکنون می‌توان تابع چگالی احتمال این متغیر را بدست آورد:

$$G_s(x) = \Pr[X_s < x] \qquad (۶)$$

دوره بازگشت $T(x)$ را می‌توان بصورت زمان متوسط بین مقدار متوالی X_s که بزرگتر از x است، تعریف نمود. رزبرگ[18] (۱۹۸۵) رابطه بین دوره بازگشت و توزیع G_s بصورت رابطه (۷) تعریف نمود:

$$G_s(x) = ۱ - \frac{۱}{\mu T(x)} \qquad (۷)$$

۱.۴ انتخاب توزیع نقاط اوج

اولین گام در تعیین توزیع این است که سری داده‌ها در شرایط استقلال، همگنی و ایستایی صدق کند دومین گام استفاده از یک توزیع مناسب و برآورد پارامترهای آن (به روش گشتاورها، حداکثر درستنمایی، و گشتاور وزنی احتمال).

توزیع نمایی و در حالت کلی GP دارای این خاصیت هستند که تابع چگالی شرطی آنها نیز به ترتیب برای هر مقدار آستانه S نمایی و GP می‌باشند. این خاصیت توسط آشکار و روسل[19] (۱۹۸۳) و ونگ[20] (۱۹۹۱) رویکردی برای استفاده از این توزیعها برای مقادیر بالاتر از یک آستانه شد. رویکرد شرطی بشرح زیر است:

$$\begin{aligned} G_{s^*}(x) &= \Pr[X_{s^*} < x] = \Pr[X_s < x | X_s > S^*] \\ &= \frac{\Pr[S^* < X_s < x]}{\Pr[X_s > S^*]} \end{aligned}$$

و

$$G_{s^*}(x) = \frac{G_s(x) - G_s(S^*)}{۱ - G_s(S^*)} = ۱ - \frac{۱ - G_s(x)}{۱ - G_s(S^*)} \qquad (۸)$$

17) Ouarda et al. 18) Rosbjerg 19) Ashkar and Rousselle 20) Wang

۵ توزیع حداکثر سیلاب سالانه

۱.۵ ارتباط بین توزیع‌های مقادیر بالاتراز آستانه و مقادیر حداکثر سالانه

متغیر X^* مقادیر حداکثر سالانه X با تابع توزیع F_x. روابط زیر توسط شان و لین[21] (۱۹۶۴) ارائه شده است:

$$F_{X^*}(x) = \Pr[X^* < x] = \sum_{k=0}^{\infty} w_k(1)[G_s(x)]^k \qquad (9)$$

در حالت عمومی

$$F_{X^*}(x) = \exp\{-\mu[1 - G_s(x)]\} \quad (Poisson\ process) \qquad (10)$$

معادله شماره (۱۰) به نتایج زیر منجر می‌شود:

۱) مدل مقادیر بالاتر ازآستانه با توزیع نمایی برای سیلابهای سالانه به سمت توزیع گامبل میل می‌کند:

$$\begin{aligned} G_s(x) &= 1 - \exp[-(x-S)/a] \\ F_x(x) &= \exp\{-\mu \exp[-(x-S)/a]\} \end{aligned}$$

۲) مدل مقادیر بالاتر از آستانه باتوریع پارتو تعمیم یافته[22](GP) برای مقادیر حداکثر سالانه به سمت توزیع مقادیر حدی تعمیم یافته[23] (GEV)

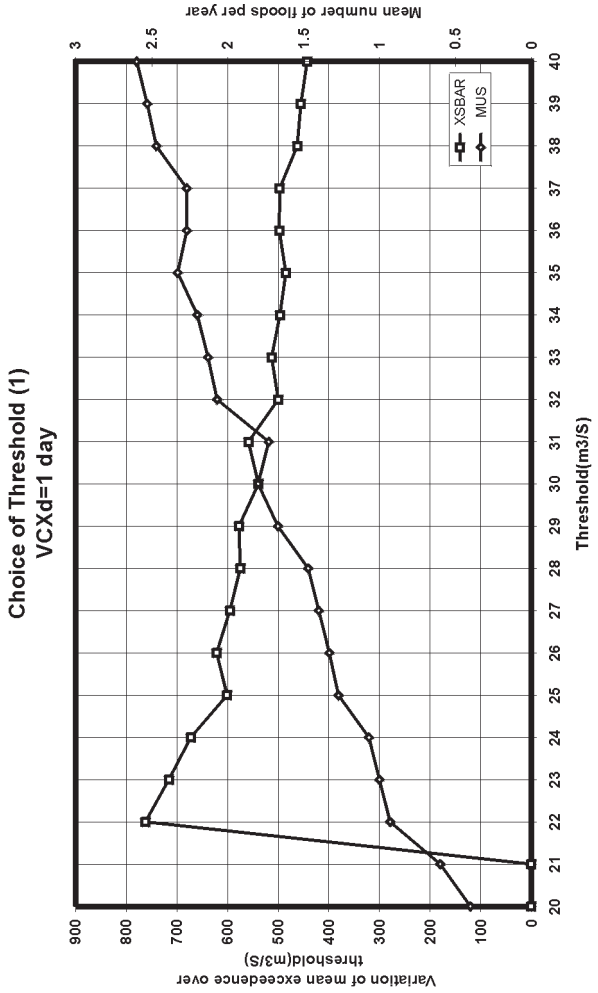
$$\begin{aligned} G_s(x) &= 1 - [1 - (k/a)(x-S)]^{1/k} \quad (if\ k \neq 0) \\ F_x(x) &= \exp\{-\mu[1 - (k/a)(x-S)]^{1/k}\} \end{aligned}$$

۶ مطالعه حالتی

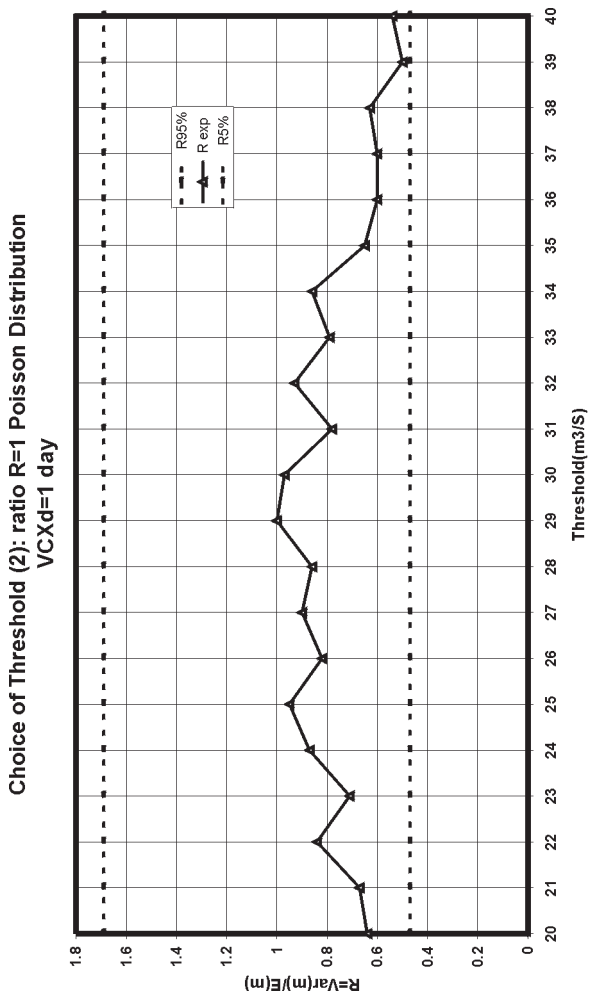
در این قسمت آمار دبی[24] متوسط روزانه یک ایستگاه هیدرومتری (در یک حوضه آبریز) اخذ شده است. در اولین بررسی سری زمانی داده‌ها رسم می‌شود، که از روی آن می‌توان محدوه‌ای را برای مقدار آستانه حدس زد. اما برای اینکه بتوان تغییرات متوسط تعداد داده‌ها و متوسط مقدار آنها را ملاحظه نمود، به کمک نرم‌افزارهای مربوطه، در محدوده‌ای که مد نظر است مقادیر متوسط‌ها محاسبه شده و در شکل شماره ۴، مشاهده می‌گردد.

شرط پواسون بودن تعداد داده‌های بالاتر از آستانه بر حسب مقادیر مختلف آن، توسط اندیس پراکنش کنترل شده است، که در شکل شماره ۵ ملاحظه می‌شود.

21) Shan and Lynn 22) Generalized Pareto 23) Generalized Extreme Value
24) discharge



شکل ۴: انتخاب مقدار آستانه (آزمون ۱: $\mu(S) = f_1(S)$، آزمون ۲: $(\overline{X}_s - S = f_2(S)$



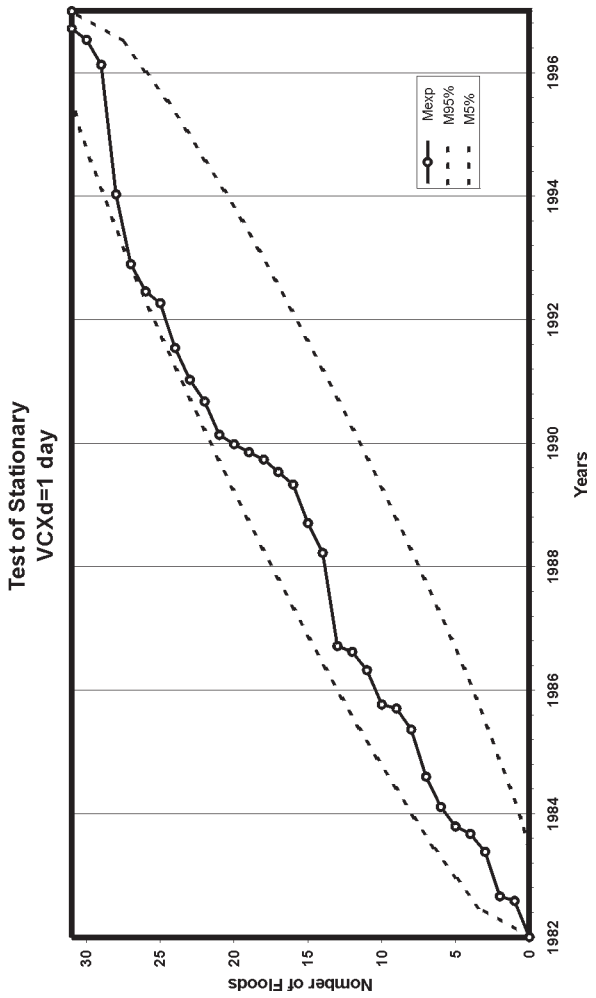
شکل ۵: آزمون اندیس پراکنش (آزمون ۳)

از روی دو شکل ۴ و ۵، مقدار $۲۵\ m^۳/s$ را برای آستانه انتخاب می‌کنیم و سپس با استفاده از این مقدار می‌توان شرط ایستایی را کنترل نمود. مورد اخیر در شکل شماره ۶ قابل رویت است.

۷ نتیجه‌گیری

به علت رویکرد چندگانه روش داده‌های بالاتر از یک آستانه معین، چالش‌هایی در این روش مطرح شده است که موجب شهرت آن در بین کاربران شده است. در این مقاله سعی بر این شد که خطوط کلی که دارای همگرایی بیشتری بین صاحب نظران است برای استفاده از روش بالاتر از آستانه ارائه گردد. مساله عمده این روش، انتخاب مقدار آستانه است، همانطور که ملاحظه شد نمی‌توان مقدار واحدی را برای آستانه بدست داد. اما می‌توان یک محدوده‌ای برای آستانه‌ها برگزید که در تجزیه و تحلیل‌های بعدی به نتایج یکسانی رسید. برای دست‌یابی به چنین محدوده‌ای باید آزمون‌های مطرح شده در این مقاله را مورد استفاده قرار داد.

سرانجام علیرغم مزایای این روش، توصیه می‌شود که تجزیه و تحلیل فراوانی برای روش مذکور و داده‌های حداکثر سالانه انجام شود، زیرا مقایسه نتایج آنها با یکدیگر مقایسه خوبی برای این دو روش می‌باشد.



شکل ۶: آزمون ایستایی فرآیند

مراجع

[1] Ashkar, F., Rousselle, J. (1983) .The effect of certain restrictions imposed on the interarrival times of flood events on the poisson distribution used for modeling flood counts., Water Resources Res. 19(2), 481-485.

[2] Ashkar, F., Rousselle, J. (1987). Partial duration series modeling under the assumption of a poissonian flood, count. J. Hydrol 90, 135-144.

[3] Cunnane, C. (1979). A note on the poisson assumption in partial duration series models, Water Resources Res. 15(2), 489-494.

[4] Davison, A.C., Smith, R.L. (1990) Models for exceedances over high threshold, J. R. Stat. Soc. B52(3), 393-442

[5] Lang, M., (1995) Les chroniques en hydrologie. PhD thesis, Université Joseph Fourier Grenoble, Cemagref Lyon, France, 296 p.

[6] Lang, M., Rassmussen, P.F., Oberlin, G., Bobée, B. (1997) Over-threshold sampling: modeling of occurrences by renewal proccesses, Revue des Sciences de l'Eau 3, 279-320

[7] Miquel, J. (1984) Guide pratique d'estimation des probabilités de crue, Eyrolles, Paris 160 p.

[8] Naden, P.S., Bayliss, A.C., (1993). Flood estimation: peak-over-threshold techniques., MaFF Conf. of River and Coastal Engineers, University of Loughborough, UK, 5-7 July, pp. 9.1.1-9.1.18

[9] Ouarda, T.B, Ashkar, F.,El-Jabi, N. (1993). Peaks over threshold model for seasonal flood variation, In: Kuo, C.Y. (Ed.), Engineering Hydrology (Pro. of symposium of San Francisco, California, 25-30 July), ASCE Publication pp. 341-346

[10] Rosbjerg, D., (1985). Etimation in partial duration series with independent and dependent peak values, J. Hydrol. 76, 183-195.

[11] Rosbjerg, D., Madsen, H., (1992). On the choice of threshold level in partial duration series, XVII Nordic Hydrological Conference, Alta, Norway, NHP Rep no. 30, pp. 604-615 23 (3), 245-256.

[12] Shane, R.M., Lynn, W.R. (1964) Mathematical model for flood risk evaluation, J. Hydraul. Div. ASCE 90 (HY6), 1-20

[13] Wang, Q.J. (1991) The POT model described by the generalized Pareto distribution with Poisson arrival rate., J. Hydrol. 129, 263-280

روش‌های نوین اقتصاد سنجی در تحلیل داده‌های مکانی

رزیتا مویدفر، آذر ابراهیمی

سازمان مدیریت و برنامه‌ریزی استان اصفهان

چکیده: همواره دانشمندان علوم اقتصادی در جهت دستیابی به تحلیل‌های منطقی و نزدیکتر به واقعیت ویژگی‌های اقتصادی جوامع، به شیوه‌های کمی در پردازش داده‌های آماری توجه داشته‌اند. این مسأله شاخه‌ای را در علم اقتصاد بوجود آورد، به نام اقتصادسنجی که توانسته در چند دهۀ اخیر علاقمندان بسیاری را به خود مشغول و از ابعاد گوناگون ماهیت و روش، پیشرفتهای قابل توجهی نماید. در این رابطه مدلهای اقتصادی به سه دسته تقسیم می‌شوند: ۱ـ ad hoc، ۲ـ time series و ۳ـ Econometrics.

استفاده از مدلهای سری زمانی با فرض همگن بودن مشاهدات (individuals) با تغییر بخش‌ها یا واحدهای اقتصادی توأم می‌باشد و این مسأله بعضا نتایج را دچار اریب می‌سازد. در حالیکه مدلهای panel data به عنوان یکی از پیشرفتهای اخیر در تخمین معادلات و تحلیل داده‌ها به روش اقتصادسنجی فرض ناهمگن بودن متغیر میان واحدهای مختلف اقتصادی را در یک روند زمانی امکان‌پذیر ساخته و با افزایش حجم مشاهدات و درجه آزادی در تخمین موجب کارائی بیشتر و واریانس کمتر مدل می‌گردد.

اما از طرف دیگر تخمین مدلهای panel به روش اقتصادسنجی مرسوم به هنگام مواجهه با داده‌های مکانی، به دلیل وجود اثرات وابستگی فضایی و ناهمسانی فضایی منجر به نقض فروض گاس ـ مارکف در تخمین حداقل مربعات معمولی می‌گردد.

یکی از تحولات و پیشرفتهای ایجاد شده در بکارگیری روشهای کمی، تکامل شاخۀ اقتصادسنجی به اقتصادسنجی فضایی است. این تکنیک با وارد ساختن ماتریس مجاورت امکان اندازه‌گیری اثرات ناشی از مجاورتهای مکانی را فراهم کرده و مشکلات ناشی از تحلیل داده‌های مکانی به روش اقتصادسنجی مرسوم را برطرف می‌سازد.

این مقاله تلاش می‌کند تا علاوه بر ارائه ادبیات نظری مدل‌های panel، به مفهوم و موضوع اقتصادسنجی فضایی پرداخته و سپس با برقراری ارتباط بین این دو تکنیک، مدل مناسبی را در تحلیل داده‌های مکانی (منطقه‌ای) معرفی می‌نماید.

واژه‌های کلیدی: داده‌های مکانی، داده‌های پانل، اقتصادسنجی فضایی

۱ مقدمه

اقتصادسنجی به عنوان ترکیبی از تئوریهای اقتصادی، اقتصاد ریاضی، آمار اقتصادی و آمار ریاضی امکان کمی‌سازی مدلهای اقتصادی را به منظور بررسی احکام و فرضیه‌های کیفی اقتصاد فراهم ساخته است.

هر چند به اعتقاد بسیاری از اقتصاددانان بیان روابط اقتصادی در قالب معادلات ریاضی و تخمین عددی آنها موجب ایجاد انحراف در پیش‌بینی مسیر اقتصاد گردیده و سیاستگذاران اقتصادی را در انتخاب سیاستهای مناسب دچار انحراف می‌سازد، اما واقعیت آن است که روش اقتصادسنجی در تایید و توسعه تئوریهای اقتصادی نقش بسزایی داشته و از این رو به عنوان یکی از مباحث پژوهشی در علم اقتصاد مطرح می‌باشد. این امر گروهی از اقتصاددانان را بر آن داشته تا اساس تحقیقات خود را بر ابداع، اصلاح و پیشرفت روشهای اقتصادسنجی و کاربرد آن در تایید یا رد تئوریهای اقتصادی استوار سازند و بدین وسیله موجب گسترش حوزه نفوذ این بخش از اقتصاد در سایر بخش‌های آن گردند. از جمله این موارد می‌توان بکارگیری روشهای اقتصادسنجی در بررسی مدلهای اقتصاد منطقه‌ای را نام برد. تئوریهای مطرح در این بخش از اقتصاد در بسیاری از موارد لزوم بررسی مشاهدات در دو بعد مکان و زمان را موجب می‌گردد و این مساله نتایج روشهای اقتصادسنجی معمول را دچار تورش می‌سازد. لذا اصلاح و پیشرفت این روشها به سمت روشهای نوین اقتصادسنجی در تحلیل داده‌های مکانی اجتناب‌پذیر بوده است. در این روند مدلهای پانل و تکنیک اقتصادسنجی فضایی با هدف کاهش خطا در تخمین مدلهای اقتصاد منطقه‌ای و تحلیل داده‌های مکانی شکل گرفت.

این مقاله می‌کوشد تا ضمن معرفی مباحث نظری این مدلها، مزیتهای عملی بکارگیری آن‌ها را طی تخمین یک مدل تجربی بر شمرد.

بخش اول این مقاله با ارائه مدل پایه اقتصادسنجی فروض اساسی تشکیل مدلهای اقتصادسنجی را معرفی می‌نماید. در بخش دوم و سوم مباحث نظری مدل پانل و تکنیک اقتصادسنجی فضایی بر مبنای خدشه‌دار شدن فروض اساسی مدلهای اقتصادسنجی معمولی در بکارگیری داده‌های مکانی شرح داده می‌شود و سپس بخش چهارم این مقاله نتایج بدست آمده در تخمین یک مدل تجربی اقتصاد منطقه‌ای را به روشهای اقتصادسنجی معمولی، پانل و اقتصادسنجی فضایی مورد مقایسه قرار می‌دهد و در نهایت بخش پنجم به نتیجه‌گیری از کل مباحث مطرح شده در این مقاله می‌پردازد.

۲ مدل پایهٔ اقتصادسنجی

علم اقتصاد شکل گرفته بر مبنای تئوریها و فرضیه‌های اقتصادی، همواره به دنبال کمی‌سازی متغیرهای کیفی و تصریح فرضیه‌ها در قالب مدلهای ریاضی به منظور فهم بهتر روابط حاکم بر اقتصاد جوامع بوده‌است.

در این رابطه مدلهای اقتصادی در سه دسته قابل طبقه‌بندی گردید:

۱ـ Ad hoc: به معنی پذیرفتن فرضی معین بدون نیاز به سنجش و اندازه‌گیری آن
۲ـ time series: به معنی وابسته بودن یک متغیر به سابقه خودش که اساس مدلهای خود رگرسیونی (AR)، میانگین متحرک (MA) و یا ترکیبی از این دو (Arma و Arima) را تشکیل می‌دهد.
۳ـ Econometrics: به معنی تخمین کمی و تجربی مدلهای اقتصادی ساخته شده بر مبنای روابط ریاضی که این مدلها می‌تواند داده‌های سری زمانی یا مقطعی و یا ترکیبی از این دو را بکار گیرد و با استفاده از نتایج بدست آمده تئوریهای اقتصادی را رد یا اثبات نمایند. به عبارت دیگر اقتصادسنجی را می‌توان به عنوان تحلیل کمی پدیده‌های اقتصادی در دنیای واقع بر مبنای بسط و توسعه همپای تئوری و مشاهده که توسط روشهای متناسب استنتاجی به یکدیگر مرتبط شده‌اند، تعریف نمود.

ابزار اصلی تحلیل اقتصادسنجی رگرسیون است که اولین بار توسط فرانسیس گالتون[1] (۱۸۸۶) و پس از آن کارل پیرسن[2] (۱۹۰۳) مطرح شد که در بررسی ارتباط بین طول قد فرزندان و والدین مورد استفاده قرار گرفت. به طور کلی می‌توان گفت تحلیل‌های رگرسیون به مطالعه وابستگی یک متغیر (متغیر وابسته) به یک یا چند متغیر دیگر (متغیر توضیحی) می‌پردازد که با تخمین یا پیش‌بینی مقدار متوسط یا میانگین مقادیر متغیر نوع اول در حالتی که مقادیر متغیر نوع دوم معلوم یا معین شده باشد (در نمونه گیری تکراری)، صورت می‌پذیرد. با توجه به این تعریف می‌توان نوشت:

$$E(Y|x_i) = f(x_i)$$

که در آن $f(x_i)$ تابعی از متغیرهای X_i می‌باشد. این تابع صرفاً بیان می‌کند که میانگین توزیع Y بر حسب x_i معلوم، به طور تبعی به x_i مربوط است. به عبارت دیگر بیان می‌دارد که چگونه مقدار میانگین Y بر حسب xها تغییر می‌کند.

یک نوع ساده از این تابع رگرسیون خطی است که به شکل زیر نوشته می‌شود:

$$E(Y|x_i) = B_1 + B_2 x_i$$

از طرف دیگر می‌دانیم که y_i خاص در جامعه لزوماً منطبق بر میانگین شرطی خودش نبوده و دارای انحراف می‌باشد. لذا می‌توانیم انحراف y_i را در اطراف امید خودش همانند ذیل بدست آوریم:

$$U_i = Y_i - E(Y|x_i)$$

بنابراین با توجه به روابط قبلی خواهیم داشت:

$$Y_i = B_1 + B_2 x_i + u_i$$

اما واقعیت این است که در بسیاری از موارد دسترسی به مشاهدات جامعه کاری است دشوار و تنها مشاهدات نمونه‌ای از متغیرهای مورد نظر قابل اندازه‌گیری می‌باشند. پس تخمین تابع

1) Francis Galton 2) Karl Pearson

رگرسیون بر مبنای مشاهدات نمونه به صورت:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + e_i$$

به عنوان یک تقریب، بیانگر رابطهٔ تابعی رگرسیون جامعه خواهد بود.

یکی از روشهای تخمین پارامترهای مدل روش حداقل مربعات مستقیم (OLS) می‌باشد که اولین بار توسط کارل فردریک گوس[3] ریاضیدان آلمانی ارائه شد. همانطور که قبلا گفته شد، مقادیر x_i داده شده در تابع رگرسیون مقادیری از Y_i را به صورت $Y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + e_i$ بدست می‌دهد که برابر با میانگین شرطی Y_i بوده و نسبت به مقادیر واقعی Y_i دارای مقدار انحرافی به اندازه e_i می‌باشد. روش حداقل مربعات معمولی بر مبنای حداقل نمودن مجموع مربعات عوامل خطا $\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$ دارای خصوصیات ویژه‌ای است که این روش را به صورت یکی از مشهورترین و قویترین روشهای تحلیل رگرسیون در آورده است. فروض اساسی حاکم بر تخمین زننده‌ها حداقل مربعات به صورت:

$۱ - E(u_i|x_i) = ۰$
$۲ - cov(u_i, u_j) = ۰$
$۳ - var(u_i|x_i) = \sigma^2$
$۴ - cov(u_i, x_i) = ۰$

مدل رگرسیون دقیقاً تصریح شده است (عدم وجود خطای تصریح یا تورش) به گونه‌ای است که مطابق قضیه گوس ـ مارکف، تخمین زننده‌های حداقل مربعات را در بین تخمین زننده‌های خطی، بدون تورش و دارای حداقل واریانس می‌سازد یا به عبارت دیگر بهترین تخمین‌زن بدون تورش خطی (BLUE) می‌باشند. اما تغییر متغیرها در دو بعد مکان و زمان و فرض ناهمگنی بین واحدهای مختلف در یک مقطع از زمان و بین زمانهای مختلف وجود یک واحد، رابطه خطی رگرسیون را به صورت:

$$Y_{it} = \alpha_i + B_i x_{it} + u_{it}$$

تغییر داده و شرایطی را فراهم می‌سازد که موجب نقض قضیه گوس - مارکف گردیده و تخمین به روش حداقل مربعات معمولی را دچار تورش می‌سازد. در این حالت نتایج حاصل از تخمین به روش حداقل مربعات معمولی، دیگر BLUE نبوده و در اینجا روشهای جایگزین برای حداقل نمودن واریانس تخمین زننده‌ها مطرح می‌گردد. در این رابطه نه تنها مدل پانل به عنوان روشی برای تخمین ضرائب رگرسیون امکان لحاظ نمودن فروض ناهمگنی بین واحدهای مختلف را در طول زمان فراهم می‌سازد، بلکه تکنیک اقتصادسنجی فضایی نیز با وارد ساختن متغیر وابستگی فضایی، موجب کاهش خطای تخمین ناشی از وابستگی و ناهمسانی فضایی می‌گردد.

بر این اساس در بخش‌های بعدی اساس مدل پانل و تکنیک اقتصادسنجی فضایی به تفصیل توضیح داده خواهدشد.

3) Kard Fredric Gauss

۳ مدلهای مشبکه‌ای(پانل)

در قسمت قبل نشان دادیم که یک معادله رگرسیون به دنبال تخمین رابطه‌ای علّی بین یک متغیر وابسته و یک یا چند متغیر مستقل می‌باشد. آنچه تاکنون بدان پرداخته شد، تخمین آماری مشاهداتی بود که تغییرات آنها در یک بعد صورت می‌گرفت. به عبارت دیگر تغییرات یا در طول زمان مشاهده می‌شد و یا در بین واحدهای مختلف. مدلهای پانل این امکان را فراهم می‌سازد که بتوانیم تفاوت بین داده‌ها در طول زمان را در میان واحدهای مختلف در نظر بگیریم. به این معنی که در یک سری زمانی مثلاً ۱۰ ساله برای هر سال می‌توان تغییرات مشاهدات را برای ۱۰ واحد ناهمگن مثلاً ۱۰ خانوار یا ۱۰ بنگاه اقتصادی و یا ۱۰ منطقه تجاری نیز مورد بررسی قرار داد. لذا در تخمین مدلهای پانل به جای استفاده از متغیرها در حالت برداری از متغیرها به صورت ماتریس استفاده می‌شود. که البته این ماتریس‌ها بر اساس مشاهدات در دسترس می‌توانند متقارن یا نامتقارن باشند و در نتیجه پانل‌های متوازن (balance) یا نامتوازن (unbalance) را خواهیم داشت.

با توجه به این توضیحات می‌توان مزایای مدلهای پانل را به صورت زیر طبقه‌بندی نمود:

۱ــ کنترل همگنی یا ناهمگنی بین واحدهای مختلف است که از تصریح نادرست (misspecification) در نتایج جلوگیری می‌کند.

۲ــ افزایش حجم مشاهدات که موجب افزایش درجه آزادی و کارایی بیشتر مدل و واریانس کمتر می‌گردد.

۳ــ مدل پانل پویائی‌های تعدیل (dynamics of adjustment) را مورد مطالعه قرار می‌دهد. مثلاً در سنجش بیکاری داده‌های مقطعی می‌توانند نسبتی از جمعیت را تخمین بزنند که در یک زمان خاص بیکار بوده اند ولی همان نسبت ممکن است در زمان دیگر بیکار نباشد. پس پانل قادر است رفتار افراد را از یک دوره به دوره دیگر مورد بررسی قراردهد.

۴ــ در بسیاری موارد اثرات متغیرها روی متغیرهای دیگر به آسانی در مدلهای مقطعی و سریهای زمانی قابل شناسائی نیست. ولی در مدل پانل با در نظر گرفتن تغییرات مقطعی و زمانی این امکان بوجود می‌آید.

۵ــ در پانل به دلیل افزایش کارائی تخمین‌ها می‌توان مدلهای رفتاری پیچیده‌تری را طراحی کرده و تخمین زد و در واقع قابلیت مدل از لحاظ آماری نیز افزایش می‌یابد.

۶ــ بررسی مسائل در سطح خرد موجب می‌شود در بسیاری از موارد داده‌ها به صورت تجمعی (aggrigate) در نظر گرفته شوند که این موجب تورش یا اریب در داده‌ها در قلمرو خرد می‌شود. در مدل پانل امکان بررسی داده‌ها به صورت غیر تجمعی فراهم شده و تورش ناشی از تجمعی بودن داده‌ها بر طرف می‌شود.

حالت کلی در مدلهای پانل به صورت:

$$y_{it} = \alpha_{it} + \beta x_{it} + u_{it}, \qquad \begin{matrix} i = ۱, \ldots, N \\ t = ۱, \ldots, T \end{matrix}$$

می‌باشد.

در اینجا α_{it} عرض از مبدأ است x_{it} مجموعه متغیرهای مستقل به صورت itمشاهده روی k متغیر توضیحی می‌باشد. B یک بردار به صورت $K \times ۱$ و نشان دهنده و در طول زمان ثابت است. مانند عامل مدیریت که برای واحدهای مختلف متفاوت ولی در زمان ثابت است. جزء بعدی شامل شیب پانل است.u_{it} جزء خطا و شامل سه قسمت است. جزئی از آن مربوط به متغیرهایی است که برای داده‌های مقطعی متفاوت متغیرهایی است در زمان مشابه، یکسان، اما در طول زمان متفاوت است. مانند قیمت، نرخ بهره و انتظارات نسبت به آینده. قسمت سوم جزئی است که نه تنها در طول زمان بلکه در مقطع هم تغییر می‌کند مثل سود که برای واحدهای مختلف و در زمان‌های مختلف، متفاوت است. در آزمون معنی‌داری ضرائب مدل پانل ۴ حالت قابل بررسی است:

حالت اول مبتنی بر همگن بودن شیب و نا همگن بودن عرض از مبداء می‌باشد به صورت:

$$H_۰ : \begin{array}{l} \alpha_i^* \neq \alpha_j^* \\ \beta_i = \beta_j \end{array}$$

که در صورت تایید فرض $H_۰$ مدل پانل به صورت $Y_{it} = \alpha_i^* + B' x_{it} + u_{it}$ خواهد بود.

حالت دوم فرض می کند عرض از مبداء همگن و شیب پانل ناهمگن می‌باشد به صورت:

$$H_۰ : \begin{array}{l} \alpha_i^* = \alpha_j^* \\ \beta_i \neq \beta_j \end{array}$$

که در صورت تایید فرض $H_۰$ خواهیم داشت $Y_{it} = \alpha_i^* + b_i' x_{it} + u_{it}$

حالت سوم هر دو عرض از مبدأ و شیب پانل را ناهمگن فرض می‌کند و به صورت زیر بررسی می‌شود.

$$H_۰ : \begin{array}{l} \alpha_i^* \neq \alpha_j^* \\ \beta_i \neq \beta_j \end{array}$$

که در صورت تایید فرضیه $H_۰$ مدل پانل به صورت $Y_{it} = \alpha_i^* + B_i^* x_{it} + u_{it}$ خواهد بود.

و حالت چهارم فرض می‌کند که هر دو عرض از مبدأ و شیب پانل همگن بوده و فرضیهٔ $H_۰$ را به صورت زیر تعریف می‌کند:

$$H_۰ : \begin{array}{l} \alpha_i^* = \alpha_j^* \\ \beta_i = \beta_j \end{array}$$

که در صورت تأیید فرضیهٔ $H_۰$ مدل پانل به صورت $Y_{it} = \alpha^* + B' x_{it} + u_{it}$ خواهد بود. در این حالت اساس همان مدل OLS تعمیم یافته است و اصطلاحاً به آن pooling data می‌گویند. نتایج مدل پانل از سه روش تخمین قابل محاسبه است:

۱ـ pooled data
۲ـ fixed effect
۳ـ Random effect

روش pooled data با فرض وجود همگنی بین ضرائب واحدهای مختلف به صورت OLS تعمیم یافته یا GLS(general least square) می‌باشد. در این روش فرضیهٔ H_0 به صورت:

$$\alpha_1 = \alpha_2 = \Lambda = \alpha_n$$

$$\beta_1 = \beta - 2 = \Lambda = \beta_n$$

تعریف می‌شود.

برای مقایسه وگزینش این روش نسبت به دو روش دیگر آمارهٔ F به صورت:

$$F\left(n-1, nT-n-k\right) = \frac{\left(R_u^2 - R_p^2\right)/n-1}{\left(1-R_u^2\right)/\left(nT-n-k\right)}$$

مورد بررسی قرار می گیرد.

در اینجا R_u^2 ضریب همبستگی حاصل از تخمین مدل به صورت دو روش دیگر و R_p^2 ضریب همبستگی حاصل از تخمین مدل به صورت روش pooled می‌باشد. در صورتی که آمارهٔ F از F جدول با تعداد مشاهدات (n) و تعداد متغیرها (k) و احتمال خطای معین بزرگتر باشد، فرضیهٔ H_0 رد شده و روش‌های منتخب برای تخمین با نتایج کاراتر fixed effect یا Random effect خواهند بود.

در این دو روش فرض تخمین مدل مبتنی بر ناهمگنی بین ضرائب عرض از مبدأ و همگنی ضرائب شیب در میان مشاهدات می‌باشد، $B_i = B_j$ و $\alpha_i \neq \alpha_j$. فرض بر این است که α_i ها دارای تابع توزیع احتمال هستند، به میانگین α و δ_α^2 واریانس و مستقل از رگرسورها توزیع شده‌اند.

آزمون هاسمن دارای توزیع χ^2 (کای دو) بوده و امکان انتخاب روش کاراتر را از میان دو روش fixed و Random فراهم می‌سازد. در این آزمون فرضیهٔ H_0 بر مبنای صحت روش Random به شکل زیر صورت می‌گیرد:

$H_0 : E(u_{it}|x_{it}) = 0$ که در آن $u_{it} = \varepsilon_{it} + \eta_i$، $\alpha_i = \alpha + \eta_i$ و $\hat{q} = \hat{B}_{RE} - \hat{B}_{FE}$

در اینجا x_{it} متغیرهای مستقل بوده و u_{it} عامل خطا می‌باشد که جزئی از آن به صورت η_i مربوط به متغیرهایی است که برای داده‌های مقطعی متفاوت و در طول زمان ثابت است بنابراین این جزء به صورت قسمتی از ضریب عرض از مبدأ (α_i) تعیین می‌شود.

$\hat{B}_{RE}$ ضرائب شیبی است که به روش Random effect به دست آمده و ضرائب شیب به دست آمده از روش fixed effect می‌باشد.

با فرض $cov(\hat{B}_{RE}, \hat{q}) = 0$ خواهیم داشت:

$$var(\hat{q}) = var(\hat{B}_{FE}) - var(\hat{B}_{RE})$$

و بر این اساس آمارهٔ هاسمن به صورت:

$$H = \hat{q}'[var(\hat{q})]^{-1} \quad \hat{q} \approx \chi_k^2$$

تعریف می‌شود که در اینجا k درجهٔ آزادی و تعداد رگرسورها می‌باشد. اگر آمارهٔ (H) کوچک باشد روش random کارا بوده و انتخاب می‌گردد و در غیر این صورت روش انتخابی اثر ثابت (fixed effect) برای تخمین ضرائب خواهد بود.

آنچه در معرفی مدل پانل به دست آمد، امکان بررسی تغییرات مشاهدات در طول زمان و در میان واحدهای مختلف به طور همزمان می‌باشد. یکی از حالت‌ها در این‌گونه تحلیل‌ها بررسی متغیرهای سری زمانی در میان مکان‌های مختلف است به عنوان مثال بررسی روابط روابط تجاری در یک دورهٔ ۲۰ ساله در میان کشورهای عضو یک منطقهٔ تجاری مانند اکو (ECO) می‌باشد. در یک چنین تحقیقاتی علاوه بر آنکه استفاده از مدل پانل اجتناب ناپذیر می‌باشد، متغیر دیگری که نمایانگر اثرات حاصل از مکان وقوع پدیده‌های تجاری است، ضرورت حضور می یابد و این نقطهٔ اتصال مدل های پانل به روش اقتصادسنجی فضایی است. از این رو در بخش بعدی به طور دقیق تکنیک اقتصادسنجی فضایی تشریح خواهد شد.

۴ اقتصادسنجی فضایی

گسترش کاربرد مدل‌های اقتصادسنجی در علوم منطقه‌ای به منظور تحلیل ارتباط موجود میان داده‌های کمی، تکنیک اقتصادسنجی فضایی را به عنوان زیر شاخهٔ اقتصادسنجی به وجود آورد. به گونه‌ای که این تکنیک اثر متقابل فضایی (خود همبستگی فضایی) و ساختار فضایی (ناهمگنی فضایی) را بررسی نموده و مدلهای رگرسیونی با داده‌های مقطعی یا پانل (panel) را مورد تخمین قرار می‌دهد.

همانطورکه در بخش یک نیز توضیح داده شد مدلهای حداقل مربعات معمولی بر پایهٔ استقرار فروض گاوس ـ مارکف به صورت استقلال همسانی واریانس مشاهدات دارای نتایجی بدون تورش خواهد بود. در حالیکه بکارگیری داده‌های مکانی در تخمین مدلهای اقتصادسنجی مرسوم نتایج را با دو مسأله مواجه خواهد ساخت:

۱ـ وابستگی فضایی موجود میان مشاهدات

۲ـ ناهمسانی فضایی و این دو مسأله باعث نقض فروض گاوس ـ مارکوف شده و نتایج بدست آمده را توأم با تورش خواهد ساخت.

۱.۴ وابستگی فضایی

وابستگی فضایی در مجموعه‌ای از مشاهدات نمونه اشاره به این حقیقت دارد که یک مشاهده مربوط به یک مکان (i) در ارتباط با مشاهدات مکانهای j $(j \neq i)$ قرار می‌گیرد. یعنی

$$Y_i = f(Y_j), \quad i = ۱, \ldots, n \quad i \neq j$$

i می‌تواند هر مقداری از ۱ تا n را بگیرد، یعنی در واقع وابستگی فضایی می‌تواند بین چند مشاهده باشد.

داده‌های مربوط به واحدهای فضایی مانند شهرها، استانها، ایالت‌ها، حوزه‌های سرشماری و . . . می‌توانند دارای خطای اندازه‌گیری باشند. این خطا به این دلیل رخ می‌دهد که مرزهای اداری برای جمع‌آوری اطلاعات دقیقاً منعکس کنندهٔ طبیعت فرآیند مربوط به ایجاد آن داده‌ها نیستند. مثلاً اندازه‌گیری نرخ بیکاری و اندازه نیروی کار، بر مبنای محل زندگی افراد می‌تواند نشانگر وابستگی فضایی باشد چون افراد برای یافتن شغل به شهرها یا استانهای مجاور می‌روند.

از طرف دیگر بعد فضایی فعالیت‌های اجتماعی-دموگرافیک، اقتصادی یا منطقه‌ای جنبهٔ مهمی از مسأله مدلسازی است. علوم منطقه‌ای بر مبنای این پیش فرض که مکان و فاصله نیروهای مهم دخیل در جغرافیای انسانی و فعالیت بازارند یا به عبارت دیگر فعالیت‌ها صرفاً تابعی از متغیرهای توضیحی شناخته شده نیست و تابعی از مجاورت و . . . است.

همهٔ این موارد در تئوری علوم منطقه‌ای موید وجود وابستگی فضایی و اثرات پراکندگی، سلسله مراتبی مکان‌ها و سرریزهای فضایی است.

۲.۴ ناهمسانی فضایی

اصطلاح ناهمسانی فضایی اشاره به تغییر روابط در طول فضا دارد. یعنی در هر نقطه از فضا انتظار وجود یک رابطه متفاوت را داریم:

$$y_i = x_i\beta_i + \epsilon_i, \quad i = ۱, \ldots, n$$

در اینجا i نشان‌دهنده نقاط موجود در فضا، x_i نشانگر بردار $(k \times ۱)$ از متغیرهای توضیحی همراه با مجموعه پارامترهای β_i مربوط به آن، y_iمتغیر وابسته در مکان i و ϵ_i بیانگر خطای تصادفی در رابطهٔ مذکور است. به عبارت دیگر خواهیم داشت:

$$y_i = f_i(x_i, \epsilon_i)$$

با توجه به رابطه مذکور هنگام حرکت در بین مشاهدات، توزیع داده‌های نمونه‌ای نشانگر میانگین و واریانس ثابتی نخواهد بود.

بر این اساس کمی‌سازی بعد مکان و ورود آن به مدلهای اقتصادسنجی به منظور رفع خطای ناشی از وابستگی فضایی و ناهمسانی واریانس مشاهدات مکانی، ما را به سمت طراحی مدلهای اقتصادسنجی فضایی رهنمون می‌سازد.

۳.۴ کمی‌سازی مکان

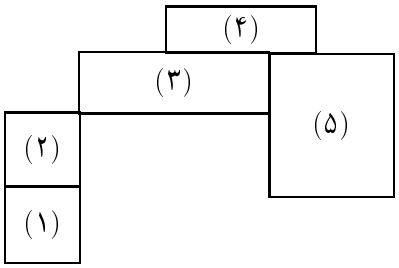
تعیین کمیت و مقدار عددی جنبه‌های مکانی به دو روش امکان‌پذیر خواهد بود.

۱ـ کمی‌سازی مجاورت فضایی

۲ـ کمی‌سازی موقعیت فضایی

۱.۳.۴ کمی‌سازی مجاورت فضایی

مجاورت منعکس کننده موقعیت نسبی واحد منطقه‌ای یک مشاهده در فضا نسبت به سایر واحدهاست. اطلاعات مجاورت را می‌توان از روی نقشه بدست آورد. وابستگی فضایی بین واحدهای مجاور بیشتر است. برای مثال یک نمونه ۵ ناحیه‌ای به شکل مقابل را درنظر می‌گیریم.



ماتریس W با ابعاد ۵ × ۵ شامل ۲۵ عنصر صفر و یک است که نشان دهنده مجاورت میان این ۵ منطقه خواهد بود. برای تعریف مجاورت راههای مختلفی در نظر گرفته می‌شود.

ـ مجاورت خطی: $W_{ij} = ۱$ برای عناصری که یک کناره مشترک بلافاصله از سمت راست یا چپ ناحیه مورد نظر با آن داشته باشند. در این نمونهٔ فرضی $W_{۵۳} = ۱$ خواهد بود.

ـ مجاورت رخ مانند: $W_{ij} = ۱$ برای نواحی که با ناحیه مورد نظر یک ضلع مشترک داشته باشد. در ۵ ناحیهٔ بالا $w_{۳۴} = ۱$ و $w_{۳۵} = ۱$ خواهد بود.

ـ مجاورت فیل مانند: $W_{ij} = ۱$ برای نواحی که یک گوشهٔ مشترک با ناحیه مورد نظر دارند. در اینجا $w_{۲۳} = ۱$ خواهد بود.

ـ مجاورت خطی دو سویه: برای نواحی که بلافاصله در سمت چپ یا راست ناحیه مورد نظر قرار دارند. برای این مثال، همان نتایج مجاورت رخ مانند بدست خواهد آمد.

ـ مجاورت رخ دو سویه: برای عناصر سمت راست، چپ، بالا و پایین $W_{ij} = ۱$ خواهد بود.

ـ مجاورت ملکه مانند: $W_{ij} = ۱$ برای نواحی که یک ضلع یا زاویه مشترک با ناحیه مورد نظر داشته باشند. انتخاب هرکدام از این تعاریف بستگی به ماهیت مسأله و اطلاعات غیر نمونه‌ای دارد. در مثال فوق ماتریس مجاورت بر حسب تعریف مجاورت رخ به صورت زیر بدست خواهد آمد.

$$W = \begin{bmatrix} ۰ & ۱ & ۰ & ۰ & ۰ \\ ۱ & ۰ & ۰ & ۰ & ۰ \\ ۰ & ۰ & ۰ & ۱ & ۱ \\ ۰ & ۰ & ۱ & ۰ & ۱ \\ ۰ & ۰ & ۱ & ۱ & ۰ \end{bmatrix}$$

این ماتریس یک ماتریس متقارن بوده و عناصر روی قطر اصلی صفر می باشد. عناصر ماتریس W غیرتصادفی و برونزا هستند. استانداردسازی عناصر این ماتریس بر حسب سطر به گونه‌ای

که مجموع عناصر هر سطر معادل یک شود موجب می‌شود تا از حاصلضرب ماتریس استاندارد مجاورت در بردار مربوط به متغیر وابسته برداری حاصل شود که عناصر آن میانگین مشاهدات نواحی مجاور باشد. به این ترتیب:

$$C = \begin{bmatrix} ۰ & ۱ & ۰ & ۰ & ۰ \\ ۱ & ۰ & ۰ & ۰ & ۰ \\ ۰ & ۰ & ۰ & \frac{۱}{۲} & \frac{۱}{۲} \\ ۰ & ۰ & \frac{۱}{۲} & ۰ & \frac{۱}{۲} \\ ۰ & ۰ & \frac{۱}{۲} & \frac{۱}{۲} & ۰ \end{bmatrix}$$

$$\begin{bmatrix} y^*_۱ \\ y^*_۲ \\ y^*_۳ \\ y^*_۴ \\ y^*_۵ \end{bmatrix} = cy = \begin{bmatrix} y_۲ \\ y_۱ \\ ۰/۵y_۴ + ۰/۵y_۵ \\ ۰/۵y_۳ + ۰/۵y_۵ \\ ۰/۵y_۳ + ۰/۵y_۴ \end{bmatrix}$$

$$\longrightarrow \quad y_i = f(y_j)$$
$$\longrightarrow \quad y = Pcy + \epsilon \qquad i \neq j$$

در اینجا c ماتریس مجاورت استاندارد شده و y^* یک متغیر توضیحی در تعریف y می‌باشد و p پارامتر نشان دهندهٔ وابستگی فضایی است. به این معنی که بخشی از کل تغییرات y در طول نمونه از طریق وابستگی آن به مناطق مجاورش توضیح داده می‌شود.

۲.۳.۴ کمی‌سازی موقعیت فضایی

مکان مربوط به مشاهدات در فضا در مدلسازی روابط ناهمسانی فضایی دارای اهمیت می‌شود. یکی از کسانی که در این زمینه به معرفی معرفی مدل پرداخت، کاستی[۴] می‌باشد. مدل مطرح شده به نام بسط فضایی است. به صورت زیر:

$$\beta = ZJ\beta_۰ \qquad Y = x\beta + \epsilon$$

در این مدل y بیانگر بردار $n * ۱$ متغیر وابسته مربوط به مشاهدات فضایی و x یک ماتریس $n * nk$ شامل اقلام x_i نشان دهندهٔ بردارهای $k * ۱$ متغیر توضیحی است.

$$y = \begin{bmatrix} y_۱ \\ y_۲ \\ \vdots \\ y_n \end{bmatrix} \qquad x = \begin{bmatrix} x'_۱ & ۰ & \dots & ۰ \\ ۰ & x'_۲ & \dots & ۰ \\ \dots & & & \dots \\ ۰ & \dots & & x'_n \end{bmatrix}$$

4) cosetti

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

اطلاعات مکانی در ماتریس z بیان شده که دارای عناصر zxi و zyi $I = 1, \ldots, n$ است که مختصات طول و عرض هر مشاهده را نشان می‌دهد.

$$Z = \begin{bmatrix} Z_{x_1} \otimes I_k & Z_{y_1} \otimes I_k & 0 & \cdots \\ 0 & \vdots & \vdots & 0 \\ \vdots & \vdots & Z_{x_1} \otimes I_k & Z_{y_1} \otimes I_k \end{bmatrix}$$

$$J = \begin{bmatrix} I_k & 0 \\ 0 & I_k \\ \vdots & \vdots \\ 0 & I_k \end{bmatrix} \qquad \beta_0 = \begin{bmatrix} \beta_x \\ \beta_y \end{bmatrix}$$

مدل نشان می‌دهد که پارامترها به صورت تابعی از مختصات طول و عرض تغییر می‌کند. در اینجا واضح است که x و z و J اطلاعات موجود مشاهدات داده را نشان می‌دهند و تنها B_0 نشان دهنده پارامترهایی در مدل است که باید برآورد گردند.

مدل ناهمسانی فضایی را از طریق ایجاد امکان انحراف در رابطه محاسبه می‌کند، به گونه‌ای که گروههای مشاهدات مجاور یا همسایه مشخص شده با مختصات طول و عرض، مقادیر پارامتر مشابهی می‌گیرند.

هنگامی که مکان تغییر می‌یابد، رابطه رگرسیون تغییر می‌کند تا بر ارزش خطی محلی در میان گروه‌های مشاهداتی که تقریب نزدیکی برای یکدیگرند تطبیق یابد.

از دیگر روشهای مطرح شده برای برآورد انحراف در طول فضا که در زمینه اقتصادسنجی دارای کاربرد است، روش رگرسیون‌های وزنی جغرافیایی (GWR)[5] است که توسط افرادی به نام کارلتون[6]، براندسون[7] و فودرینگهام[8]، طراحی و معرفی شده در این مدل y نشان دهنده بردار $n * 1$ مشاهدات متغیر وابسته که از n نقطه در فضا به دست آمده، می‌باشد و x ماتریس $n * k$ متغیرهای توضیحی و بردار $n * 1$ خطاهای نرمال که دارای واریانس ثابت است. با فرض اینکه W_i نشانگر ماتریس قطری $n * n$ شامل وزنهایی بر مبنای فاصله برای مشاهده i باشد که منعکس کننده فاصله میان مشاهده i و سایر مشاهدات دیگر است می‌توان مدل GWR را به صورت زیر نوشت:

$$w_{iy} = w_i x \beta_i + w_i \epsilon_i$$

اندیس i در B_i نشان می‌دهد که بردار $k * 1$ پارامتر مربوط به مشاهدهٔ i است. مدل GWR، n مورد از چنین بردارهای مربوط به برآوردهای پارامترها را ایجاد می‌کند که هر یک برای یک

5) Geographically Weighted Regressions 6) Charlton 7) Brundson
8) Fotheringham

مشاهده است. این برآوردها با استفاده از رابطه زیر ایجاد می‌گردند:

$$B_i = (x'w_i^{۲}x)^{-۱}(x'w_i^{۲}y)$$

۴.۴ یک مثال تجربی

با توجه به آنچه در دو قسمت قبل توضیح داده شد، هدف ما در اینجا این است که با ارائه نتایج یک تخمین تجربی بر این نکته متمرکز شویم، هرگاه در تئوریهای اقتصاد، مدل نظری علاوه بر زمان، تحت تأثیر مکان وقوع داده‌ها نیز قرار داشته باشد، و یا به عبارت دیگر امکان فرض همگنی در بین واحدهای مختلف برای یک یا چند متغیر وجود نداشته باشد، نه تنها پردازش مدل به روش پانل نتایج واقعی‌تری را به دست می‌دهد بلکه تکنیک اقتصادسنجی فضایی را در نظر گرفتن وابستگی‌های فضایی، تکنیک مناسب‌تری جهت تخمین مدل خواهد بود.

همگرائی در درآمد سرانه بین مناطق مختلف مبتنی بر تئوری رشد نئوکلاسیک به صورت:

$$\frac{\delta g_k}{\delta k} = \frac{sf(k)k - sf(k)}{k^{۲}} = -s[f(k) - kf(k)]/k^{۲} < ۰$$

$$\therefore In(y_{it}/y_{it-۱}) = a - (۱ - e^{-B})Iny_{it-۱} + u_{it}$$

یک نمونه از الگوهایی است که تحت تأثیر زمان و مکان وقوع داده‌ها می‌باشد.

در اینجا k سرمایه سرانه، s نرخ میل به پس‌انداز، g_k نرخ رشد سرمایه سرانه، B ضریب همگرائی و y_{it} تولید سرانه در مکان i در زمان t می‌باشد.

با توجه به الگوی ارائه شده، فرضیهٔ همگرائی مطرح می‌کند که اقتصادها با سطوح پایین‌تر در درآمد سرانه به سمت رشد سریع‌تری در متغیرهای سرانه (درآمد سرانه و سرمایه سرانه) پیش می‌روند.

آزمون این فرضیه برای درآمد سرانه استانهای ایران طی دوره ۷۰ـ۱۳۸۰ به منظور مقایسه طی سه مرحله ۱ـ حداقل مربعات معمولی ۲ـ مدل پانل ۳ـ مدل پانل با استفاده از تکنیک اقتصادسنجی فضایی، صورت گرفت.

در مرحلهٔ اول میانگین نرخ رشد درآمد سرانه هر استان طی دورهٔ ۷۰ـ۱۳۸۰ محاسبه گردید و سپس به روش حداقل مربعات معمولی با استفاده از مدل بین مقطعی (cross section) ارتباط بین میانگین نرخ رشد درآمد سرانه و درآمد سرانه سال پایه (۱۳۷۰) برای استانهای ایران محاسبه گردید. نتایج به دست آمده به صورت زیر مشاهده می‌شود:[۹]

$$\ln(y_{i۸۰}/y_{i۷۰}) = \underset{(۰/۵۱)}{۰/۲۰۶} - (۱ - e^{-۰/۰۱۷})\underset{(-۰/۴۷)}{\ln(y_{i۷۰})} + \epsilon$$

۹) مراجعه شود به پیوست (۱)

$$R^۲ = \%۰/۹ \qquad D.W = ۱/۸۶$$

هر چند ضریب مثبت B تأیید کننده وجود همگرایی در بین استانهای ایران می‌باشد، اما این ضریب معنی دار نبوده و تابع از قدرت توضیح دهندگی کافی برخوردار نمی‌باشد.

در مرحلهٔ دوم با استفاده از مدل پانل برای هر متغیر یک ماتریس که تعداد سطرهای آن برابر تعداد استانها و تعداد ستونها برابر تعداد سالهای مورد نظر بود در نظر گرفته شده و به این ترتیب حجم مشاهدات از ۲۴ مشاهده در مرحلهٔ اول به ۲۴۰ مشاهده در مرحلهٔ دوم افزایش یافت. تخمین مدل به روش غیر خطی در محیط Tsp 4.3 نتایج زیر را به دست می‌داد:

$$\ln(y_{it}/y_{it-۱}) = \underset{۷/۲}{۳/۳۲} - (۱ - e^{\underset{-۷/۲}{-۰/۳۶}}) \ln(y_{it-۱}) + \epsilon$$

$$R^۲ = \%۱۹$$

همانطورکه مشاهده می‌شود $R^۲$ مدل نسبت به مرحلهٔ قبل افزایش یافته و معنی‌داری ضرایب به میزان قابل توجهی بالا رفته است و این نشان می‌دهد، روش تخمین روش مناسبی است و تنها متغیر مستقل از قدرت کافی برای توضیح تغییرات متغیر وابسته برخوردار نمی‌باشد.

در مرحلهٔ بعد با استفاده از تکنیک اقتصادسنجی فضایی، ماتریس مجاورت (w) به عنوان متغیر توضیحی وابستگی فضایی به صورت:

$$\ln(y_{it}/y_{it-۱}) = a - (۱ - e^{-B}) \ln y_{it-۱} + Pw \ln(y_{it}/y_{it-۱}) + \epsilon$$

وارد شد. نتایج به دست آمده به صورت زیر مشاهده می‌شود:

$$\ln(y_{it}/y_{it-۱}) = \underset{۷/۸}{۳/۰۰۲} - (۱ - e^{\underset{-۷/۸}{-۰/۳۱}}) \ln y_{it-۱} + \underset{۱۱/۱۷}{۰/۷۰۱w} \ln(y_{it}/y_{it-۱}) + \epsilon$$

$$R^۲ = \%۴۷$$

در این مدل نه تنها معنی‌داری ضرائب افزایش یافته بلکه $R^۲ = \%۴۷$ نیز نشان دهنده افزایش قدرت توضیحی متغیرهای مستقل در تغییرات متغیر وابسته می‌باشد از طرف دیگر ضریب مثبت متغیر وابستگی فضایی نمایانگر وجود اثرات مثبت ناشی از مجاورت در نرخ رشد در آمد سرانه هر استان بوده است.

به طورکلی میزان ضریب همگرایی این نتیجه را به دست می‌دهد که استانهای ایران با سرعت ۳۱ درصد در سال به سمت نقطهٔ تعادل پایدار در رشد اقتصادی همگرا بوده و نیمی از شکاف رشد اقتصادی موجود بین آنها طی ۲/۵ سال از میان می‌رود.[۱۰]

۱۰) مراجعه شود به پیوست (۱)

۵ نتیجه گیری

آنچه در این مقاله بدان پرداختیم معرفی مبنای نظری مدل پانل و کاربرد آن در تحلیل روابط بین داده‌هایی بود که در دو بعد بخش و زمان دارای ناهمگنی بوده و تخمین روابط خطی مابین آن‌ها جز در قالب ماتریس‌های پانل، نتایج را دچار تورش خواهد ساخت. هدف ما از طرح این مسأله پرداختن به داده‌هایی بود که به طور خاص نه تنها در طول زمان متغیر بوده بلکه تحت تأثیر مکان وقوع خود نیز می‌باشند و در اینجا بود که تکنیک اقتصادسنجی فضایی در اندازه‌گیری اثرات فضایی موضوعیت می‌یافت. در روش اقتصادسنجی معمول زمانی که داده‌ها در دو بعد زمان و مکان به صورت یک ماتریس در دسترس می‌باشند صرفنظر کردن از اثرات فضایی، نتایج را به دلیل وجود وابستگی و ناهمسانی واریانس فضایی بین داده‌ها، دچار خطا می‌سازد که این مسأله با وارد کردن متغیر فضایی با توجه به مبنای نظری روش اقتصادسنجی فضایی برطرف می‌گردد.

این موضوع در آزمون تجربی همگرایی بین درآمد سرانه استان‌های ایران طی دورهٔ ۷۰ـ۱۳۸۰ به خوبی ثابت گردید و نتایج نشان داد، کاربرد مدل پانل با استفاده از تکنیک اقتصادسنجی فضایی، بهترین گزینه در تخمین ضرائب توابعی است که ارتباط بین داده‌ها دو بعدی از نظر زمان و مکان می‌باشد.

مراجع

[۱] رائو و میلر، اقتصادسنجی کاربردی، ترجمهٔ حمید ابریشمی، مؤسسهٔ تحقیقات پولی و بانکی، تهران، ۱۳۷۰.

[۲] عسگری، علی و نعمت‌اله اکبری، روش‌شناسی اقتصادسنجی فضایی، تئوری و کاربرد، مجلهٔ پژوهشی دانشگاه اصفهان، جلد دوازدهم، شمارهٔ ۱ و ۲، ۱۳۸۰، صص ۹۳ـ۱۲۲.

[۳] گجراتی، دامودار، مبانی اقتصادسنجی، ترجمهٔ حمید ابریشمی، مؤسسهٔ انتشارات و چاپ دانشگاه تهران، تهران، ۱۳۷۷.

[4] Anselin , Luc , " Spatial Econometrics " , Bruton Center School of Social Sciences , University of Texas at Dallas , Richardson , 1999

[5] Baltaji , Badi. H , Econometric Analysis of Panel Data , John Wiley & sons Ltd , 2001

[6] Bond , Stephen.R, "Dynamic Panel Data Models: A Guide to Micro Data Methods and Practice" , Nuffield College , Oxford and Institute for Fiscal Studies , 2002

[7] Greene , William. H , " Econometric Analysis" , Prentice Hall , Fifth Edition , 2003

[8] Levine , Ned , "Spatial Statistics and GIS , Software Tools to Quantify Spatial Patterns" , Computer Report , 1996

[9] Rey , Sergio. J and Brett D. Montouri , "Us Regional Income Convergence : A Spatial Econometric Perspective , Department of Geography San Diego State University , San Diego , CA , USA , Regional Studies , vol 33 –2 , pp : 143-156 , 1997

پیوست (۱)

```
Dependent Variable: Y
Method: Least Squares
Date: 05/10/04   Time: 19:14
Sample: 1 24
Included observations: 24
Variable  Coefficient    Std. Error   t-Statistic       Prob.
================================================================
 C       0.206257       0.404214     0.510267        0.6149
 X      -0.017040       0.036280    -0.469672        0.6432
================================================================
R-squared           0.009927  Mean dependent var     0.016429
Adjusted R-squared -0.035076  S.D. dependent var     0.028592
S.E. of regression  0.029089  Akaike info criterion -4.157281
Sum squared resid   0.018615  Schwarz criterion     -4.059110
Log likelihood     51.88737   F-statistic            0.220592
Durbin-Watson stat  1.864353  Prob(F-statistic)      0.643211
================================================================

Current sample:  1 to 240

 PANEL DATA ESTIMATION
 =====================

Balanced data:  NI=   24, T=   10, NOB=   240

TOTAL (plain OLS) Estimates:

Dependent variable: Y

Mean of dependent variable =.016429
Std. error of regression = .131270
Std. dev. of dependent var. = .143271      R-squared = .164017
Sum of squared residuals=4.10118  Adjusted R-squared = .160504
Variance of residuals = .017232

 Estimated    Standard
Variable  Coefficient     Error       t-statistic
X          -.262174      .037805       -6.93481
C          2.91224       .418877        6.95249
Standard Errors are heteroskedastic-consistent (HCTYPE=1).

BETWEEN (OLS on means) Estimates:


Dependent variable: Y

Mean of dependent variable = .016429
Std. error of regression = .022971
Std. dev. of dependent var. = .028592      R-squared = .382588
Sum of squared residuals=.011609  Adjusted R-squared = .354524
Variance of residuals = .527662E-03
```

```
 Estimated    Standard
Variable  Coefficient     Error       t-statistic
 X         -.101579      .027512       -3.69224
 C         1.13841       .303912        3.74586


WITHIN (fixed effects) Estimates:

Dependent variable: Y

Sum of squared residuals = 3.54057        R-squared = .278291
Variance of residuals = .016468  Adjusted R-squared = .197729
Std. error of regression = .128327

 Estimated    Standard
Variable  Coefficient     Error       t-statistic
 X         -.498626      .062129       -8.02563
Standard Errors are heteroskedastic-consistent (HCTYPE=1).

F test of A,B=Ai,B:  F(23,215) = 1.4801,  P-value = [.0791]
Critical F value for diffuse prior (Leamer, p.114) =   6.4579

Variance Components (random effects) Estimates:

VWITH (variance of Uit)   =  0.16468E-01
VBET  (variance of Ai)    =  0.76408E-03
(computed from small sample formula)
THETA (0=WITHIN, 1=TOTAL) =  0.68307

Dependent variable: Y

Sum of squared residuals = 3.99415        R-squared = .185835
Variance of residuals = .018577  Adjusted R-squared = .094951
Std. error of regression = .136299

 Estimated    Standard
Variable  Coefficient     Error       t-statistic
 X         -.299537      .041642       -7.19321
 C         3.32492       .460056        7.22722

Hausman test of H0:RE vs. FE:CHISQ(1)=18.644,  P-value=[.0000]

************************************************************

END OF OUTPUT.

 MEMORY USAGE:    ITEM:    DATA ARRAY  TOTAL MEMORY
 UNITS:  (4-BYTE WORDS) (MEGABYTES)
 MEMORY ALLOCATED         :    500000       4.0
 MEMORY ACTUALLY REQUIRED :      7247       2.1
 CURRENT VARIABLE STORAGE :      51
```

```
Current sample:  1 to 240

 PANEL DATA ESTIMATION
 =====================

Balanced data:  NI=   24, T=   10, NOB=   240

TOTAL (plain OLS) Estimates:

Dependent variable: Y

Mean of dependent variable = .016429
Std. error of regression = .107360
Std. dev. of dependent var. = .143271     R-squared = .443169
Sum of squared residuals=2.73171 Adjusted R-squared = .438470
Variance of residuals = .011526

 Estimated    Standard
Variable  Coefficient     Error       t-statistic
 X         -.235172      .031862       -7.38107
 G         .702933       .075243        9.34215
 C         2.60531       .353358        7.37301
Standard Errors are heteroskedastic-consistent (HCTYPE=1).

BETWEEN (OLS on means) Estimates:

Dependent variable: Y

Mean of dependent variable = .016429
Std. error of regression = .020700
Std. dev. of dependent var. = .028592     R-squared = .521400
Sum of squared residuals=.899863E-02 Adjusted R-squared=.475819
Variance of residuals = .428506E-03

 Estimated    Standard
Variable  Coefficient     Error       t-statistic
 X         -.081598      .026081       -3.12866
 G         -.859328      .348195       -2.46795
 C         .928331       .286796        3.23690

WITHIN (fixed effects) Estimates:

Dependent variable: Y

Sum of squared residuals = 2.31672        R-squared = .527762
Variance of residuals = .010826   Adjusted R-squared = .472594
Std. error of regression = .104047

 Estimated    Standard
Variable  Coefficient     Error       t-statistic
 X         -.411161      .049500       -8.30634
 G         .676947       .074284        9.11299
Standard Errors are heteroskedastic-consistent (HCTYPE=1).
```

```
F test of A,B=Ai,B:  F(23,214) = 1.6667,  P-value = [.0329]
Critical F value for diffuse prior (Leamer, p.114) =   6.4279

Variance Components (random effects) Estimates:

VWITH (variance of Uit)   =  0.10826E-01
VBET  (variance of Ai)    =  0.70043E-03
(computed from small sample formula)
THETA (0=WITHIN, 1=TOTAL) =  0.60716

Dependent variable: Y

Sum of squared residuals = 2.61332        R-squared = .467302
Variance of residuals = .012212  Adjusted R-squared = .405070
Std. error of regression = .110507

 Estimated    Standard
Variable  Coefficient     Error       t-statistic
 X         -.271065      .034930       -7.76019
 G         .701419       .062810       11.1672
 C         3.00178       .385995       7.77674

Hausman test of H0:RE vs. FE: CHISQ(2)=16.861,P-value=[.0002]

************************************************************

END OF OUTPUT.

 MEMORY USAGE:    ITEM:    DATA ARRAY  TOTAL MEMORY
 UNITS:  (4-BYTE WORDS) (MEGABYTES)
 MEMORY ALLOCATED         :    500000       4.0
 MEMORY ACTUALLY REQUIRED :      7948       2.1
 CURRENT VARIABLE STORAGE :      5266
```

هم انباشتگی کسری در سری‌های زمانی و بررسی نرخ تورّم در ایران

سید محمود میرجلیلی[۱]، قاسم تارمست[۲]

[۱] **دانشجوی کارشناسی ارشد دانشگاه شهید چمران اهواز***
[۲] **هیئت علمی دانشگاه شهید چمران اهواز**

چکیده: اغلب روشهای تجزیه و تحلیل سری‌های زمانی بر روی سری‌های زمانی ایستا انجام می‌شود در حالی که در بسیاری از زمینه‌ها مانند اقتصاد سریهائی به صورت ناایستا وجود دارد. روشهای کلاسیک تجزیه و تحلیل سری‌های زمانی ابتدا سری‌های ناایستا را به ایستا تبدیل می‌کنند که این کار با تفاضل‌گیری انجام می‌شود. روش تفاضل‌گیری در سری‌های بلند مدّت باعث حذف روند بلند مدت سری می‌شود و در سری‌های کوتاه مدت باعث افزایش خطا می‌شود. برای حل این مشکل روش هم‌انباشتگی (cointegration) برای اولین بار توسط انگل ـ گرنجر (۱۹۸۷) معرفی شد بر اساس این روش هرگاه بتوان ترکیب خطی از چند سری ناایستا با درجه ناایستائی مشابه (انباشتگی Integration)، I(d) را طوری پیدا کرد که آن ترکیب خطی ایستا باشد (I(0)) به آن متغییرها هم‌انباشته می‌گویند. با این روش علی‌رغم ناایستائی سری‌های تحت مطالعه، نتایج قابل تفسیر هستند و دیگر مسأله رگرسیون کاذب پیش نمی‌آید.

در روشهای سنّتی سری‌های زمانی اندازه درجه انباشتگی یعنی d را یک عدد صحیح (اغلب عدد ۱ یا ۲) در نظر می‌گرفتند. دراین تحقیق عدد d را یک عدد حقیقی در نظر می‌گیریم که، اگر d یک عدد حقیقی بین صفر تا دو باشد به سری انباشته کسری می‌گوئیم و برای اینکه تبدیل به سری ایستا شود باید تفاضل‌گیری کسری به اندازه d بر روی سری انجام شود. در رابطه هم‌انباشتگی نیز وقتی سری‌های ناایستا با هم ترکیب می‌شوند باید ترکیب خطی آنها یک سری ایستا یعنی I(0) باشد در حالی که در روش هم‌انباشتگی کسری ترکیب خطی آنها می‌تواند انباشته از درجه $\frac{1}{2}-$ تا $\frac{1}{2}$ باشد.

در این تحقیق بعد از معرفی روش هم‌انباشتگی کسری، رابطه بین نرخ تورم، نرخ ارز، حجم پول و تولید ناخالص داخلی از سال ۱۳۵۰ تا سال ۱۳۸۱ در ایران را بررسی می‌کنیم.

واژه‌های کلیدی: انباشتگی کسری، تفاضل‌گیری کسری، هم‌انباشتگی کسری

*) smirjalily@yahoo.ca

۱ مقدمه

سریهای زمانی به داده‌هائی گفته می‌شود که در طول زمان و با فواصل مساوی اندازه‌گیری می‌شوند بطور کلی سری‌های زمانی را می‌توانیم به دو کلاس تقسیم کنیم کلاس اول شامل سری‌های است که طول زمان دارای میانگین ثابتی هستند و یا به عبارتی دیگر حرکت میانگین آنها به زمان بستگی ندارد، تابع خودهمبستگی آنها به سرعت کاهش می‌یابد، واریانس آنها محدود و ثابت است و در طول زمان تغییر نمی‌کند و چگالی طیفی آنها نیز برای تمام فرکانس‌ها وجود دارد، این‌گونه سریها را ایستای کواریانس می‌گویند.

کلاس دوم شامل سری‌های زمانی ناایستا است، سری‌های ناایستا در طول زمان میانگین ثابتی ندارند تابع خودهمبستگی آنها به کندی نزول می‌کند، دارای واریانس نامحدود هستند و تابع چگالی طیفی آنها برای تمام فرکانس‌ها وجود ندارد. اغلب روشهایی که برای تجزیه و تحلیل سری‌های زمانی وجود دارد بر روی سری‌های زمانی ایستا انجام می‌شود و در صورت مواجه شدن با سری‌های ناایستا، با استفاده از روش تفاضل‌گیری آنها را تبدیل به سری‌های ایستا می‌کنند. اگر سری ناایستا از درجه یک باشد با یکبار تفاضل‌گیری تبدیل به سری ایستا می‌شود و اگر ناایستا از درجه دو باشد با دوبار تفاضل‌گیری و همین‌طور الی آخر. بنابراین اگر $\{x_t; t = ۱, ..., n\}$، یک سری زمانی باشد و داشته باشیم

$$(۱ - B)^d x_t = \epsilon_t \qquad \epsilon_t \sim WN(۰, \sigma^۲) \tag{۱}$$

یعنی سری x_t با d بار تفاضل‌گیری به یک سری اغتشاش خالص ϵ_t که دارای توزیع نرمال با میانگین صفر و واریانس ثابت $\sigma^۲$ است تبدیل می‌شود. اگر x_t بصورت معادله (۱) باشد آنرا انباشته از درجه d می‌گوئیم و آنرا با $x_t \sim I(d)$ نشان می‌دهیم. در اغلب کارهای کاربردی و روشهای سنتی سری‌های زمانی d را صفر، یک یا حداکثر دو در نظر می‌گیرند. اما در دو دهه اخیر یک کلاس دیگر از سری‌های زمانی توسط هاسکینگ[۱] (۱۹۸۱)، انگل[۲] (۱۹۸۳) و فاکس تاکو[۳] (۱۹۸۶) و غیره معرفی شده است. این کلاس از سری‌های زمانی بین سری‌های ایستا و ناایستا قرار گرفته‌است که d یعنی پارامتر تفاضل‌گیری می‌تواند هر عدد طبیعی در بازه $(-\frac{۱}{۲}, ۲)$ را بپذیرد. این دسته از سری‌های زمانی را سری‌های انباشته کسری می نامند. برای اینگونه سری‌ها تفاضل‌گیری در معادله (۱) را می‌توانیم بصورت زیر نشان دهیم:

$$(۱ - B)^d = \sum_{j=۰}^{\infty} \frac{\Gamma(j - d)}{\Gamma(-d)\gamma\ (j + ۱)} L^j \tag{۲}$$

بطوریکه

$$\Gamma(a) = \begin{cases} \int_۰^\infty x^{(a-۱)} e^{-e} dx & a > ۰ \\ \frac{(-۱)^n}{n!} & a < ۰ \\ \Gamma(a + ۱) = a\Gamma(a) & a = ۱, \ldots, n \end{cases}$$

1) Husking 2) Engle 3) Fax and Taqqu

سریهایی که پارامتر تفاضل‌گیری آنها در بازه $-\frac{1}{2} < d < \frac{1}{2}$ قرارگیرد را ایستا و سریهایی که $|d| \geq \frac{1}{2}$ باشد را ناایستا می‌نامند.

روشهای چند متغیره سری‌های زمانی که ارتباط بین متغیرها را بررسی می‌کنند اغلب بر اساس فرض ایستایی سریها انجام می‌شوند و برای سری‌های ناایستا نیز ابتدا آنها را با روش تفاضل‌گیری تبدیل به سری‌های ایستا می‌کنند، روش تفاضل‌گیری در سری‌های بلند مدت باعث حذف روند سری و در سری‌های کوتاه مدت باعث افزایش خطا می‌شود. برای رفع این مشکل و بدست آوردن رابطه بلند مدت بین متغیرها، روش هم‌انباشتگی برای اولین بار توسط انگل ـ گرنجر[۴] (۱۹۸۷) معرفی شد. بر اساس این روش اگر بتوان ترکیب خطی از متغیرهای ناایستا را طوری پیدا کرد که آن ترکیب خطی ایستا باشد به آن متغیرها هم‌انباشته می‌گویند. فرض کنید X_t یک بردار $p \times 1$ متغیر از سری‌های زمانی باشد و داشته باشیم

$$\beta' X_t = \epsilon_t \qquad (۳)$$

که β یک بردار $p \times 1$ از پارامترها باشد. در تعریف ابتدایی انگل و گرنجر (۱۹۸۷) سری‌های X_t همگی $I(1)$ بودند و ϵ_t می‌بایست $I(0)$ باشد تا رابطه هم‌انباشتگی وجود داشته باشد. آنها در مقاله خود اشاره‌ای به هم‌انباشتگی کسری نیز کردند. در رابطه هم‌انباشتگی کسری سری‌های X_t می‌توانند انباشته کسری، $I(d)$ باشند و باقیمانده‌های ϵ_t، باید انباشته از درجه $I(d-b)$، که $b > 0$ است باشند.

حال مساله اصلی که پیش می‌آید تشخیص مقادیر d و b است. انگل و گرنجر (۱۹۸۷) برای اینکار از آزمونهای ریشه واحد دیکی و فولر[۵] و دیکی فولر تعمیم یافته[۶] (۱۹۸۱) استفاده می‌کردند. هدف اصلی ما نیز در این مقاله بررسی برآورد پارامتر تفاضل‌گیری یعنی d است. روشهای اولیه برای برآورد d، ابتدا توسط جواک و پوتر هارک[۷] (۱۹۸۶) و کانچ[۸] (۱۹۸۷) معرفی شدند و سپس توسط ساول[۹] (۱۹۹۲) و رابینسون[۱۰] (a،b ۱۹۹۵) و غیره گسترش یافتند. در زمینه هم‌انباشتگی کسری هم بعد از اشاره انگل و گرنجر (۱۹۶۷) مطالعه و تحقیق در این زمینه توسط چان وترین[۱۱] (۱۹۹۵)، رابینسون (۱۹۹۴) رابینسون و ماریناسی[۱۲] (۱۹۹۸) و ماریناسی و رابینسون (۲۰۰۱) و رابینسون و یاجیما[۱۳] (۲۰۰۲) صورت گرفته است.

مدل‌بندی فرآیندهای کسری امکان بیشتری را برای وجود رابطه هم‌انباشتگی بین فرآیندها بوجود می‌آورند و تحقیقات جدید و خاصی را در این زمینه طلب می‌کنند. روش‌های مختلفی برای هم‌انباشتگی که در آنها درجه انباشتگی عدد صحیح فرض شده گسترش یافته‌اند، ولی این روشها، برای درجه‌های مختلف انباشتگی معتبر نیستند و باید برای فرآیندهای کسری تعمیم داده

4) Engle and Granger 5) Dicky-Fuller 6) Augmented Dicky-Fuller 7) Geweke and Poter Hudack 8) Kunsch 9) Sowell 10) Robinson ,P.M 11) Chan and Terrin 12) Robinson and Murrinaci 13) Yajimma

شوند. اما این که ما می‌توانیم بی نهایت مقدار برای d در نظر بگیریم کار را سخت می‌کند و برای حل این مشکل باید توسط آزمونهایی مقادیر d را از قبل تعیین کنیم. یک مساله پیچیده‌تر و سخت‌تر این است که هر یک از متغیرها دارای درجه انباشتگی متفاوت باشند و بخواهیم رابطه هم‌انباشتگی آنها را بررسی کنیم. که در این زمینه تحقیقات کمی صورت گرفته است.

مقاله حاضر روشهایی را برای استنباط در مورد فرآیندهای کسری و نیز امکان رابطه هم‌انباشتگی کسری را بررسی می‌کند و در مورد برآورد پارامترهای هم‌انباشتگی بحث می‌کند. در ادامه این مقاله در بخش دوم تعاریف انباشتگی و هم انباشتگی کسری را مطرح می کنیم. در قسمت سوم برآورد پارامتر تفاضل‌گیری و آزمون فرض در مورد آن را با روشهایی که در حوزه فرکانس و برای فرکانس‌های نزدیک صفر، توسط رابینسون (۱۹۹۵)، لوبیتو[14] (۱۹۹۶) و لوبیتو و رابینسون (۱۹۹۸) معرفی شده‌اند را بررسی می‌کنیم . در قسمت چهارم برآورد بردار رگرسیونی پارامترهای هم انباشتگی که در حوزه فرکانس و در یک باند کوچک حول فرکانس صفر توسط رابینسون و مارینا‌سی (۱۹۹۹) معرفی شده‌اند را بررسی می‌کنیم و بالاخره در قسمت پنجم، رابطه هم‌انباشتگی کسری بین نرخ تورم، نرخ ارز، حجم پول و تولید ناخالص داخلی در سالهای ۱۳۸۱ ـ ۱۳۵۰ در ایران را بررسی می‌کنیم.

۲ تعاریف انباشتگی و هم انباشتگی کسری

تعریفهای متفاوتی برای فرآیندهای انباشته کسری $I(d)$ وجود دارد یکی از آنها این طور بیان می کند که فرآیند $\{X_t; \quad t \in Z\}$ را $I(d)$ می‌گوئیم اگر یک فرآیند $\{\epsilon_t \quad t \in Z\}$ که $I(۰)$ است وجود داشته باشد بطوری‌که:

$$X_t = \mu + \Delta^{-d}\epsilon_t \mathbb{1}(t > ۰) \qquad (۴)$$

که $\mathbb{1}(۰)$ تابع نشانگر، $\Delta = (۱ - B)$، B عملگر وقفه و μ میانگین X_t است. فرض کنید X_t یک بردار $p \times ۱$ متغیره باشد که هر کدام از آنها انباشته از درجه $d_i = ۱, ۲, ..., p$ باشند، برای وجود رابطه هم‌انباشتگی حداقل باید دو سری زمانی داشته باشیم که درجهٔ انباشتگی آنها مشابه باشد. حال اگر بیش از دو سری زمانی داشته باشیم درجهٔ انباشتگی آنها می‌تواند متفاوت باشد. امّا درجهٔ انباشتگی حداقل دو سری که دارای بزرگترین درجهٔ انباشتگی هستند باید با هم مشابه باشد. رابینسون و یاجیما (۲۰۰۲) بیان می‌کنند برای بررسی بهتر روابط هم‌انباشتگی بین سریها می‌توانیم آنها را با توجه به درجه انباشتگی‌شان در s مجموعه بصورت زیر افراز کنیم.

$$d_۱ = ... = d_{i_۱} > d_{i_۱+۱} = ... = d_{i_۲} > ... > d_{i_{s-۱}+۱} = ... = d_{i_s} \qquad (۵)$$

14) Lobato

که در آن $i_0 = 0$ و $1 \leq i_1 < i_2 < ... < i_s = p$ است. فرض کنید

$$\beta = (\beta_1, \beta_2, ..., \beta_p)' \qquad (۶)$$

یک بردار p بعدی از ضرایب باشد. بردار β را نیز مانند (۵) افراز می‌کنیم.

$$\beta = (\beta^{(1)'}, ..., \beta^{(s)'})' \qquad (۷)$$

که در آن $p_l = i_l - i_{l-1}$ و $\beta^{(l)} = (\beta_{i_{l-1}+1}, ..., \beta_{i_l}), l = 1, 2, ..., s$ تعداد سری‌های زمانی در مجموعهٔ iام است و به همین صورت بردار X_t را افراز می‌کنیم.

$$X_t = (X_t^{(1)'}, ..., X_t^{(s)'})' \quad , \quad X_t^{(l)} = (X_{i_{l-1}+1,t}, ..., X_{i_l,t}) \qquad (۸)$$

حال با توجه به افراز بالا تعاریف هم‌انباشتگی کسری را که توسط نویسندگان مختلف بیان شده است در اینجا می‌آوریم.

تعریف انگل ـ گرنجر (۱۹۸۷): اگر $d_1 = d_2 = ... = d_p = d = 1$ باشد و یک بردار β وجود داشته باشد بطوریکه βX_t $I(d_u)$ باشد و $d_u = d - b = 0$ باشد آنگاه این سری هم‌انباشته از درجهٔ $d - b$ یا $CI(d, b)$ هستند.

تعریف جوهانسون (۱۹۹۶): مؤلفه‌های بردار X_t را هم‌انباشته گوییم اگر یک بردار β وجود داشته باشد بطوریکه βX_t، $I(d_u)$ باشد و $d_u < d_1$ باشد. در این تعریف d_u که در درجهٔ انباشتگی βX_t است باید از d_1 که بزرگترین درجهٔ انباشتگی بین سریها است کوچکتر باشد.

تعریف فلورز و زافارز[15] (۱۹۹۶): مؤلفه‌های بردار X_t را هم‌انباشته گوییم اگر یک بردار β وجود داشته باشد بطوریکه $\beta(1) \neq 0$ و $\beta(X_t)$، $I(d_u)$ باشد و $d_u < d_1$ باشد. در این تعریف ضریب متغیری که دارای بزرگترین درجهٔ انباشتگی است نباید صفر باشد.

تعریف رابینسون و ماریناسی (۱۹۹۸): مؤلفه‌های بردار X_t را هم‌انباشته گوییم اگر یک بردار $\beta \neq 0$ وجود داشته باشد بطوریکه $\beta(X_t)$، $I(d_u)$ باشد و $d_u < d_p$ باشد. در این تعریف d_u درجهٔ انباشتگی βX_t باید از d_p که کوچکترین درجهٔ انباشتگی بین سریها است کوچکتر باشد.

تعریف رابینسون و یاجیما (۲۰۰۲): مؤلفه‌های بردار X_t را هم‌انباشته گوییم اگر یک بردار غیرصفر $\beta(l)$ وجود داشته باشد بطوریکه $\beta'(l)X_t^{(l)}$، $I(d_u)$ باشد و $d_u < d_{i_l}$ باشد. در این تعریف بردار هم‌انباشتگی بصورت $\beta = (o', o', ..., o', \beta(l)', o', ..., o')'$ است. در ادامه سعی می‌کنیم با استفاده از یک مثال تعاریف بالا را نشان دهیم.

مثال: بردار $p = 5$ متغیره x_t را بصورت زیر در نظر بگیرید

$$x_t = (u_t + \varepsilon_{1t}, au_t + \varepsilon_{2t}, v_t + \varepsilon_{3t}, bv_t + \varepsilon_{4t}, \varepsilon_{5t}) \qquad (۹)$$

15) Flores and Szafarz

که در آن $0 < e < d, v_t \sim I(e), u_t \sim I(d), a, b \neq 0$ طبق تعریف فلورز و سافارز (۱۹۹۶) $r = 1$ تعداد رابطه‌های هم‌انباشتگی است و بردار هم‌انباشتگی بصورت زیر است

$$\beta = (0, 0, -1, 0, 0)'$$

چون که درجه انباشتگی سریها متفاوت است تعریف انگل ـ گرنجر (۱۹۸۷) کاربردی ندارد.

طبق تعریف جوهانسن (۱۹۹۶) $r = 4$ تعداد روابط هم‌انباشتگی است که بصورت زیر نشان می‌دهیم

$$\beta = (a, -1, , 0, 0)', \beta = (0, 0, b, -1, 0)', \beta = (0, 0, 0, 1, 0)', \beta = (0, 0, 0, 0, 1)'$$

طبق تعریف فلورز و زافارز (۱۹۹۶) تعداد روابط هم‌انباشتگی $r = 1$ است که بصورت زیر نمایش می‌دهیم

$$\beta = (0, -1, 0, 0, 0)'$$

طبق تعریف رابینسون و مارینا‌سی (۱۹۹۸) تعداد روابط هم‌انباشتگی $r = 0$ است. چون ترکیب خطی وجود ندارد درجه انباشتگی آن کمتر از کوچکترین درجه انباشتگی یعنی صفر باشد.

طبق تعریف رابینسون و یاجیما (۲۰۰۴) تعداد روابط هم‌انباشتگی $r = 2$ است که بصورت زیر نشان می‌دهیم

$$\beta = (a, -1, 0, 0, 0)' \qquad \beta = (0, 0, b, -1, 0)'$$

در تعریف رابطه هم‌انباشتگی باید به نکات زیر توجه کنیم:

الف) اگر یک بردار هم انباشتگی را بصورت نرمال تبدیل[16] کنیم آنگاه آن بردار منحصر به‌فرد می‌شود

ب) خطاهای حاصل از روابط هم انباشتگی می‌توانند انباشته از درجه‌های متفاوت باشند.

ج) تعاریف بالا فقط شامل ترکیبهای خطی می‌شوند و ترکیبهای غیر خطی را پوشش نمی‌دهند

۳ استنباط در مورد درجه انباشتگی

استنباط و برآورد مرتبه‌های انباشتگی کسری یک مساله کلیدی در هم‌انباشتگی کسری است. روشهای برآورد d همگی در حوزه فرکانس انجام می‌شوند. روشهائی که در حوزه زمان وجود دارند بیشتر برای آزمون فرض در مورد d بکار می‌روند در حوزه فرکانس نیز اغلب روشهائی بکار گرفته شده که بصورت نیمه پارامتری هستند و فرکانسهای نزدیک صفر در دوره نگار (برآورد تابع

16) Normalization

چگالی طیف) را برای برآورد d بکار می‌برند روشهای برآورد نیمه پارامتری روشهای نیرومندی هستند که بدون اینکه نیازی به تعیین مدل داشته باشند می‌توانند سازگاری برآوردها را تامین کنند. اما در روشهای پارامتری مشخص نبودن پارامترهای مدل می‌تواند باعث ناسازگاری در برآوردها شود. تعریف انباشتگی که در معادله (۴) بیان کردیم فقط برای یک متغیر بود اما اگر یک بردار از متغیرها داشته باشیم باید به مدلی فکر کنیم که بطور توام همه فرایندهای انباشته کسری را شامل شود. برآورد درجه انباشتگی برای سریهائی که دارای فرض ایستائی درکواریانس باشند و چگالی طیفی آنها موجود باشد بطور اساسی گسترش یافته‌اند. ما نیز در اینجا همین موضوع را ادامه می‌دهیم چون که خاصیت مجانبی سری‌های انباشته کسری با $|d| \leq \frac{۱}{۲}$ شبیه به سری‌های ایستا است. برای سری‌های ناایستا $|d| > \frac{۱}{۲}$ نیز می‌توانیم با یک مرتبه تفاضل‌گیری خاصیت مجانبی سری‌های ایستای کواریانس را بوجود آوریم و پس از آنکه درجه انباشتگی را برآورد کردیم به آن مقدار یک اضافه کنیم [رابینسون (b و ۱۹۹۵a) لوبیتو (۱۹۹۶)]. توجه کنید که در فرایندهای AR استنباط روی داده‌های تفاضل‌گیری شده باعث عدم کارائی می‌شود ولی در فرآیندهای کسری وقتی فرض ما روی داده‌های تفاضل‌گیری شده باشد استنباط در مورد فرایندها هنوز دارای کارائی لازم است (رابینسون (b ۱۹۹۴).

برای برآورد درجه انباشتگی خطاهای e_t، d_e، می‌توانیم روشهایمان را روی یک فرایند ξ_t که با $\hat{e}_t = y_t - \hat{\beta}x_t$ برآورد می‌شود انجام دهیم $\hat{\beta}$ هم با یکی از روشهائی که در بخش (۴) توضیح می‌دهیم برآورد می‌شود

یک بردار $p \times ۱$ متغیره x_t که ایستا درکواریانس با تابع چگالی طیفی $f_x(\lambda)$ را در نظر بگیرید. (k,l)امین مولفه $f_x(\lambda)$ بصورت زیر است

$$f_{kl}(\lambda) \sim g_{kl} e^{i(\frac{\pi}{۲})(\delta_k - \delta_l)} \lambda^{-\delta_k - \delta_l} \qquad \lambda \longrightarrow ۰^+ \qquad (۱۰)$$

که در آن $k, l =, ۱, ۲, ..., p$. $\sim$ نشان می‌دهد که نسبت بخش حقیقی و موهومی سمت راست و چپ معادله به سمت یک میل می‌کند، برای $k = ۱, ۲, ..., p$ $\frac{-۱}{۲} < d < \frac{۱}{۲}$ ماتریس $G = [g_{kl}]$ برای حالتی که هم‌انباشتگی بین سریها وجود نداشته باشد معین مثبت و در غیر این صورت نیمه معین مثبت است و برای $k = ۱, ۲, ..., p$ $g_{kk} > ۰$ است.

دو راهکار کلی برای برآورد $d = (d_۱, d_۲, ..., d_p)$ وجود دارد اولی براساس لگاریتم دوره نگارکه توسط جواک و پوتر هاک (۱۹۸۳) معرفی شده و خاصیت‌های مجانبی آن توسط رابینسون (a ۱۹۹۵) و هرویچ و دیگران[۱۷] بدست آمده است.

فرض کنید $\{x_t, t = ۱, ۲, ..., n\}$ یک سری زمانی باشد تبدیل فوریه گسسته بصورت زیر انجام می‌شود

$$w_x(\lambda) = \frac{۱}{(۲\pi n)^{\frac{۱}{۲}}} \sum_{t=۱}^{n} x_t e^{it\lambda} \qquad (۱۱)$$

17) Hurvich et al.

همچنین اگر یک سری $\{y_t, t =, ۱, ۲, ..., n\}$ داشته باشیم دوره نگار متقابل بصورت زیر می‌شود

$$I_{xy} = w_x(\lambda)w_y(\lambda) \qquad (۱۲)$$

حال فرض کنید $I_{kk}(\lambda)$، kامین مولفه قطری $I_x(\lambda)$ باشد. برای اعداد صحیح s، $y_{kj} = \log(I_{kk}(\lambda_j))$ که $s < [\frac{n}{۲}], j = ۱, ۲, ..., s \quad k = ۱, ۲, ..., q$ علامت $[.]$ جزء صحیح و s پارامتر پنجره برای هموارسازی دوره نگار است را در نظر بگیرید، در این صورت d_k بصورت زیر برآورد می‌شود

$$\tilde{d}_k = \frac{\sum_{j=۱}^{s} \nu_j Y_{kj}}{۲ \sum_{j=۱}^{s} \nu_j Y_{kj}} \qquad \nu_j = \log\, j - \frac{۱}{s}\sum_{j=۱}^{s} \log j \quad k = ۱, ..., q \qquad (۱۳)$$

توزیع مجانبی d_k توسط رابینسون (a ۱۹۹۵) بررسی شده که بصورت زیر است

$$\tilde{d}_k \sim N(d_k, \frac{\pi^۲}{۲۴s})$$

برای آزمون فرض $H_۰ : \Pi d = \rho$ می‌توانیم از آماره والد بصورت زیر استفاده کنیم

$$W = ۴s(\Pi d - \rho)'(\Pi\Omega\Pi')^{-۱}(\Pi d - \rho) \qquad (۱۴)$$

فرض صفر در صورتی رد می‌شود که این آماره بزرگ‌تر از $\chi^۲_u$ (چندک uام توزیع $\chi^۲$ باشد). در این آماره $d = (d_۱, d_۲, ..., d_p)$ و Ω یک برآورد سازگار از واریانس حدی ماتریس $۲s(\tilde{d}_k - d)$ است (رابینسون a ۱۹۹۵). توجه کنید که روشهای نیمه پارامتری[۱۸] برای برآورد درجه انباشتگی دارای نرخ سازگاری $n^{\frac{۱}{۲}}$ هستند. بنابراین یک استنباط قابل اعتبار در این زمینه نیاز به حجم نمونه بزرگ دارد.

برای برآورد d می‌توانیم قیدهائی را روی d_k اعمال کنیم که در این صورت می‌توان برآوردهای کاراتری از d را که در ادامه کار رابینسون (a ۱۹۹۵) که توسط ماریناسی و رابینسون (۲۰۰۱) معرفی شده‌اند بدست آورد. فرض کنید این قیود بصورت $d_۱ = d_۲ = ... = d_p$ باشند مقدار مشترک d_* با برآوردهای GLS بصورت زیر بدست می‌آید

$$\tilde{d}_* = -\frac{\sum_{j=۱}^{s} ۱'\Omega^{-۱}Y_j\nu_j}{۲۱'_p\tilde{\Omega}^{-۱}۱_p \sum_{j=۱}^{s} \nu_j^۲} \qquad (۱۵)$$

18) semiparametric

که در آن $y_j = (y_{1j}, ..., y_{pj})$است. توزیع مجانبی برآورد کننده $\tilde{d}_*$ نیز توسط مارینا‌سی و رابینسون (۲۰۰۱) بصورت زیر بدست آمده است

$$\tilde{d}_* \sim N(d_*, \frac{1'_p \tilde{\Omega}^{-1} 1_p}{4s})$$

بنابراین می‌توانیم آزمون والد را برای این مقدار مشترک d_*انجام دهیم.

کارائی برآوردهای $\tilde{d}$ و d_* خیلی کمتر از برآوردهای کلاس دیگری از این برآورد کننده‌ها یعنی برآوردهای نیمه پارامتری گوسی، یا برآورد وایتل[19] که توسط کانچ[20] (۱۹۸۷) معرفی شد و سپس توسط رابینسون (b ۱۹۹۵) و لوبیتو (۱۹۹۸) گسترش یافت. این برآوردها با استفاده از بهینه کردن تابع بزرگنمائی گوسی در فرکانسهای فوریه نزدیک به صفر بدست می‌آیند. برای برآوردهای تک متغیره d_* از تابع هدف زیر[21] که بصورت تک متغیره است استفاده می‌شود

$$R_k(d_k) = \log\left(\frac{1}{s}\sum_{j=1}^{s} I_{kk}(\lambda_j) j^{2d_k}\right) - \frac{2d_k}{s}\sum_{j=1}^{s}\log j \qquad (16)$$

و d_k بصورت $\bar{d}_k = arg\ min(R_k(d_k))$ برای $k = 1, 2, ...p$با مینیمم کردن $R_k(d_k)$برای $d \in (\frac{-1}{2}, \frac{1}{2})$ که ایستائی و وارون‌پذیری را بیان می‌کند برآورد می‌شود و هرکدام از $\bar{d}_k$ها دارای توزیع تقریبی $N(d_k, \frac{1}{4s})$ هستند، بنابراین این برآوردکننده خیلی کاراتر از $\tilde{d}_k$است [رابینسون (b ۱۹۹۵)]. یک برآورد کاراتر دیگر نیز توسط لوبیتو (۱۹۹۶) برای برآورد چند متغیره بصورت زیر معرفی شده است

$$R(\delta) = \log\left|\frac{1}{s}\sum_{j=1}^{s} A_j(d)\right| - \frac{2}{s}\sum_{k=1}^{p} d_k \sum_{j=1}^{s}\log j \qquad (17)$$

که درآن $A_j(d) = Re(\Lambda_j(d) I_{xx}(\lambda_j) \Lambda_j(d))$و $\Lambda_j(d) = diag(e^{\frac{i\pi d_1}{2}} j^{d_1}, ..., e^{\frac{i\pi d_p}{2}} j^{d_p})$ و برآورد $\hat{d}$ بصورت $\hat{d} = arg\ \ min(R(d))$با مینیمم کردن $R(d)$ در بازه فشرده $d \in (\frac{-1}{2}, \frac{1}{2})^p$ بدست می‌آید. $\hat{d}_k$ و $\tilde{d}_*$ برخلاف $\bar{d}_k$ و $\tilde{d}$ در یک فرم بسته تعریف نمی‌شوند. بهر حال یک روش دیگر برای برآورد d_k روش مرحله‌ای نیوتن است. در گام اول آن باید یک برآورد سازگار d با نرخ همگرائی $s^{\frac{1}{2}}$ بدست آوریم. در این روش هر چند در مراحل بعدی کارائی مرتبه اول بهبود پیدا نمی‌کند اما کارائی مراتب بالاتر بهبود پیدا می‌کند [رابینسون (۱۹۹۸)]. در این برآورد $\hat{d}$ در مرحله $v+1$ بصورت زیر بدست می‌آید

$$\hat{d}^{[\nu+1]} = \hat{d}^{[\nu]} - \left\{2\left(I_q + \hat{G}^{[\nu]} \circ \hat{G}^{[\nu]^{-1}}\right)\right\}^{-1} \frac{\partial R\left(\hat{d}^{[\nu]}\right)}{\partial R} \qquad (18)$$

19) Whittle Estimate 20) Kunsch 21) Objective function

که در آن $v \geq ۰$ و $\hat{d}^{[v]}$ برآورد مرحله قبلی آن، $\circ$ نشان دهنده ضرب هادامارد[۲۲]، $\hat{G}^{[v]} = \hat{G}^{[v]}(\hat{d}^{[v]})$ و $\hat{G}(d) = s^{-۱}\sum_{j=۱}^{s} A_j(d)$ است. در معادله (۱۷) ماتریس داخل براکت یک برآورد سازگار از احتمال حدی $\frac{\partial^۲ R(\hat{d}^{[\nu]})}{\partial d \partial d'}$ است. توجه کنید که $۲(I_p + \hat{G}^{[v]} \circ \hat{G}^{[v]^{-۱}})$ وقتی که $p = ۱$ باشد به ۴ کاهش پیدا می‌کند [مارینا‌سی و رابینسون (۲۰۰۱)].

یک انتخاب برای $\hat{d}^{[۰]}$، $\tilde{d}$ است، چون که مانند $\hat{d}$ دارای نرخ سازگاری $s^{\frac{۱}{۲}}$ است با این حال باز دارای کارائی کمتری است. خصوصیاتی که در بالا ذکر کردیم برای فرض عدم هم‌انباشتگی بود. اما در برآوردهای انفرادی که اول توضیح دادیم فرض عدم هم‌انباشتگی لازم نیست.

در نهایت یک برآورد کارای دیگر که با توجه به قید $d_۱ = d_۲ = ... = d_p = d_*$ توسط مارینا‌سی و رابینسون (۲۰۰۱) معرفی شده و d_* در آن نامعلوم است بصورت زیر است

$$R_*(d) = \log \left| \frac{۱}{s} \sum_{j=۱}^{s} A_j(d_* ۱_p) \right| - \frac{۲p}{s} \sum_{j=۱}^{s} \log j \qquad (۱۹)$$

و $\hat{d} = arg \ \ min(R_*(d))$، برآورد مرحله‌ای نیوتن با همین کارائی بصورت زیر بدست می‌آید

$$\hat{d}_*^{[\nu+۱]} = \hat{d}_*^{[\nu]} - \frac{۱}{۴p} \frac{\partial R(\hat{d}_*^{[\nu]} ۱_p)}{\partial d} \qquad \nu \geq ۰ \qquad (۲۰)$$

و برای $\hat{d}_*^{[۰]}$ یک برآورد با $s^{\frac{۱}{۲}}$ سازگاری در نظر می‌گیریم. برای انجام آزمون فرض برای d_*، $\hat{d}$، $\bar{d}$ و برآوردهای مرحله‌ای نیوتن مطابق بااینها می‌توانیم از آماره والد استفاده کنیم

$$\hat{W}_{\Pi d=\rho} = (\Pi\hat{d} - \rho) \left[۲\Pi \left(I_p + \hat{G}(\hat{d}_۰) \circ \hat{G}(\hat{d}_۰)^{-۱} \right) \Pi' \right] (\Pi\tilde{d} - \rho) \qquad (۲۱)$$

که دارای توزیع حدی $\chi_u^۲$ است و به همین صورت برای آزمون d_* می‌توان بااستفاده از تقریب $\hat{d}_* \sim N(d_*, (۴ps)^{-۱})$ عمل نمود.

تابعهای هدف (۱۵)، (۱۶) و (۱۸) پیشنهاد می‌کنند از آزمونهائی بر اساس اصل چند گانه لاگرانژ[۲۳] (LM) یا اصل نسبت درستنمائی[۲۴] (LR) استفاده کنیم .یک آزمون LM برای $d_۱ = d_۲ = ... = d_p$، که برای فرض مقابلی متفاوت از (۱۰) بکار می‌رود، توسط لوبیتو و رابینسون (۱۹۹۸) معرفی شده است و یک آزمون LR برای حالت $q = ۱$ توسط رابینسون (۱۹۹۸) معرفی شده است. برای آزمون فرض $\Pi d = \rho$ از آماره LM زیر استفاده می‌کنیم

$$LM_{\Pi d=\rho} = s \frac{\partial R(d_۰)}{\partial d} \left[۲\Pi \left(I_p + \hat{G}(\hat{d}_۰) \circ \hat{G}(\hat{d}_۰)^{-۱} \right) \Pi \right] \frac{\partial R(d_۰)}{\partial d} \qquad (۲۲)$$

22) Hadumard 23) Lagrang Multiplier 24) Likelihood Ratio

که در آن $\hat{d}_\circ$، می‌تواند برآوردی باشد که $R(d)$ را تحت فرض $\Pi d = \rho$ مینیمم می‌کند یا می‌تواند یک تقریب نیوتن از d باشد. طبق تئوری مجانبی رابینسون (b ۱۹۹۵) و لوبیتو (۱۹۹۶) آماره (۲۲) دارای توزیع مجانبی χ^2_u است. برای آزمون $H_\circ : d_* = d_\circ$ یک آماره LM بصورت زیر است

$$LM_{d_*=d_{*\circ}} = s(\frac{\partial R(d_\circ)}{\partial d})^2/4p \qquad (۲۳)$$

که تحت فرض صفر دارای توزیع حدی χ^2_1 است. آماره LR برای فرضهای $H_\circ : \Pi d = \rho$, $H_o : d_* = d_\circ$ بترتیب بصورت زیر است

$$LR_{\Pi d=\rho} = 2s\left(R(\hat{d}_\circ) - R(\hat{d})\right) \qquad LR_{d_*=d_\circ} = 2s\left(R(\hat{d}_{*\circ}) - R(\hat{d}_*)\right) \qquad (۲۴)$$

که تحت فرض صفر دارای توزیع مجانبی χ^2_u و χ^2_1 هستند.

۴ برآورد بردارهای هم‌انباشتگی

بعد از برآورد درجه انباشتگی بین متغیرها برای بررسی رابطه بین آنها برآورد β را در رابطه زیر توضیح می‌دهیم

$$y_t = \beta' x_t + e_t \qquad (۲۵)$$

برای بردار قابل مشاهده $z_t = (x'_t, y'_t)$ و طبق رابطه‌های (۱۱) و (۱۲) دوره نگار میانگین‌گیری شده متقابل[25] رابصورت زیر تعریف می‌کنیم

$$F_{xy}(m) = 2Re\left[\frac{2\pi}{n}\sum_{j=1}^{m} I_{xy}(\lambda_j)\right] - \frac{2\pi}{n} I_{xy}(\pi) l(m = \frac{n}{2}) \qquad (۲۶)$$

که درآن $l(.)$ تابع نشانگر و عدد صحیح m در شرط $1 \le m \le \frac{n}{2}$ صدق می‌کند. آخرین جمله معادله (۲۳) وقتی وجود دارد که n زوج باشد و m بیشترین مقدار خود یعنی $\frac{n}{2}$ را بگیرد وقتی که $m = [\frac{n}{2}]$ ($[.]$ تابع جزء صحیح است) عبارت زیر را می‌توانیم استنتاج کنیم

$$\hat{F}_{xy}([\frac{n}{2}]) = \frac{1}{n}\sum_{t=1} n(x_t - \bar{x})(y_t - \bar{y})' \qquad (۲۷)$$

که آن را کواریانس تصحیح شده میانگین[26] می‌نامیم و $\bar{x}$ و $\bar{y}$ میانگین نمونه هستند همانطور که مشاهده می‌کنیم $\hat{F}_{xy}$ فرکانسهای $[\lambda_1, \lambda_m]$ را برای تولید کواریانس نمونه با هم سهیم می‌کند؛

25) Averaged (cross-) preiodogram 26) Mean-corrected sample covarince

معادلات (۷) و (۸) را برای حالتی که تصحیح میانگین صورت نگرفته باشد را نیز می‌توانیم بدست آوریم که بدلیل حجیم شدن مطلب از آوردن آن در اینجا خودداری می‌کنیم.

حال برآورد β به روش کمترین مربعات در حوزه فرکانس[۲۷] (FDLS) را بصورت زیر معرفی می‌کنیم

$$\hat{\beta}_m = \hat{F}_{xx}(m)^{-1}\hat{F}_{xy}(m) \qquad (۲۸)$$

و فرض می‌کنیم که معکوس $\hat{F}_{xx}(m)$ وجود داشته باشد شاموی[۲۸] (۱۹۸۶). یک حالت خاص برای معادله (۲۴) براورد کمترین مربعات (OLS) بصورت زیر است

$$\hat{\beta}_{[\frac{n}{2}]} = \left(\sum_{t+1}^{n}(x_t-\bar{x})(x_t-\bar{x})'\right)^{-1}\sum_{t=1}^{n}(x_t-\bar{x})(y_t-\bar{y})' \qquad (۲۹)$$

وقتی $m \le \frac{n}{2}$ است دو حالت برای رفتار مجانبی توسط ماریناسی و رابینسون (۲۰۰۱) برای این برآورد بصورت زیر در نظرگرفته شده است

$$m \sim Cn \qquad 0 \le C \le \frac{1}{2} \qquad (۳۰)$$

و

$$\frac{1}{m} + \frac{m}{n} \to 0 \qquad (۳۱)$$

در حالت (۲۷) یک دسته از فرکانسهای غیر قابل تولید استفاده می‌شود. این حالت $\hat{\beta}_m$ بوسلیه هانن[۲۹] (۱۹۶۳) معرفی شد و سپس بوسیله رابینسون (۱۹۷۳) و انگل (۱۹۷۴) بررسی شد آنها این روش را«رگرسیون طیفی پنجره‌ای[۳۰]» نامیدند. در حالت (۲۸) یکسری از فرکانسهای فوریه افزایشی دوباره استفاده می‌شوند. اما برآورد β فرکانسهای قابل تولید در اطراف فرکانس صفر انجام می‌شود.

برآورد OLS (۲۶) برای تحلیل سری‌های اتو رگرسیو (AR)، $I(1)$ و $I(2)$ که با هم هم‌انباشته باشند کاربرد فراوانی دارد. در ابتدا این برآورد، برآوردی مطلوب برای مبتکران هم‌انباشتگی [انگل و گرنجر (۱۹۸۷)] بود و سپس بعنوان یک برآورد اولیه برای محاسبه باقیمانده‌ها که برای برآوردهای کاراتری از β بکار می‌رفتند استفاده شد [مثلاً فیلیپس (۱۹۹۱)، فیلیپس و هانسن (۱۹۹۰)]. یک خصوصیت مطلوب برآورد FDLS این است که فرضهای متعامد بودن x_tها با e_t که در برآوردهای رگرسیونی ضروری است در اینجا نیازی نیست با این وجود باز هم $\hat{\beta}_m$ سازگار است چونکه x_t بطور مجانبی بر e_t مشرف است. به هر حال عدم وجود تعامد بین x_t وe_t در برآورد کمترین مربعات و وقتی که x_t ایستا ($d \le \frac{1}{2}$) است باعث بوجود آمدن اریبی در معادلات

27) Frequency Domain Least Square 28) Shumway 29) Hanan 30) Band spectrum regression

همزمان می‌شود. این باعث ایجاد انگیزه در رابینسون (a ۱۹۹۴) شد تا حالت (۲۸) را مورد بررسی قرار دهد. او نشان داد علاوه وجود هم بستگی بین x_t و e_t باز هم $\hat{\beta}_m$ برای β سازگار است. البته وقتی بررسی اطراف فرکانس صفر باشد و چگالی طیفی x_t بر e_t مشرف باشد. (این خصوصیت برای حالت (۲۷) وجود ندارد) بنابراین در سری‌های ایستا برآورد (۲۸) خیلی بهتر از برآورد OLS (۲۶) است. در شرایط کاربردی ممکن است از ایستائی یا ناایستائی مطمئن نباشیم، مخصوصاً در حالتی که سریها انباشته کسری باشند تشخیص ایستائی و ناایستائی خیلی است. به همین دلیل رابینسون و ماریناسی (۱۹۹۷، ۱۹۹۹، ۲۰۰۰) $\hat{\beta}_m$ در حالت نایستائی محض بررسی کردند. دانشمندان دیگری نیز بررسی این برآورد را آغاز کردند. چونکه رابطه هم‌انباشتگی که در بالا توضیح دادیم یک پدیده در فرکانسهای پائین است سهیم کردن فرکانسهای بالا در این برآورد ضرورتی ندارد. در واقع تکنیکهای اصلی که در این زمینه وجود دارند از وارد کردن فرکانسهای خیلی بزرگتر از صفر خودداری می‌کنند. بنابراین فرایندهای ورودی معمولاً باید $\mathrm{I}(0)$ باشند. خصوصیات $\hat{\beta}_m$ وابسته به مرتبه‌های انباشگی $d_1, d_2, \cdots d_p$ و d_e هم است و این خصوصیات در هرکدام از نواحی $d_1, d_2, \cdots d_p$ و d_e درای حالت متفاوتی هستند.

برای اینکه توضیحات ما ساده‌تر شود دیگر به جزئیات و شرایط خاص نمی‌پردازیم و فرض می‌کنیم که مقادیر d برای تمام سریها مشترک باشند. در واقع توضیحات رابینسون و ماریناسی (a ۱۹۹۹)، که کار اصلی آنها نیز روی دوره نگار میانگین‌گیری شده بود، فقط برای حالت $p = ۲$ بررسی شد. ما مؤلفه iام $\hat{\beta}_m$ را با $\hat{\beta}_{im}$ نشان می‌دهیم و فرض می‌کنیم که (۲۸) همچنان برقرار باشد. ما در اینجا فقط نرخ همگرائی را گزارش می‌کنیم که در آن $X_n \sim h(n)$ نشان دهنده این است که $h(n)^{-۱} x_n$ در توزیع همگرا به توزیع یک متغیر تصادفی غیر قابل تولید است، که بیان می‌کند $X_n = O_p(h(n))$ است. حال خصوصیات مجانبی $\hat{\beta}_{im}$ که توسط مارینا‌سی و رابینسون (۲۰۰۱) بیان شده‌اند را در زیر می‌آوریم:

الف) نایستائی کمتر از ریشه واحد باشد $d_i > \frac{۱}{۲}, d_e \geq ۰, d_i + d_e < ۱$

$$\hat{\beta}_{im} - \hat{\beta}_i \sim_d n^{d_e - d_i} m^{۱ - d_{min} - d_e} \qquad \hat{\beta}_{i[\frac{n}{۲}]} - \hat{\beta}_i \sim_d n^{۱ - d_i - d_{min}}, \tag{۳۲}$$

ب) حالت باندی[31] $\frac{۱}{۲} \leq d_i = ۱ - d_e \leq ۱$

$$\hat{\beta}_{im} - \hat{\beta}_i \sim_d n^{۲d_e - ۱} \log m \qquad \hat{\beta}_{i[\frac{n}{۲}]} - \hat{\beta}_i \sim_d n^{۲d_e - ۱} \log n \tag{۳۴}$$

ج) حالت $d_۱ = d_۲ = \cdots = d_p, I(۰)/I(۱)$

$$\hat{\beta}_{im} - \hat{\beta}_i \sim_d n^{-۱} \qquad \hat{\beta}_{i[\frac{n}{۲}]} - \hat{\beta}_i \sim_d n^{-۱} \tag{۳۵}$$

31) Boundary case

د) حالتی که نا ایستائی بزرگتر از ریشه واحد باشد $d_1 = d_2 = ... = d_p > \frac{1}{2}$

$$d_i + d_e > 1 \qquad d_e > 0$$

$$\hat{\beta}_{i[\frac{n}{2}]} - \hat{\beta}_i \sim_d n^{d_e - d_i} \qquad \hat{\beta}_{im} - \hat{\beta}_i \sim_d n^{d_e - d_i} \qquad (۳۶)$$

در اینجا توزیع حدی متغیرهای تصادفی غیر استاندارد است، جزئیات بیشتر در این مورد را در رابینسون و مارینا‌سی (۱۹۹۹) ببینید، مارینا‌سی و رابینسون (۲۰۰۰) نیز تابع حدی آنها استفاده کرده‌اند. در اینجا حالت سوم، حالت عمومی $I(1)/I(0)$ است که در متنهای مربوط به هم‌انباشتگی بر اساس مدل AR وجود دارد. مطالب مربوط به FDLS نیز نشان می‌دهند که یکسری نقصهائی مانند اریبی درجه دوم مخصوصاً نسبت به حالت OLS در آنها وجود دارد. جزئیات در این ورد نیز توسط رابینسون و مارینا‌سی (۱۹۹۹) آمده‌است. معادله (۳۰) در اینجا نشان می‌دهد که نرخ همگرائی FDLS یا بعبارتی توزیع حدی آن بهتر از روش OLS است. در مورد (۲۹) برتری روش FDLS واضح است. حالت (۳۲) نشان می‌دهد وقتی که از یک مقدار بزرگ m برای وارد کردن فرکانسها استفاده کرده‌ایم حذف فرکانسهای بالا تغییری در سرعت همگرائی برآوردها بوجود نمی‌آورد. نقص فرکانسهای بالا نه تنها نقصی ایجاد نمی‌کند حتی باعث بهبود برآوردها نیز می‌شود به همین دلیل است که حذف فرکانسهای بالا باعث ایجاد اریبی در معادلات همزمان می‌شود، به عبارتی دیگر در واریانس فقط فرکانسهای پائین شرکت می‌کنند.

ما در اینجا حالتی که مؤلفه‌های قطعی مانند جمله ثابت، روند خطی و ... در مدل وجود داشته باشند را بررسی نمی‌کنیم. ولی رابینسون و مارینا‌سی (۲۰۰۰) این حالت را بررسی کرده‌اند. اگر مؤلفه‌های قطعی تحت تأثیر مؤلفه‌های تصادفی باشند نتایج حالات (۳۲-۲۸) برای این حالت نیز برقرار است.

۵ بررسی نرخ تورم در ایران

در این قسمت برای نشان دادن کاربرد رابطه هم‌انباشتگی رابطه بین شاخص قیمت کالاها CPI، نرخ ارز غیر رسمی ER، تولید ناخالص داخلی GDPو حجم پول Mدر ایران را بررسی کرده‌ایم. داده‌های مورد استفاده در این مقاله مربوط به لگاریتم سری‌های نام برده از سال ۱۳۸۱ ـ ۱۳۵۰ در ایران است.

برای بررسی ایستائی سریها ابتدا از آزمونهای استاندارد دیکی فولر و دیکی فولر تعمیم یافته که درکارهای عملی بسیار از آنها استفاده می‌شود را بکار برده‌ایم. مقادیر آماره آزمون برای آزمون دیکی فولر تعمیم یافته در جدول (۱) آورده‌ایم. هیچکدام از این مقادیر بزرگتر مقادیر آزمون ADF نیستند تا بتوانیم فرض صفر ناایستائی را رد کنیم. بنابراین هر چهار سری ناایستا از درجه یک هستند. پس از اینکه متوجه شدیم هر چهار سری ناایستا از درجه مشابه یک هستند می‌توانیم آزمون هم‌انباشتگی به روش انگل ـ گرنجر و جوهانسون[۳۲](۱۹۹۱) را روی این سری‌ها انجام دهیم.

32) Juohonsen

جدول ۱: مقادیرآماره آزمون ADF برای لگاریتم سری‌های شاخص قیمت کالاها CPI، نرخ ارز غیر رسمی ER، تولید ناخالص داخلی GDPو حجم پولM

سری زمانی	آمارهآزمونADF	مقادیر بحرانی
Lcpi	۲٫۱۹-	۳٫۵۷-
LM	۳٫۱۴-	۳٫۵۷-
LER	۲٫۷-	۲٫۹-
LGDP	۱٫۷-	۳٫۵۷-

در روش انگل ـ گرنجر آزمون ADF را روی باقیمانده‌های حاصل از معادله رگرسیون زیر انجام می‌دهیم

$$LCPI = ۲٫۶ + ۰٫۱۶۴LER - ۰٫۰۸LGDP + ۰٫۶۸LM + \varepsilon_t \qquad (۳۷)$$

آماره ADF برای باقیمانده‌ها $۲٫۴-$ و مقدار بحرانی آن $۲٫۹-$ بدست آمد که نشان دهنده ایستا بودن باقیماندها است، بنابراین رابطه هم‌انباشتگی بین چهار سری وجود دارد. آزمون جوهانسون نیز رابطه هم‌انباشتگی را تائید کرد که بدلیل زیاد بودن مطلب از آوردن آن خوداری می‌کنیم.

برای بررسی رابطه هم‌انباشتگی کسری ابتدا درجه انباشتگی هرکدام از سریها را برآورد کرده‌ایم که نتایج آن در جدول (۲) نشان داده شده است. مقادیر نشان داده شده در ستون دوم جدول

جدول ۲: برآورد درجه انباشتگی سری‌های زمانی مورد بررسی. δ درجه انباشتگی سری‌های تفاضل‌گیری شده‌است و $\delta + ۱$ درجه انباشتگی سری‌های زمانی اصلی است

سری زمانی	δ	$\delta + ۱$
Lcpi	۰٫۳۹۵	۱٫۳۰۵
LM	۰٫۳۲۸	۱٫۳۲۸
LER	۰٫۰۳۹۳	۱٫۰۳۹۳
LGDP	۰٫۳۲۸	۱٫۳۲۸

(۲) مربوط درجه انباشتگی سری‌های تفاضل‌گیری شده است که آن را با δ نشان داده‌ایم و مقادیر نشان داده شده در ستون سوم جدول درجه انباشتگی سری‌های اصلی است در واقع به δ عدد یک اضافه کرده‌ایم.

پس از برآورد درجه انباشتگی سری‌ها و انجام آزمون فرض معلوم شد که سه متغیر GDP،M و CPIدارای درجه انباشتگی مشابه هستند ولی درجه انباشتگی ER با بقیه متفاوت است. بنابراین رابطه رگرسیونی را بین متغیرهای زیر برآورد کرده‌ایم

$$LCPI = -۲/۶ - ۰/۳LGDP + ۰/۸LM + \varepsilon_t \qquad (۳۸)$$

حال برای اینکه رابطه هم‌انباشتگی بین این متغیرها وجود داشته باشد درجه انباشتگی باقیماندهای مدل بالا باید کوچکتر از درجه انباشتگی متغیرها باشد. پس از برآورد درجه انباشتگی باقیمانده‌ها مقدار آن ۰/۰۳۷ بدست آمد که رابطه هم‌انباشتگی کسری بین این سه متغیر وجود دارد.

۶ نتایج

پس از تعاریف رابطه هم‌انباشتگی، انباشتگی کسری و رابطه هم‌انباشتگی کسری نشان دادیم که تعداد بردارهای هم‌انباشتگی با توجه به تعریف آنها می‌تواند متفاوت باشد.

کار اصلی در بررسی رابطه هم‌انباشتگی کسری برآورد درجه انباشتگی هر کدام از سری‌ها است. ما در اینجا نشان دادیم که برآوردهای نیمه پارامتری در این زمینه کاراتر از برآوردهای پارامتری هستند. پس از برآورد درجه انباشتگی سری‌ها برآورد بردارهای هم‌انباشتگی اهمیت پیدا می‌کند. ما همچنین نشان دادیم که یک روش کارا برای برآورد بردارهای هم‌انباشتگی روش FDLS است که به مراتب کاراتر از روش OLSاست.

ما کاربرد رابطه هم‌انباشتگی و هم‌انباشتگی کسری را برای بررسی رابطه بین لگاریتم سریهای شاخص قیمت کالاها LCPI، نرخ ارز غیر رسمی LER، تولید ناخالص داخلی LGDPو حجم پول LMدر ایران را بررسی کردیم. در رابطه هم‌انباشتگی طبق تعریف انگل ـ گرنجر و همچنین تعریف جوهانسون معلوم شد یک رابطه هم‌انباشتگی بین این سری‌ها وجود دارد که آن را در معادله (۳۷) نشان دادیم. این معادله نشان می‌دهد نرخ تورم در بلند مدت با حجم پول و نرخ ارز غیر رسمی رابطه مستقیم و با تولید ناخالص داخلی رابطه عکس دارد. بنابراین دولت ایران می‌تواند با کاهش حجم پول در دست مردم و کاهش واردات و ارز و همچنین با بالا بردن تولید ناخالص داخلی یعنی تولید کالاهای داخلی وکاهش صادرات نفت نرخ تورم راکاهش دهد.

مراجع

[1] Chan,N.H , trrin ,N.(1995),Infrence for unstable log-memory processes whit applications to fractional unit root autoregressions. Annals of Statistics 23 ,1662-1683 .

[2] Chueng ,Y.W.,Lai K,S, 1993 A fractional cointegration anlysis of purchasing power parity. Journal of Bussines and Economic Statistics 11 ,103-112.

[3] Dicky , D.A.,Fuller,W.A, 1981 Likelihood ratio statistics for autoregressive time series whit unit root .Econometrica 49, 1057-1072.

[4] Engle ,R.F Granger ,C.W.G. ,1987 Cointegration and error correction : representation , estimation and testing . Econometrica 57 , 251-276.

[5] Engle ,R.F Granger ,C.W.G. ,1991 Long-Run Economic Relationship , Oxford University Press , Oxford.

[6] Flores Jr., Szafarz,A., 1996 An enlarge definition of cointegration . Economic Letters 50 , 193-195.

[7] Fox, R.,Taqqu ,M.S., 1986 Large-sample properties of parameter estimate for strongly dependent stationary Gayssian time series .Annals of Statistics 14 ,517-532.

[8] Geweke ,J., Poter-Hudak ,S., 1983 ,The estimstion snd application of long memory time series models .Journal of Time Series Analysis 4 ,221-238.

[9] Hannan E.J., 1963 Regression of time series whit errors of measurment . Biometrika 50 ,293-302 .

[10] Hausman ,J.1978 Misspecification tests in econometrics . Econometrica 46 ,1251-1271.

[11] Hurvich ,C.M ,Deo ,R.,Brodsky ,J.,1998 The mean squer error of Geweck and Poter-hudak's estimate of the memory parameter of long memory time series .Journal of Time Series Analysis 19 ,19-46.

[12] Johansen,S., 1991, Estimation and hypothesis testing of cointegrating vectors in Gaussian vectors in autoregressive models . Econometrica 59,1551-1580.

[13] Johansen,S., 1996,Likelihood-Based infrence in conintegrated vector autoregressive models. 2nd Printing .Oxford University Press ,Oxford. Kunsch ,H., 1987 ,Statistical of aspects of self-similar processes . INn : Prohorov ,Yu,Sazonov ,V.V(Eds.), Proceeding of the first world Congress of the Bernolli Society 1.VNU Science Press Utrecch, pp . 67-74.

[14] Lobato ,I 1995 Multivariate analysis of long memory seies inthe frequency domain . PH.D Thesis , London School of Economics.

[15] Lobato ,I., 1999 A semiparametric tow-step estimator in the multivariate long memory models . Journal of Econometrics 90 129-153.

[16] Marinucci ,D., Robinson,P.M., 1999a ,Alternative forms of fractional brownian motion . Journal of Statistical Planning and Infrence 90 ,129-153.

[17] Marinucci ,D., Robinson,P.M., 1999b ,Finite sample improvements in statistical infrence whit I(1) processes . Journal of Econometrics , forthcoming.

[18] Marinucci ,D., Robinson,P.M., 2000 ,Weake convergence of multivariate fractional processes. Stochastic processes and their application 86 ,103-120.

[19] Marinucci ,D., Robinson,P.M., 2001 ,Semiparametric fractional cointegration nalysis. Journal of Econometrics 105 ,225-247

[20] Philips ,P.C.B.,Ouliaris ,S., 1990 ,Asymptotic properties of residual based cointegration test.Econometrica 58 ,165-193 .

[21] Robinson ,P.M., 1973,The stochastic difference equation whit nonintegral differences . Advance In Applied Probability 6 ,524-545

[22] Robinson ,P.M.,1988 The stochastic difference between econometric statistics . Econometrica 56 ,531-148

[23] Robinson ,P.M.1994a ,Semiparametric analysis of long memory time sries .Annals of Statistics 22 ,519-539

[24] Robinson ,P.M.1994b , Efficient test of nonstationary hypothesis .Journal of the American Statistical Association 89, 1420-1437

[25] Robinson ,P.M.1995a ,Log-priodogram regression of time sreies whit long range dependence.Annals of Statistics 23,1048-1072

[26] Robinson ,P.M.1995b ,Gaussian semiparametric estimation of long range dependence Annals of Statistics 23,1630-1661

[27] Robinson ,P.M.1998 ,comment onreal and spurios long memory properties of stock market data ,by I.G.Lobato and N.E Savin .Journal of Bussines and Economic Statistics 16 ,261-182

[28] Robinson ,P.M., Marinucci ,D.1997 ,Semiparametric frequency-domain analysis of fractional cointegration ,preprint.

[29] Robinson ,P.M., Marinucci ,D.1999 , em Narrow-band analysis of nonstionary processes ,preprint

[30] Robinson ,P.M., Marinucci ,D.2000, The averaged priodogram of nonstionary vector time series . Statistical Infrence for Stochastic Processes 3 ,149-160

[31] Robinson ,P.M.,Yajima . 2002 , Determination of cointegration rank in fractional systems . Journal of Econometrics 106,217-241

معیارهای انتخاب متغیر در داده‌های چند متغیره بر اساس ساختار کوواریانس و همبستگی

نادر نعمت‌الهی[۱]، قدرت‌اله رحمتی[۲]

[۱] گروه آمار دانشگاه علامه طباطبایی
[۲] کارشناسی ارشد آمار دانشگاه علامه طباطبایی

چکیده: مؤلفه‌های اصلی یک ترکیب خطی از متغیرهای اولیه می‌باشند. در عمل هنگامیکه تحلیل مؤلفه‌های اصلی روی تعداد زیادی از متغیرها به کار گرفته می‌شود، ممکن است نتوان مؤلفه‌های اصلی به دست آمده را به آسانی تفسیر کرد. در این حالت می‌توان با انتخاب یک زیر مجموعه از متغیرها که تقریباً همان اطلاعات کلی را داشته باشند، تفسیر مؤلفه‌های اصلی حاصل را راحتتر انجام داد. در این مقاله معیارهای مختلف انتخاب متغیر برای ساختار کوواریانس و همبستگی معرفی و بررسی شده است. چون هر معیار یک زیر مجموعه از متغیرها را انتخاب می‌کند، در نتیجه با استفاده از معیار کارایی توضیح داده شده، مناسب‌ترین زیر مجموعه انتخاب می‌شود.

واژه‌های کلیدی: مؤلفه‌های اصلی، معیارهای انتخاب متغیر، ساختار همبستگی، اندازه کارایی، تحلیل خوشه‌ای

۱ مقدمه

در بسیاری از مباحث علمی از قبیل پزشکی، مدیریت، جامعه شناسی و اقتصاد مطالعات بر روی تعداد زیادی از متغیرهای تصادفی انجام می‌پذیرد که همبستگی متقابلی بین این متغیرها وجود دارد. به دلیل همین همبستگی چنین به نظر می‌رسد که در اینگونه مسایل بایستی بعد مساله را با از دست دادن حداقل اطلاعات کاهش داد. یکی از روش‌هایی که برای کاهش بعد وجود دارد روش تجزیه و تحلیل مؤلفه‌های اصلی[۱] می‌باشد که هتلینگ[۲] ارایه داده است. مؤلفه‌های اصلی $Y_i \quad , i = ۱, ۲, \ldots, p$ یک ترکیب خطی از متغیرهای اولیه $X_۱, X_۲, \ldots, X_p$ به صورت زیر می‌باشند:

$$Y_i = \sum_{j=۱}^{p} e_{ji} X_i \qquad , i = ۱, ۲, \ldots, p$$

1) Principal Components 2) Hotelling

که معمولاً بوسیله بارهای e_{ji} مؤلفه‌های اصلی تعبیر و تفسیر می‌شوند. حال اگر تعداد متغیرهای اولیه یعنی p بزرگ باشد آنگاه تفسیر مؤلفه‌های اصلی آسان نیست. برای رفع این مشکل انتخاب یک زیر مجموعه از متغیرهای اولیه که تقریباً همان اطلاعات کلی را داشته باشند راه حل عملی می‌باشد. در اینجا یک سوال پیش می‌آید که کدامیک از متغیرها ضروری هستند و در صورت ضروری نبودن بعضی از متغیرها آیا می‌توان آنها را حذف کرد؟ یک راه حل برای این سوال پیدا کردن معیارهایی برای انتخاب زیر مجموعه‌ای از متغیرها است که اولین بار توسط بیل[3] و دیگران (۱۹۷۶) بررسی شده است. این معیارها به طور کلی به چهار دسته تقسیم می‌شوند.

۱) معیارهای که بر اساس همبستگی چندگانه بدست می‌آیند.

۲) معیارهایی که از مؤلفه‌های اصلی به دست می‌آیند،

۳) معیارهایی که از متغیرهای اولیه به دست می‌آیند،

۴) معیارهایی که از تحلیل خوشه‌ای به دست می‌آیند.

بعد از بیل و دیگران، جولیف[4] و الکاندری[5] در مقالات مختلف در سال‌های ۱۹۷۲، ۱۹۷۳، ۲۰۰۱ و ۲۰۰۲ به بررسی معیارهای مختلف انتخاب متغیر پرداخته و مک کیب[6] نیز در سال (۱۹۸۴) با استفاده از متغیرهای اصلی معیارهای دیگری را نیز عرضه کرد. کادیما[7] (۱۹۹۵) نیز به بررسی ضعف استفاده از روش معمول انتخاب متغیر بر اساس مؤلفه‌های اصلی پرداخت.

در این مقاله سعی شده است بر اساس اینکه متغیرها هم واحد باشند و یا نه، معیارهای انتخاب متغیر بر اساس ساختار کوواریانس و ساختار همبستگی معرفی و بررسی شوند. بر این اساس در بخش دوم معیارهای انتخاب متغیر برای ساختار کوواریانس ارایه می‌شوند. در بخش سوم نیز معیارهای انتخاب متغیر برای ساختار همبستگی معرفی می‌شوند و کاربرد معیارهای انتخاب متغیر در یک مثال بررسی می‌شود.

۲ معیارهای انتخاب متغیر برای ساختار کوواریانس

در تحلیل چند متغیره بر حسب اینکه متغیرها از لحاظ واحد اندازه‌گیری هم واحد باشند یا نه از ساختار کوواریانس یا همبستگی استفاده می‌شود. در این بخش برای ساختار کوواریانس معیارهای انتخاب متغیر را بررسی می‌کنیم. برای این منظور ابتدا سه خاصیت مؤلفه‌های اصلی که توسط مک کیب (۱۹۸۴) اثبات شده است را بیان می‌کنیم. اکثر معیارهایی که در این بخش و بخش بعد آمده‌اند بر اساس خواص زیر می‌باشند.

خاصیت ۱: به ازای هر عدد صحیح q، $p \leq q \leq 1$ تبدیل زیر را در نظر بگیرید

$$\underset{\sim}{\mathbf{Z}} = \mathbf{B}'\underset{\sim}{\mathbf{X}}$$

3) Beale 4) Jolliffe 5) AL-Kandari 6) Mc Cabe 7) Cadima

که $\underset{\sim}{\mathbf{Z}}$ یک بردار q-بعدی و $\mathbf{B}'$ یک ماتریس $q \times p$ می‌باشند و $\mathbf{B}'\mathbf{B} = \mathbf{I}_q$. در این صورت اثر ماتریس $\sum_Z$ یعنی $tr(\sum_Z)$ ماکزیمم است اگر $\mathbf{B} = \mathbf{P}_q$ باشد، که در آن $\mathbf{P}_q$ شامل q ستون اول ماتریس بردارهای ویژه ماتریس واریانس - کوواریانس متغیرهای اولیه است $\sum_Z = \mathbf{B}'\sum\mathbf{B}$ و ماتریس واریانس - کوواریانس $\underset{\sim}{\mathbf{Z}}$ است.

خاصیت ۲: با در نظرگرفتن همان ترکیب خطی متعامد $\underset{\sim}{\mathbf{Z}} = \mathbf{B}'\underset{\sim}{\mathbf{X}}$ در خاصیت ۱، $tr(\sum_Z)$ مینیمم است اگر $\mathbf{B} = \mathbf{P}_q^*$ که $\mathbf{P}_q^*$ شامل q ستون آخر ماتریس بردارهای ویژه ماتریس واریانس - کوواریانس متغیرهای اولیه است.

خاصیت ۳: همانند خواص ۱ و ۲، تبدیل $\underset{\sim}{\mathbf{Z}} = \mathbf{B}'\underset{\sim}{\mathbf{X}}$ را در نظر بگیرید، در این صورت دترمینان ماتریس $\sum_Z$ یعنی $\det(\sum_Z)$ ماکزیمم است اگر $\mathbf{B} = \mathbf{P}_q$ باشد.

حال معیارهای انتخاب متغیر را برای ساختار کوواریانس بیان می‌کنیم.

معیار ترکیب واریانس (VC): این معیار زیر مجموعه‌ای از متغیرها را انتخاب می‌کند که دارای واریانس‌های بزرگی باشند. یعنی اگر واریانس متغیرها را به ترتیب نزولی مرتب کنیم آنگاه q متغیری که دارای q واریانس اول باشند، انتخاب می‌شوند.

معیار ترکیب بار (LC): در این معیار ابتدا بایستی تحلیل مؤلفه‌های اصلی را روی داده‌ها انجام داده و سپس k مؤلفه اصلی اول که دارای بیشترین سهم در کل واریانس باشند را در نظر می‌گیریم و بار p متغیر اولیه را در این k مؤلفه اصلی اول به طور نزولی با هم مرتب می‌کنیم. سپس آن q متغیری که (بدون تکرار) دارای بزرگترین بار باشند را نگه می‌داریم و بقیه متغیرها را حذف می‌کنیم.

معیار $\mathbf{B}_1$: در این معیار ابتدا تحلیل مؤلفه‌های اصلی را روی p متغیر اولیه انجام داده، سپس آن تعداد از مؤلفه‌های اصلی که بیشتر از α درصد از درصد تجمعی تغییرات کل را توضیح دهند را در نظر گرفته و از هر مؤلفه مورد نظر، متغیری که دارای بیشترین بار باشد را حذف می‌کنیم. تعداد متغیرهای حذف شده برابر است با تعداد مؤلفه‌های اصلی که بیشتر از α درصد تجمعی را توضیح دهند. جولیف (۱۹۷۲) با شبیه‌سازی مقدار مناسب $\alpha = ۰/۸$ را پیشنهاد کرد. اما جولیف در عمل $\alpha = ۰/۸$ را مناسب نمی‌داند و باید با توجه به تعداد متغیرهای مورد نظر مقدار α را تعیین کرد.

معیار $\mathbf{B}_2$: در این معیار ابتدا تحلیل مؤلفه‌های اصلی را روی p متغیر اولیه انجام داده، سپس آن تعداد از مؤلفه‌های اصلی که کمتر از α درصد از درصد تجمعی تغییرات کل را توضیح دهند را در نظر می‌گیریم. در آخر از هر مؤلفه مورد نظر متغیری که دارای بیشترین بار باشد را نگه می‌داریم. برای تعیین مقدار مناسب α همانند معیار B_1 عمل می‌شود.

معیار همبستگی چندگانه (R): این معیار بوسیله یک روش گام به گام به دست می‌آید. بدین صورت که ابتدا متغیری که با $(p-1)$ متغیر دیگر دارای همبستگی چندگانه ماکزیمم باشد را حذف می‌کنیم، زیرا به خوبی توسط $(p-1)$ متغیر دیگر توضیح داده شده است. در گام دوم متغیری که با $(p-2)$ متغیر دیگر دارای همبستگی چندگانه ماکزیمم باشد را حذف می‌کنیم. پس در هر گام هنگامی که m متغیر باقی می‌ماند، متغیری که دارای ماکزیمم همبستگی چندگانه با

$(m-1)$ متغیر دیگر باشد را حذف می‌کنیم و این فرآیند را تا رسیدن به q متغیر ادامه می‌دهیم. شرط توقف برای این روش آن است که تمام همبستگی‌های چندگانه بین متغیرها باقیمانده کمتر از یک مقدار R_0 باشد. جولیف با استفاده از شبیه‌سازی مقدار مناسب R_0 را برابر ۰٫۱۵ پیشنهاد کرده است.

معیار دسته‌بندی همبستگی (CC): در این معیار متغیرهایی که بیشترین ارتباط را با q مؤلفه اصلی اول داشته باشند را نگه می‌داریم. برای انجام این فرآیند به این صورت عمل می‌کنیم که ابتدا مقدار همبستگی بین متغیرها و q مؤلفه اصلی اول را با هم به ترتیب نزولی مرتب می‌کنیم، سپس متغیرهایی که دارای مقدار همبستگی بزرگی باشند را نگه می‌داریم.

تعمیم معیار دسته‌بندی بارها (NLC): در این معیار میانگین قدر مطلق بار متغیرها را در q مؤلفه اصلی اول در نظر می‌گیریم و متغیرهایی که دارای بیشترین میانگین قدر مطلق بار باشند را نگه می‌داریم.

تعمیم معیار دسته‌بندی همبستگی (NCC): در این معیار آن q متغیر را نگه می‌داریم که دارای بیشترین میانگین قدر مطلق مقادیر همبستگی با مؤلفه‌های اصلی مورد نظر باشند.

معیار $\mathbf{M_1}$: در این معیار آن q متغیری نگه داشته می‌شوند که دارای کوچکترین دترمینان ماتریس کوواریانس شرطی متغیرهای حذف شده به شرط متغیرهای نگه داشته شده باشند. یعنی q متغیری که کوچکترین دترمینان ماتریس $\mathbf{S}_{22,1} = \mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$ را به دست می‌دهند، که در آن $\mathbf{S}_{22}$ ماتریس کوواریانس متغیرهای حذف شده و $\mathbf{S}_{22}$ ماتریس کوواریانس متغیرهای نگه داشته شده می‌باشند.

معیار $\mathbf{M_2}$: در این معیار آن q متغیری نگه داشته می‌شوند که دارای کوچکترین اثر ماتریس کوواریانس شرطی متغیرهای حذف شده به شرط متغیرهای نگه داشته شده باشند.

معیار$\mathbf{M_3}$: در این معیار آن q متغیری نگه داشته می‌شوند که دارای کوچکترین نرم ماتریس کوواریانس شرطی متغیرهای حذف شده به شرط متغیرهای نگه داشته شده باشند.

معیار تحلیل خوشه‌ای (C): این معیار بر اساس منتسب کردن p متغیر اولیه به q خوشه از متغیرها عمل می‌کند. در این روش متغیرها بر اساس تحلیل خوشه‌ای سلسله مراتبی به q خوشه تقسیم می‌شوند و سپس از هر خوشه یک متغیر را انتخاب می‌کنیم. انتخاب متغیر از هر خوشه را می‌توان به سه روش زیر انجام داد.

۱ــ آخرین متغیر انتخاب شده در هر خوشه،
۲ــ انتخاب یک متغیر از متغیرهای اولیه یا متغیر داخلی هر خوشه،
۳ــ انتخاب تصادفی یک متغیر از هر خوشه.

الگوریتم تحلیل خوشه‌ای سلسله مراتبی به صورت زیر است.

۱ــ تعریف یک اندازه تشابه r_{XY}، بین هر دو گروه از متغیرهای X و Y (در مرحله اول هر متغیر یک گروه محسوب می‌شود)،
۲ــ محاسبه r_{XY} برای تمام زوج متغیرهای گروهها که تعداد آنها در مرحله اول $\frac{1}{2}p(p-1)$

می‌باشد،
۳ــ ترکیب دو گروه از متغیرها که دارای ماکزیمم r_{XY} باشند،
۴ــ محاسبه r_{XY} برای گروههای جدید و برگشت به گام سوم.

این فرآیند تا هنگامی که تمام r_{XY} بین گروهها کمتر از یک مقدار r_0 باشند ادامه پیدا می‌کند. روشهای متفاوتی برای در نظر گرفتن اندازه تشابه r_{XY} وجود دارد. در این مقاله از دو روش اتصال متوسط[8] و اتصال تکی[9] استفاده می‌کنیم که در آن r_{XY} و r'_{XY} به صورت زیر به دست می‌آیند

$$r_{XY} = \frac{\sum_{i\in X}\sum_{j\in Y} r_{ij}}{n_1 n_2}$$
$$r'_{XY} = \max_{i\in X j\in Y} r_{ij}$$

که n_1 وn_2 تعداد متغیرهای موجود در هر گروه X و Y، r_{ij} ضریب همبستگی بین متغیر iام و متغیر jام می‌باشند. جولیف (۱۹۷۲) با استفاده از شبیه‌سازی مقدار r_0 را برای روش‌های اتصال متوسط و اتصال تکی به ترتیب برابر ۰٫۴۵ و ۰٫۵۵ در نظر گرفته است.

اندازه کارایی: بعد از اینکه هر معیار انتخاب متغیر یک زیر مجموعه از متغیرها را معرفی نمود، بایستی یکی از این زیر مجموعه‌ها را انتخاب کنیم. در اینجا یک سوال به وجود می‌آید که کدام یک از این زیر مجموعه‌ها مناسب است؟ برای پاسخ به این سوال، نیاز به معرفی یک اندازه کارایی داریم که نزدیک بودن نتایج مؤلفه‌های اصلی روی مجموعه کامل و زیر مجموعه‌ای از آنها را بسنجد. در این مقاله دو اندازه کارایی را در نظر می‌گیریم.

اندازه کارایی اول عبارت است از

$$E_1 = \frac{\frac{\sum_{j=1}^{k}\lambda_j^n}{tr(\mathbf{S})}100}{\frac{\sum_{j=1}^{k}\lambda_j}{tr(\mathbf{S})}100}$$

که در آن λ_j، λ_j^n و $\mathbf{S}$ به ترتیب واریانس jامین مؤلفه اصلی اولیه، واریانس jامین مؤلفه اصلی جدید و ماتریس واریانس کوواریانس نمونه می‌باشند که برای k مؤلفه اصلی اول مقایسه می‌شوند.

برای به دست آوردن اندازه کارایی دوم، ابتدا درصدی از تغییرات کل که توسط زیر مجموعه‌ای از متغیرها توضیح داده می‌شود را در نظر می‌گیریم، که به صورت زیر است

$$\frac{\sum_{i=1}^{q} S_{i,r}^2 + \sum_{h=1}^{p-q}(R_{h,r}^2 \times S_{h,d}^2)}{tr(\mathbf{S})}$$

8) Average Linkage 9) Single Linkage

که در آن $S^2_{i,r}$ واریانس نمونه iامین متغیر در متغیرهای نگه داشته شده، $S^2_{h,d}$ واریانس نمونه hامین متغیر در متغیرهای حذف شده و $R^2_{h,r}$ نیز مربع همبستگی چندگانه hامین متغیر حذف شده با q متغیر نگه داشته شده می‌باشند. سپس رابطه بالا را با رابطه زیر مقایسه می‌کنیم

$$\frac{\sum_{i=1}^{p}(S_i^2 \times R^2_{i,y_1^n,y_2^n,\cdots,y_k^n})}{tr(\mathbf{S})}$$

که در آن $R^2_{i,y_1^n,y_2^n,\cdots,y_k^n}$ مربع همبستگی چندگانه iامین متغیر با k مؤلفه اصلی اولیه می‌باشد. یعنی

$$E_2 = \frac{\frac{\sum_{i=1}^{q} S^2_{i,r} + \sum_{h=1}^{p-q}(R^2_{h,r} \times S^2_{h,d})}{tr(\mathbf{S})}100}{\frac{\sum_{i=1}^{p}(S_i^2 \times R^2_{i,y_1^n,y_2^n,\cdots,y_k^n})}{tr(\mathbf{S})}100}$$

۳ معیارهای انتخاب متغیر مربوط به ساختار همبستگی

هنگامیکه داده‌های چند متغیره هم واحد نباشند، تحلیل مؤلفه اصلی بر اساس ماتریس همبستگی که بر اساس متغیرهای استاندارد شده $z_i = \frac{x_i - \bar{x}}{s}$ بدست می‌آید، انجام می‌شود. در این حالت برای انتخاب یک زیر مجموعه از متغیرها بایستی معیارهای جدیدی تعریف شوند. البته بعضی از معیارها انتخاب متغیر از قبیل معیار همبستگی چندگانه (R)، معیارهای M_1، M_2، M_3و معیار تحلیل خوشه‌ای (C) برای ساختار کوواریانس و همبستگی مشترک می‌باشند. در ادامه معیارهایی که فقط مربوط به ساختار همبستگی هستند ارایه می‌شود.

معیار B_1^*: در این معیار، ابتدا تحلیل مؤلفه‌های اصلی را روی تمام p متغیر اولیه انجام داده و سپس آن تعداد از مؤلفه‌های اصلی که مقادیر ویژه آنها کمتر از λ_0 باشد را در نظر گرفته و از هر مؤلفه مورد نظر، متغیری که دارای بیشترین بار باشد را حذف می‌کنیم. تعداد متغیرهای حذف شده برابر تعداد مقادیر ویژه کوچکتر از λ_0 می‌باشد. جولیف (۱۹۷۲) با شبیه‌سازی مقدار مناسب $\lambda_0 = 0.7$ را پیشنهاد کرده است.

معیار B_2^*: در این معیار نیز ابتدا تحلیل مؤلفه‌های اصلی را روی تمام p متغیر اولیه انجام داده و سپس آن تعداد از مؤلفه‌های اصلی که مقادیر ویژه آنها بزرگتر از λ_0 باشد را در نظر گرفته و از هر مؤلفه مورد نظر، متغیری که دارای بیشترین بار باشد را نگه می‌داریم. تعداد متغیرهای نگه داشته شده برابر تعداد مقادیر ویژه بزرگتر از λ_0 می‌باشد. جولیف (۱۹۷۲) با شبیه‌سازی مقدار مناسب $\lambda_0 = 0.7$ را پیشنهاد کرده است.

معیار M_1^*: در این معیار آن q متغیری را نگه می‌داریم که دارای کوچکترین دترمینان ماتریس

همبستگی شرطی متغیرهای حذف شده به شرط متغیرهای نگه داشته شده باشند. یعنی q متغیری که کوچکترین دترمینان ماتریس $R_{۲۲/۱} = R_{۲۲} - R_{۲۱}R_{۱۱}^{-۱}R_{۲۱}$ را به دست می‌دهند، که در آن $R_{۲۲}$ ماتریس همبستگی متغیرهای حذف شده و $R_{۱۱}$ ماتریس همبستگی متغیرهای نگه داشته شده می‌باشند.

معیار $M_۲^*$: در این معیار آن q متغیری را نگه می‌داریم که دارای کوچکترین اثر ماتریس همبستگی شرطی متغیرهای حذف شده به شرط متغیرهای نگه داشته شده باشند.

معیار $M_۳^*$: در این معیار آن q متغیری را نگه می‌داریم که دارای کوچکترین نرم ماتریس همبستگی شرطی متغیرهای حذف شده به شرط متغیرهای نگه داشته شده باشند.

اندازه کارایی: در این حالت برای انتخاب زیر مجموعه مناسبی از متغیرها از اندازه کارایی زیر استفاده می‌کنیم.

$$E_۱^* = \frac{\frac{\sum_{j=۱}^{k} \lambda_j^n}{p} ۱۰۰}{\frac{\sum_{j=۱}^{k} \lambda_j}{p} ۱۰۰}$$

که این رابطه برای k مؤلفه اصلی اول محاسبه می‌شود.

کاربرد مباحث فوق را روی داده‌های مثال زیر که از کتاب «آشنایی با روش‌های آماری چند متغیره» تالیف مانلی[۱۰]، ترجمه دکتر محمد مقدم و دیگران (۱۳۷۳) گرفته شده است، به کار می‌بریم.

مثال: مقدار پروتیین موجود در ۹ گروه غذایی برای ۲۵ کشور اروپایی اندازه گرفته شده، که این ۹ گروه غذایی عبارتند از گوشت قرمز ($X_۱$)، گوشت سفید ($X_۲$)، تخم مرغ ($X_۳$)، شیر ($X_۴$)، ماهی ($X_۵$)، غلات ($X_۶$)، غذاهای نشاسته‌ای ($X_۷$)، آجیل و دانه‌های روغنی ($X_۸$) و میوه‌جات و سبزیجات ($X_۹$). نتایج مربوط به استفاده از معیارهای انتخاب متغیر برای انتخاب زیر مجموعه‌ای از متغیرها بر اساس ساختار کوواریانس در جدول ۱ ارایه شده است.

با توجه به جدول ۱، بر اساس اندازه کارایی $E_۱$ و $E_۲$، مناسب‌ترین زیر مجموعه عبارت است از

$$X_۱, X_۲, X_۴, X_۵, X_۶, X_۷, X_۸, X_۹$$

نتایج مربوط به استفاده از معیارهای انتخاب متغیر برای انتخاب زیر مجموعه‌ای از متغیرها بر اساس ساختار همبستگی در جدول ۲ ارایه شده است.

با توجه به اندازه کارایی جدول ۲، مناسب‌ترین زیر مجموعه عبارت است از

$$X_۱, X_۲, X_۴, X_۵, X_۶, X_۷, X_۸, X_۹$$

توجه کنید که برای تعیین زیر مجموعه‌ای از متغیرها توسط معیارهای $M_۱$، $M_۲$ و $M_۳$ (و یا معادلاً $M_۱^*$، $M_۲^*$ و $M_۳^*$) بایستی تمام افرازهای ممکن ماتریس واریانس کوواریانس (یا همبستگی)

10) Manly

جدول ۱: معیارهای انتخاب متغیر و زیر مجموعه های انتخاب شده بر اساس ساختارکوواریانس

معیار انتخاب متغیر	متغیرهای انتخاب شده	E_1	E_2
VC	X_2, X_4, X_6	۰/۹۶	۰/۹۴
LC	X_1, X_2, X_4, X_6	۰/۹۷	۰/۹۹
B_1	X_1, X_2, X_4, X_6	۰/۹۷	۰/۹۹
B_2	X_1, X_2, X_4, X_6	۰/۹۷	۰/۹۹
R	X_1, X_2, X_5, X_6, X_9	۰/۷۲	۰/۸۹
CC	X_1, X_2, X_4, X_6	۰/۹۷	۰/۹۹
NLC	X_1, X_2, X_4, X_6	۰/۹۷	۰/۹۹
NLC	X_1, X_2, X_4, X_8	۰/۴۱	۰/۵۷
M_1	$X_1, X_2, X_4, X_5, X_6, X_7, X_9$	۰/۹۸	۱/۰۵
M_2	$X_1, X_2, X_4, X_5, X_6, X_7, X_8, X_9$	۱	۱/۰۶
M_3	$X_1, X_2, X_4, X_5, X_6, X_7, X_8, X_9$	۱	۱/۰۶
C	$X_1, X_2, X_4, X_5, X_6, X_7, X_8, X_9$	۱	۱/۰۶

جدول ۲: معیارهای انتخاب متغیر و زیر مجموعه‌های انتخاب شده بر اساس ساختار همبستگی

معیار انتخاب متغیر	زیرمجموعه‌های انتخاب شده	E_1^*
B_1^*	X_2, X_3, X_6, X_7, X_9	۰/۶۴
B_2^*	X_1, X_5, X_6, X_7, X_8	۰/۶۴
R	X_1, X_2, X_5, X_6, X_9	۰/۶۴
M_1	$X_1, X_2, X_4, X_5, X_6, X_7, X_9$	۰/۸
M_2	$X_1, X_2, X_4, X_5, X_6, X_7, X_8, X_9$	۰/۹۰۵
M_3	$X_1, X_2, X_4, X_5, X_6, X_7, X_8, X_9$	۰/۹۰۵
M_1^*	X_7	۰/۱۲
M_2^*	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_9$	۰/۹
M_3^*	$X_1, X_2, X_4, X_5, X_6, X_7, X_8, X_9$	۰/۹۰۵
C	$X_1, X_2, X_5, X_6, X_7, X_8, X_9$	۰/۸۱

نمونه را بدست آورده و سپس دترمینان، اثر و یا نرم ماتریس $\mathbf{S}_{۲۲,۱}$ یا $(\mathbf{R}_{۲۲,۱})$ را برای هر کدام حساب کرده و مقدار مینیمم را بدست آوریم. برای انجام این روش برنامه‌ای با استفاده از MATLAB تهیه شده است که نتایج این برنامه در جداول ۱ و ۲ مشاهده می‌گردد.

نتیجه: در تحلیل داده‌های چند متغیره، اگر تعداد متغیرها زیاد باشد آنگاه بوسیله معیارهای انتخاب متغیر گفته شده در فوق زیر مجموعه‌هایی از این متغیرها انتخاب می‌شود. سپس با استفاده از معیارهای کارایی توضیح داده شده مناسب‌ترین معیار و در نتیجه مناسب‌ترین زیر مجموعه از متغیرها جهت تحلیل مؤلفه‌های اصلی انتخاب می‌شود.

مراجع

[۱] مانلی، آشنایی با روش‌های آماری چند متغیره، ترجمه دکتر محمد مقدم و دیگران، ۱۳۷۳، انتشارات پیشتاز علم.

[2] AL-Kandari, N. (2002). Correlation structures and variable selection in principal component analysis. unpublished manuscript.

[3] AL-Kandari, N. and Jolliffe, I. (2001).Variable selection and interpretation of covariance principal components. Communication in statistics: simulation and computation, 30(2), 339-354.

[4] Beale, E.M.L. Kendall, M.G. and Mann, D.W. (1967) .The discarding of variable in multivariate analysis. Appl. Statist., 16, 223-236.

[5] Cadima, J.F.C.L., Jolliffe, I.T. (1995). Loading and correlation in the interpretation of principal components. Journal of applied statistics, 22(2), 203-314.

[6] Jolliffe, I.T. (1972). Discarding variables in a principal component analysis : artificial data. Appl. Statist. 21, 160-173.

[7] Jolliffe, I.T. (1973). Discarding variables in a principal component analysis : artificial data. Appl. Statist. 22, 21-31.

[8] Mc Cabe, G.P. (1984). Principal variables. Tecnometrics, 26(2), 137-144.

پیشگویی در نمونه‌گیری از جامعه متناهی تحت مدل‌های خطای اندازه‌گیری و کاربرد آن در برآورد هزینه و درآمد خانوار

نادر نعمت‌الهی[۱]، پوریا رضاسلطانی[۲]

[۱] گروه آمار دانشگاه علامه طباطبایی
[۲] سازمان مدیریت و برنامه‌ریزی کشور

چکیده: در برخی از بررسی‌های نمونه‌ای عملی ممکن است مقادیر واقعی مشخصه‌ها (متغیرها) مشاهده نشوند، بلکه مقادیر دیگری که با خطاهای اندازه‌گیری آمیخته شده‌اند، مشاهده شوند. در این مقاله با فرض مدل خطای اندازه‌گیری ساده‌ای بین مقدار واقعی و مقدار مشاهده شده مشخصه، پارامترهای جامعه متناهی به روش‌های کلاسیک، بیز و مینیماکس پیشگویی می‌شوند. سپس با استفاده از داده‌های طرح هزینه و درآمد خانوار، میانگین هزینه سالیانه خانوارها با فرض مدل خطای اندازه‌گیری پیشگویی و با برآورد جاری که بدون در نظر گرفتن خطای اندازه‌گیری محاسبه شده، مقایسه می‌شود.

واژه‌های کلیدی: خطاهای اندازه‌گیری، پیشگویی، جامعه متناهی، نمونه‌گیری دومرحله‌ای، پیشگوی بیز و پیشگوی مینیماکس

۱ مقدمه

مبحث نمونه‌گیری یکی از مباحث مهم علم آمار و دارای کاربرد فراوان است. در یک بررسی نمونه‌ای (نمونه‌گیری) با دو نوع خطا، خطای نمونه‌گیری[۱] و خطای غیرنمونه‌گیری[۲] سروکار داریم.

در یک بررسی نمونه‌ای، اطلاعات به دست آمده از نمونه مبنای نتیجه‌گیری دربارهٔ جامعه را تشکیل می‌دهد. از آنجایی که نمونه صرفاً بخشی از جامعه است، این اطلاعات کامل نبوده و بنابراین هرگونه نتیجه‌گیری همراه با خطا خواهد بود. این نوع خطا که به دلیل استفاده از یک نمونه به جای سرشماری ایجاد می‌شود، اصطلاحاً خطای نمونه‌گیری گفته می‌شود. بدیهی است که تنها راه حذف این خطا، انجام یک سرشماری است.

در مراحل اجرای یک نمونه‌گیری امکان بروز خطاهایی از قبیل خطای پوشش، خطای اندازه‌گیری، خطا در تکمیل پرسشنامه و خطای پردازش داده‌ها وجود دارد. تأثیر این گونه خطاها بر نتایج بررسی‌ها، اصطلاحاً خطای غیرنمونه‌گیری گفته می‌شود.

1) Sampling Error 2) Nonsampling Error

در یک بررسی نمونه‌ای معمولاً کاربران برای ارزیابی دقت روش به کار گرفته شده به خطاهای استاندارد [۳] برآورد شده متکی هستند. خطاهای استاندارد برآورد شده تنها منعکس‌کنندهٔ خطاهای نمونه‌گیری هستند و برای استنباط‌های آماری از قبیل بازه‌های اطمینان و آزمون فرض‌ها استفاده می‌شوند. نکتهٔ مهمی که بایستی متذکر شد این است که در روش‌های متفاوت نمونه‌گیری که معمولاً مطالعه می‌شوند، فرمول‌هایی که ارائه می‌شوند مبتنی بر این فرض هستند که وقتی واحدهای نمونه مشخص می‌شوند و به اندازه‌گیری مشخصهٔ آنها پرداخته می‌شود و براساس این اندازه‌ها، فرمول‌ها به صورت ریاضی و احتمالی بیان می‌شوند، خطای اندازه‌گیری وجود ندارد و تنها خطای موجود، خطای نمونه‌گیری است. در حالی که در برخی از بررسی‌های نمونه‌ای عملی ممکن است مقدار واقعی مشخصه، مشاهده نشود بلکه مقادیر دیگری که با خطاهای اندازه‌گیری آمیخته شده‌اند، مشاهده شوند.

خطای اندازه‌گیری به مشاهدهٔ مشخصه‌هایی که در یک بررسی اندازه‌گیری می‌شوند مربوط است و بعضی وقت‌ها به عنوان «خطای مشاهده‌ای» نسبت داده می‌شود. خطای اندازه‌گیری به عنوان قسمتی از خطای گردآوری داده‌ها، در مقابل خطای نمونه‌گیری، بی‌پاسخی، پوشش یا پردازش داده‌ها پدید می‌آید. خطای اندازه‌گیری می‌تواند از چهار منبع به وجود آید: پرسشنامه، روش گردآوری داده‌ها، مصاحبه‌گر و پاسخگو.

در یک بررسی نمونه‌ای، اگر یک مدل خطای اندازه‌گیری بین مقدار واقعی و مقدار مشاهده شده مشخصه، در نظر گرفته شود آن‌گاه می‌توان پارامترهای جامعه متناهی را به طور واقعی‌تری پیشگویی نمود.

در خصوص به کار بردن مدل‌های خطای اندازه‌گیری، ابتدا اسپرنت[۴](۱۹۶۶) مدل رگرسیونی که در آن متغیرهای تبیینی با خطای اندازه‌گیری آمیخته شده‌اند را معرفی کرد و روشی براساس رهیافت کمترین توان‌های دوم تعمیم‌یافته، برای برآورد ضریب‌های رگرسیونی خطی پیشنهاد کرد. بولفارین[۵](۱۹۹۱)؛ موخوپادیای[۶] (۱۹۹۲ و ۱۹۹۴)؛ باتاچاریا[۷](۱۹۹۷)؛ پیشگوی کلاسیک میانگین و واریانس جامعهٔ متناهی را تحت مدل مکان خطای اندازه‌گیری در نظر گرفتند. در مورد پیشگویی بیز و مینیماکس بولفارین و زاکس[۸](۱۹۹۱) مقاله‌ای ارائه کردند.

در این مقاله با فرض مدل خطای اندازه‌گیری ساده‌ای بین مقدار واقعی و مقدار مشاهده شده مشخصه، پارامترهای جامعه متناهی به روشهای کلاسیک، بیز و مینیماکس پیشگویی می‌شوند. سپس با استفاده از داده‌های طرح هزینه و درآمد خانوار، میانگین هزینه سالیانه خانوارها پیشگویی و با برآورد جاری مقایسه می‌شود.

بر این اساس در بخش ۲ بطور خلاصه به نمونه‌گیری مبتنی بر طرح و مبتنی بر مدل اشاره شده و در بخش ۳ پیشگویی تحت مدل خطای اندازه گیری جمعی مطرح می‌گردد. در بخش ۴ پیشگویی بیز و مینیماکس میانگین جامعه مورد بررسی قرار گرفته و پیشگویی میانگین هزینه سالیانه خانوارها در بخش ۵ ارائه می‌شود.

---

3) Standard Errors 4) Sprent 5) Bolfarine 6) Mukhopadhyay
7) Bhattacharyya 8) Zacks

۲ نمونه‌گیری مبتنی بر طرح و نمونه‌گیری مبتنی بر مدل

در رهیافت مبتنی بر طرح در بررسی نمونه‌ای، مقادیر متغیر (مشخصه) مورد نظر (Y ـ مقادیر) جامعه به عنوان کمیت‌های ثابت در نظر گرفته می‌شوند. در این رهیافت احتمال‌های انتخاب نمونه‌های مرتبط با طرح نمونه‌گیری مطرح هستند که در به دست آوردن امید ریاضی، واریانس، اریبی و ویژگی‌های دیگر برآوردگرها استفاده می‌شوند.

از طرف دیگر، در رهیافت مبتنی بر مدل، مقادیر متغیرهای مورد نظر جامعه به عنوان متغیرهای تصادفی در نظر گرفته می‌شوند و ویژگی‌های برآوردگرها به توزیع توأم این متغیرهای تصادفی وابسته هستند. به طور مثال وقتی فشار خون اندازه‌گیری می‌شود، مقدار مشاهده شده برآوردی از مقدار واقعی است. بنابراین مقدار واقعی فشار خون واحدها به طور صحیح مشاهده نمی‌شوند بلکه مقادیر مشاهده شده، برآوردهایی از فشار خون واقعی واحدها هستند.

فرض کنید $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$ به برداری از Y مقادیر مربوط به N واحد جامعه اشاره کند. از دیدگاه طرح، این Y ـ مقادیر، مقادیر ثابتی هستند. از دیدگاه طرح، این Y ـ مقادیر، متغیرهای تصادفی با توزیع توأم ξ هستند. فرض کنید $p(s)$ به احتمال انتخاب نمونه s تحت یک طرح نمونه‌گیری اشاره کند. s دنباله‌ای یا زیرمجموعه‌ای از واحدهای جامعه است. به وسیلهٔ این نمونه می‌توان برخی از کمیت‌های جامعه (Y_0) را برآورد یا پیشگویی کرد. Y_0 می‌تواند میانگین جامعه یا مقدار کل جامعه باشد. پیشگو یا برآوردگر $\hat{Y}_0$، تابعی از Y ـ مقادیر نمونه است.

برآوردگر یا پیشگوی $\hat{Y}_0$ برای Y_0 طرح ـ نااریب[9] گفته می‌شود اگر امید ریاضی شرطی آن به شرط تحقق N، Y مقادیر جامعه، برابر با مقدار تحقق‌یافتهٔ Y_0 باشد. یعنی

$$E(\hat{Y}_0|\mathbf{Y}) = Y_0$$

که در آن $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$ است. برآوردگر یا پیشگوی $\hat{Y}_0$ برای Y_0 مدل ـ نااریب[10] گفته می‌شود اگر به شرط هر نمونه s، امید ریاضی شرطی $\hat{Y}_0$ برابر امید ریاضی Y_0 باشد. یعنی

$$E(\hat{Y}_0|s) = E(Y_0)$$

پیشگوی مدل ـ نااریب $\hat{Y}_0$ برای Y_0 تحت توزیع توأم ξ جامعه و هر طرح نمونه‌گیری انتخاب نمونه s به شرط نمونه s، نااریب است. در نتیجه این نااریبی به طرح نمونه‌گیری که با استفاده از آن نمونه s انتخاب شده است، بستگی ندارد.

برآوردگر یا پیشگوی $\hat{Y}_0$ برای Y_0 نااریب است (یعنی نااریب غیرشرطی)، اگر امید ریاضی $\hat{Y}_0$ برابر امید ریاضی Y_0 باشد $(E(\hat{Y}_0) = E(Y_0))$.

9) Design-Unbiased 10) Model-Unbiased

۳ پیشگویی تحت مدل خطای اندازه‌گیری جمعی

فرض کنید جامعهٔ متناهی P با تعداد مشخص N واحد با برچسب‌های $1,\ldots,i,\ldots,N$ نمایش داده شده باشد. در ارتباط با واحد i ـ ام جامعه، y_i مقدار واقعی مشخصهٔ مورد مطالعه است که نمی‌توان آن را به طور صحیح مشاهده کرد اما یک مقدار متفاوت Y_i که با خطای اندازه‌گیری آمیخته شده است، مشاهده می‌شود. فولر[11] (۱۹۸۷، ۱۹۸۹) رفتار کلی برای استنباط تحت مدل‌های خطای اندازه‌گیری جمعی را مورد مطالعه قرار داده است. بولفارین (۱۹۹۱)، موخوپادیای (۱۹۹۲ و ۱۹۹۴) و باتاچاریا (۱۹۹۷) و دیگران مدل ذیل را برای پیشگویی میانگین و واریانس جامعهٔ متناهی تحت مدل خطای اندازه‌گیری در نظر گرفتند.

$$\begin{aligned} y_i &= \mu + e_i \quad , \quad e_i \sim N(0, \sigma_{ee}), \\ Y_i &= y_i + u_i \quad , \quad u_i \sim N(0, \sigma_{uu}), \end{aligned} \qquad (1)$$

که در آن μ (میانگین مقدار واقعی مشخصهٔ مورد نظر در جامعه)، $(\sigma_{ee} > 0)$ و $(\sigma_{uu} > 0)$ مقادیر ثابت هستند و نماد $e_i \sim N(0, \sigma_{ee})$ به متغیرهای تصادفی نرمال مستقل و دارای توزیع یکسان $(i.i.d.)$[12]، با میانگین صفر و واریانس σ_{ee} اشاره می‌کند و e_i و u_j برای $i, j = 1, 2, \ldots, N$ از یکدیگر مستقل هستند.

برای پیشگویی میانگین جامعه، ردهٔ پیشگوهای خطی ذیل در نظر گرفته می‌شود

$$g(s, \mathbf{Y}_s) = b_s + \sum_{i \in s} b_{is} Y_i$$

که در آن b_s و b_{is} مقادیر ثابتی هستند که به Y ـ مقادیر بستگی ندارند و $\mathbf{Y}_s = (Y_1, Y_2, \ldots, Y_n)$ به مجموعهٔ مقادیر مشاهده شدهٔ واحدهای نمونه s اشاره می‌کند.

پیشگوی $g(s, \mathbf{Y}_s)$ برای میانگین جامعه $(\bar{y})$ نااریب خواهد بود اگر و تنها اگر

$$E\big(g(s, \mathbf{Y}_s)\big) = E(b_s + \sum_{i \in s} b_{is} Y_i) = E(\bar{y}) = \mu$$

با توجه به این‌که میانگین نمونه $(\bar{Y}_s = \frac{1}{n} \sum_{i \in s} Y_i)$ در شرط فوق صدق می‌کند از این‌رو پیشگوی نااریب میانگین جامعه $(\bar{y})$ خواهد بود. همچنین با توجه به این‌که Y_i ـ ها از یکدیگر مستقل و دارای توزیع نرمال هستند، با استفاده از قضیهٔ برآورد (پیشگو) MVU[13] رائو[14] (۱۹۷۳)، میانگین نمونه $(\bar{Y}_s)$ پیشگوی نااریب با مینیمم واریانس جامعه $(\bar{y})$ است.

نتیجه حاصل از بخش قبل این است که در هر طرح نمونه‌گیری ناآگاهی‌بخش با اندازهٔ مؤثر ثابت نمونه (n) میانگین نمونه پیشگوی بهینهٔ میانگین جامعه است. بنابراین $\bar{Y}_s$ برای هر طرح

11) Fuller 12) independent identically distributed 13) Minimum Variance Unbiased 14) Rao

نمونه‌گیری ناآگاهی‌بخش با اندازهٔ مؤثر ثابت نمونه (n) و تحت مدل (۱) پیشگوی بهینه برای میانگین جامعه $(\bar{y})$ است.

از طرف دیگر با معرفی این‌که V به واریانس متناظر با عملیات توأم مدل (۱) و طرح نمونه‌گیری p اشاره می‌کند، واریانس پیشگوی میانگین جامعه به صورت ذیل حاصل می‌شود

$$V(\bar{Y}_s - \bar{y}) = (\frac{۱}{n} - \frac{۱}{N})\sigma_{ee} + \frac{۱}{n}\sigma_{uu} = \frac{r_۰}{N^۲} \qquad (۲)$$

که در آن $r_۰ = N^۲\left\{(\frac{۱}{n} - \frac{۱}{N})\sigma_{ee} + \frac{۱}{n}\sigma_{uu}\right\}$ است.

مدل‌هایی که با استفاده از مدل (۱) توصیف شدند به پارامترهای σ_{ee} و σ_{uu} بستگی دارند که آنها نامعلوم هستند، از این‌رو ساختاری از برآوردگرهای سازگار برای پارامترهای مدل به اطلاعات اضافی از جامعه نیاز دارد. اکنون دو حالت ذیل برای برآورد این پارامترها در نظر گرفته می‌شود.

حالت اول فرض می‌شود واریانس خطای اندازه‌گیری، σ_{uu}، معلوم است. همان‌گونه که در فولر (۱۹۸۷) اشاره شده است، با اندازه‌گیری‌های تکراری می‌توان مقادیر معقولی را برای σ_{uu} به دست آورد. با استفاده از روش برآوردگر گشتاوری، برآوردگر σ_{ee} به صورت $\hat{\sigma}_{ee} = s_Y^۲ - \sigma_{uu}$ حاصل می‌شود، که در آن $s_Y^۲ = \frac{۱}{(n-۱)}\sum_{i\in s}(Y_i - \bar{Y}_s)^۲$ است. حالت دوم که ممکن است در موارد خاصی روی دهد این است که نسبت $\delta = \frac{\sigma_{ee}}{\sigma_{uu}}$ معلوم باشد. در این حالت برآوردگر روش گشتاوری σ_{uu} برابر با $\hat{\sigma}_{uu} = \frac{۱}{۱+\delta}s_Y^۲$ است.

برای به دست آوردن پیشگوی بهینهٔ واریانس جامعه $(V(y) = \frac{۱}{N}\sum_{i=۱}^{N}(y_i - \bar{y})^۲)$ همانند پیشگوی میانگین جامعه عمل می‌شود و ردهٔ پیشگوهای ناااریب درجهٔ دوم ذیل در نظر گرفته می‌شود.

$$g(s, \mathbf{Y}_s) = b_s + \sum_{i\in s} b_{is}Y_i^۲ + \sum_{i\neq} \sum_{i'\in s} b_{ii's}Y_iY_{i'}$$

پیشگوی $g(s, \mathbf{Y}_s)$ برای واریانس جامعه $(V(y))$ ناااریب خواهد بود اگر و تنها اگر

$$E\big(g(s, Y_s)\big) = V(y)$$

با توجه به این‌که $s_Y^۲ = \frac{۱}{(n-۱)}\sum_{i\in s}(Y_i - \bar{Y}_s)^۲$ در شرط فوق صدق می‌کند بنابراین پیشگوی ناااریب واریانس جامعه $(V(y))$ خواهد بود.

همچنین Y_iها از یکدیگر مستقل و دارای توزیع نرمال هستند و همانند پیشگوی ناااریب با مینیمم واریانس میانگین جامعه، با استفاده از قضیهٔ برآورد (پیشگو) MVU رائو (۱۹۷۳) نتیجه می‌شود که واریانس نمونه $(s_Y^۲)$، پیشگوی ناااریب با مینیمم واریانس $V(y)$ است.

۴ پیشگویی بیز و مینیماکس

در بخش قبل در مدل (۱)، μ کمیتی ثابت در نظر گرفته شده بود. در روش بیزی μ متغیری تصادفی در نظر گرفته می‌شود و تغییرات آن توسط یک توزیع احتمال بیان می‌گردد. در این بخش پیشگوی بیز و مینیماکس مقدارکل جامعه در نمونه‌گیری تصادفی ساده و دو مرحله‌ای مورد بررسی قرار می‌گیرد.

۱.۴ پیشگویی بیز و مینیماکس مقدارکل جامعه در نمونه‌گیری تصادفی ساده

موخوپادیای (۲۰۰۰) با در نظرگرفتن مدل (۱) و نظر به این که در نمونه‌های بزرگ، توزیع اغلب متغیرهای وابسته به موضوع‌های اقتصادی و اجتماعی (حداقل به طور تقریبی) نرمال هستند، یک توزیع پیشین نرمال برای μ در نظر گرفته است.

$$\mu \sim N(\tau, \theta^2) \qquad (۳)$$

در قضیهٔ ذیل پیشگوی بیز مقدارکل جامعه تحت مدل (۱) و توزیع پیشین (۳) به دست آورده می‌شود.

قضیه۱.۴ (موخوپادیای ۲۰۰۰) تحت مدل (۱) و توزیع پیشین (۳) برای μ، اگر $\mathbf{Y}_s = (Y_1, Y_2, \ldots, Y_n)$ یک نمونهٔ تصادفی ساده به اندازهٔ n انتخاب شده از جامعه Pباشد، آنگاه پیشگوی بیز مقدارکل جامعه $(T(y) = \sum_{i=1}^{N} y_i)$ تحت تابع زیان درجهٔ دوم خطا، عبارت است از

$$\hat{T}_B = \frac{n\bar{Y}_s\sigma_{ee}}{\sigma^2} + \left[\frac{n\sigma_{uu} + (N-n)\sigma^2}{\sigma^2}.\frac{\tau\sigma^2 + n\bar{Y}_s\theta^2}{\sigma^2 + n\theta^2}\right] \qquad (۴)$$

که در آن $\sigma^2 = \sigma_{ee} + \sigma_{uu}$ و مخاطرهٔ بیز آن عبارت است از

$$r_\theta(\hat{T}_B) = n\sigma_0^2 + (N-n)\sigma_{ee} + (N - \frac{n\sigma_{ee}}{\sigma^2})^2.\frac{\theta^2\sigma^2}{\sigma^2 + n\theta^2} \qquad (۵)$$

برای به دست آوردن پیشگوی مینیماکس مقدارکل جامعه از روش بیز حدی استفاده می‌شود. چون

$$\lim_{\theta\to\infty} r_\theta(\hat{T}_B) = \frac{N(N-n)}{n}\sigma^2 + N\sigma_{uu} = r_0$$

و همچنین مخاطرهٔ پیشگوی $N\bar{Y}_s$ با توجه به رابطهٔ (۲) برابر با r_0 است لذا پیشگوی $N\bar{Y}_s$ پیشگوی مینیماکس مقدارکل جامعه است.

پیشگویی بیز واریانس جامعه در نمونه‌گیری تصادفی ساده

باتاچاریا (۱۹۹۷) پیشگوی بیز واریانس جامعه $(V(y) = \frac{۱}{N}\sum_{i=۱}^{N}(y_i - \bar{y})^۲$ را تحت مدل (۱) و با در نظرگرفتن توزیع پیشین (۳) به صورت ذیل به دست آورد.

قضیه۲.۴ (باتاچاریا ۱۹۹۷) تحت مدل (۱) و توزیع پیشین (۳) برای μ، پیشگوی بیز واریانس جامعه تحت تابع زیان درجهٔ دوم خطا عبارت است از

$$\begin{aligned}E(V(y)|Y_s) &= E\left[\left(\frac{n}{N}s_y^۲ + \frac{(N-n)}{N}s_{yr}^۲ + \frac{n(N-n)}{N^۲}(\bar{y}_r - \bar{y}_s)^۲\right)|Y_s\right] \\ &= \frac{\sigma_۰^۲}{N}(n - ۱ + ۲\lambda) + \frac{(N-n-۱)}{N}\sigma_{ee} + \frac{n(N-n)}{N^۲}\times \\ &\quad \left\{\sigma_۱^۲ + \frac{\sigma_{ee}^۲}{\sigma^۴}\left[\bar{Y}_s^۲ - ۲\bar{Y}_s\left(\frac{\tau\sigma^۲ + n\bar{Y}_s\theta^۲}{\sigma^۲ + n\theta^۲}\right) + \right.\right. \\ &\quad \left.\left.\left(\frac{\tau\sigma^۲ + n\bar{Y}_s\theta^۲}{\sigma^۲ + n\theta^۲}\right)^۲ + \frac{\theta^۲\sigma^۲}{\sigma^۲ + n\theta^۲}\right]\right\}\end{aligned}$$

که در آن

$$\lambda = \frac{(n-۱)\sigma_{ee}^۲.s_Y^۲}{\sigma^۴\sigma_۰^۲} \quad , \quad s_Y^۲ = \frac{۱}{(n-۱)}\sum_{i\in s}(Y_i - \bar{Y}_s)^۲,$$
$$\sigma_۰^۲ = \frac{\sigma_{uu}.\sigma_{ee}}{\sigma^۲} \quad , \quad \sigma^۲ = \sigma_{ee} + \sigma_{uu} \quad , \quad Y_s = (Y_۱, Y_۲, \ldots, Y_n).$$

۲.۴ پیشگویی بیز در نمونه‌گیری دومرحله‌ای

اخیراً در جامعه‌های انسانی، بررسی‌های نمونه‌ای دومرحله‌ای کاربرد فراوان پیدا کرده‌اند. بولفارین (۱۹۹۱) و موخوپادیای (۱۹۹۵) پیشگویی بیز در نمونه‌گیری دومرحله‌ای از جامعهٔ متناهی تحت مدل خطای اندازه‌گیری را مورد مطالعه قرار داده‌اند.

در این قسمت جامعهٔ متناهی P که شامل K زیرجامعهٔ (خوشهٔ) دوبه‌دو مجزا به اندازه‌های M_h، $h = ۱, ۲, \ldots, K$ است، در نظر گرفته می‌شود. همچنین y_{hj} به مشخصهٔ مورد نظر مربوط به واحد j ـ ام در خوشهٔ h ـ ام جامعه، $j = ۱, ۲, \ldots, M_h$ و $h = ۱, ۲, \ldots, K$ اشاره می‌کند. برای انتخاب نمونه از جامعهٔ متناهی P، در مرحلهٔ نخست یک نمونهٔ n خوشه‌ای از K خوشهٔ جامعه انتخاب می‌شود، این نمونهٔ n خوشه‌ای با s نمایش داده می‌شود. در مرحلهٔ دوم از هر یک از خوشه‌ها که در مرحلهٔ نخست انتخاب شده‌اند، نمونهٔ s_h ای به اندازه m_h از آن خوشه انتخاب می‌شود. پس از این که نمونه از جامعه انتخاب شد، مقدار کل جامعه را می‌توان

به صورت ذیل نوشت

$$T(y)=\sum_{h\in s}\sum_{j\in s_h}y_{hj}+\sum_{h\in s}\sum_{j\notin s_h}y_{hj}+\sum_{h\notin s}\sum_{j=1}^{M_h}y_{hj} \qquad (۶)$$

برای پیشگویی مقدار کل جامعه مانند قسمت‌های قبلی مدل ذیل در نظرگرفته می‌شود

$$\begin{aligned} y_{hj}&=\mu_h+e_{hj},\\ \mu_h&=\mu+v_h,\\ Y_{hj}&=y_{hj}+u_{hj} \end{aligned} \qquad (۷)$$

که در آن μ، میانگین مقدار واقعی مشخصهٔ مورد نظر در جامعه، و μ_h، میانگین مقدار واقعی مشخصه در خوشهٔ h ـ ام جامعه است و e_{hj}، v_h و u_{hj} دوبه‌دو از یکدیگر مستقل هستند، $e_{hj}\sim N(0,\sigma_{eeh})$، $v_h\sim N(0,\delta^2)$ و $u_{hj}\sim N(0,\sigma_{uuh})$، $h=1,2,\ldots,K$ و $j=1,2,\ldots,M_h$. در این مدل Y_{hj} به مقدار مشاهده شدهٔ مشخصهٔ مربوط به واحد j ـ ام در خوشهٔ h ـ ام اشاره می‌کند.

مقادیر نمونه که به روش دومرحله‌ای از جامعهٔ متناهی انتخاب شده است به صورت ذیل نمایش داده می‌شود

$$\mathbf{Y}_s=\{Y_{hj};j\in s_h,h\in s\}$$

قضیه۳.۴ (موخوپادیای ۱۹۹۵) تحت مدل (۷) و با توجه به این که در روش نمونه‌گیری دومرحله‌ای μ_h دارای توزیع $N(\mu,\delta^2)$ است و با فرض یک توزیع پیشین ناآگاهی‌بخش برای μ (یعنی، $P(\mu)\propto const.$)، اگر $\mathbf{Y}_s=\{Y_{hj};j\in s_h,h\in s\}$ یک نمونهٔ تصادفی باشد که به روش دومرحله‌ای از جامعهٔ متناهی انتخاب شده باشد آنگاه پیشگوی بیز مقدار کل جامعه $(T(y)=\sum_{h=1}^{K}\sum_{j=1}^{M_h}y_{hj})$ تحت تابع زیان درجهٔ دوم خطا عبارت است از

$$\begin{aligned} \hat{T}_B=&E(T|Y_s)=\sum_{h\in s}m_h\frac{\sigma_{eeh}}{\sigma_{eeh}+\sigma_{uuh}}\bar{Y}_{sh}\\ &+\sum_{h\in s}m_h\frac{\sigma_{uuh}}{\sigma_{eeh}+\sigma_{uuh}}\{\lambda_h\bar{Y}_{sh}+(1-\lambda_h)\bar{Y}_{s\cdot}\}\\ &+\sum_{h\in s}(M_h-m_h)\{\lambda_h\bar{Y}_{sh}+(1-\lambda_h)\bar{Y}_{s\cdot}\}+\bar{Y}_{s\cdot}\sum_{h\notin s}M_h \end{aligned} \qquad (۸)$$

و مخاطرهٔ بیز آن عبارت است از

$$V(T|Y_s)=\sum_{h\in s}B_h^2c_{hh}+\sum_{h\notin s}M_h^2c_{hh}+\sum_{h\in s}\sum_{q\neq h\in s}B_hB_qc_{hq}$$

$$+\sum_{h\notin s}\sum_{q\neq h\notin s} M_h M_q c_{hq} + \sum_{h\in s}\sum_{q\neq h\notin s} B_h M_q c_{hq}$$
$$+\sum_{h\in s} m_h \frac{\sigma_{uuh}.\sigma_{eeh}}{\sigma_{eeh}+\sigma_{uuh}} + \sum_{h\in s}(M_h - m_h)\sigma_{eeh} + \sum_{h\notin s} M_h \sigma_{eeh}. \quad (۹)$$

که در آن $B_h = m_h \frac{\sigma_{uuh}}{\sigma_{eeh}+\sigma_{uuh}} + (M_h - m_h)$ است.

۵ پیشگویی هزینه سالیانه خانوارها

یکی از شیوه‌های ارزیابی سطح زندگی افراد یک جامعه، دسترسی به الگوهای متفاوت هزینه و درآمد خانوارها با ویژگی‌های متفاوت اجتماعی و اقتصادی است که این هدف با اجرای طرح «آمارگیری از هزینه و درآمد خانوار» تحقق می‌یابد.

با توجه به این که نتایج آمارگیری از هزینه و درآمد خانوار می‌تواند بیان‌کنندهٔ میزان و نحوهٔ توزیع هزینه‌ها و درآمدها و چگونگی استفادهٔ خانوارها از تسهیلات زندگی در سطح ملی و منطقه‌ای باشد، لذا این طرح کاربرد بسیار وسیعی در ارزیابی عملکرد و نتایج تصمیم‌گیری‌ها و سیاست‌های گذشته و برنامه‌ریزی و سیاست‌گذاری اقتصادی، اجتماعی و فرهنگی کشور دارد.

اهمیت روزافزون کاربرد نتایج طرح هزینه و درآمد خانوار درکشور ایجاب می‌کند که مطالعاتی برای بهبود روش‌های کار در مراحل متفاوت تهیه، اجرا و استخراج نتایج این طرح انجام شود و در صورت لزوم برای افزایش درستی و دقت نتایج، برخی از این روش‌ها مورد تجدید نظر واقع شوند.

با توجه به اهمیت موضوع به ویژه در مرکز آمار ایران که متولی تولید آمار رسمی کشور است، مدیریت پژوهش‌های آماری مرکز آمار ایران در سال ۱۳۷۷ یک طرح تحقیقاتی تحت عنوان «مقایسهٔ دو شیوهٔ اطلاع‌گیری از هزینه و درآمد خانوار» انجام داده است. در این طرح تحقیقاتی از تعدادی از خانوارهای معمولی ساکن شهر تهران، اطلاعات یکسانی به دو روش، روش مصاحبهٔ حضوری (روش فعلی اجرای طرح) و روش ثبت اطلاعات توسط خانوارکه انتظار می‌رود دارای دقت و درستی بیشتری باشد، در دو مرحله گردآوری شده است. با مقایسهٔ نتایج حاصل از این دو مرحله و انجام تجزیه و تحلیل‌های لازم، روش فعلی گردآوری اطلاعات طرح «هزینه و درآمد خانوار» مورد ارزیابی قرار گرفته است.

در این بخش با استفاده از اطلاعات طرح تحقیقاتی فوق‌الذکر، مدلی برای خطای اندازه‌گیری هزینهٔ خانوارها در نظر گرفته و تحت این مدل میانگین هزینهٔ سالیانه خانوارها پیشگویی می‌شوند. لازم به ذکر است که در آمارگیری جاری هزینه و درآمد خانوار، برآوردها با فرض این که خطای اندازه‌گیری صفر است، محاسبه می‌شوند.

مشخصات کلی

هدف اصلی این بخش پیشگویی کلاسیک میانگین هزینه سالیانه خانوارهای معمولی ساکن شهر تهران تحت مدل خطای اندازه‌گیری و مقایسه آن با برآوردهای شیوه جاری است. جامعه آماری کلیه خانوارهای معمولی ساکن شهر تهران است و هر یک از این خانوارها به عنوان یک واحد آماری هستند. چارچوب آماری مورد استفاده، فهرست کلیه بلوکهای شهر تهران در سرشماری عمومی نفوس و مسکن سال ۱۳۷۵ به انضمام تعداد خانوار هر یک از این بلوکهاست و زمان آمارگیری برای سؤال‌های متفاوت، بهمن ماه سال ۱۳۷۷ و سال مورد نظر (از اول اسفند ماه ۱۳۷۶ لغایت روز سی‌ام بهمن ماه ۱۳۷۷) است.

روش نمونه‌گیری و اندازهٔ نمونه

در این طرح، روش نمونه‌گیری دومرحله‌ای است. مرحلهٔ اول مربوط به انتخاب بلوک‌های نمونه و مرحلهٔ دوم مربوط به انتخاب خانوارهای نمونه در هر یک از این بلوکهاست. با توجه به محدودیت‌های اجرایی، امکانات طرح و اهداف مورد نظر، تعداد خانوار نمونهٔ این طرح ۵۰۰ خانوار در نظر گرفته شده است. در مرحلهٔ اول نمونه‌گیری، ۵۰ بلوک انتخاب شده است و در مرحلهٔ دوم نمونه‌گیری از هر یک از بلوک‌ها که در مرحلهٔ اول انتخاب شده‌اند، ۱۰ خانوار انتخاب می‌شود.

یادآور می‌شود همان‌طور که در فوق مطرح شد، هدف اصلی پیشگویی میانگین هزینهٔ خانوارهای معمولی ساکن شهر تهران است. ولیکن برای پیشگویی پارامترها در رهیافت مبتنی بر مدل، پارامترها با توجه به مدل (۱) پیشگویی می‌شوند که در این مدل فرض شده است مانده‌ها (متغیرها) مستقل و دارای توزیع یکسان نرمال هستند، لذا باید ابتدا این فرض‌ها آزمون شوند. با آزمودن مانده‌ها (متغیرها) مشخص شد که فرض نرمال بودن برای آنها صادق نیست ولیکن با تغییر متغیر ریشهٔ چهارم هزینهٔ خوراکی و لگاریتم هزینهٔ غیرخوراکی سالیانهٔ خانوارها فرض نرمال بودن صادق شد. در این قسمت به همین دلیل ابتدا برآوردها و پیشگوهای میانگین ریشهٔ چهارم هزینهٔ خوراکی سالیانهٔ خانوارها و لگاریتم هزینهٔ غیرخوراکی سالیانه خانوارها محاسبه می‌شوند و سپس با تبدیل عکس، میانگین هزینهٔ خوراکی سالیانهٔ خانوارها و میانگین هزینهٔ غیرخوراکی سالیانهٔ خانوارها پیشگویی خواهند شد.

مقادیر مربوط به برآورد (بدون خطای اندازه‌گیری) میانگین هزینه‌های خوراکی و غیرخوراکی به همراه خطای استاندارد آنها و همچنین پیشگوی کلاسیک (با در نظر گرفتن خطای اندازه‌گیری) میانگین هزینه‌های خوراکی و غیرخوراکی به همراه خطای استاندارد آنها در جداول ذیل آمده‌اند. برای محاسبهٔ مقادیر این جداول از روابط بخش ۳ استفاده شده است.

مقادیر مربوط به هزینهٔ خوراکی سالیانهٔ خانوارها

روش	میانگین	برآورد واریانس	خطای استاندارد
برآورد معمولی (بدون خطای اندازه‌گیری)	۶۵۰۸۴۲۷٫۶۲۴	۹٫۰۴۵۳۳۴۲۸۳×۱۰^۱۰	۳۰۰۷۵۴٫۶۲۲۳
پیشگوی کلاسیک (با خطای اندازه‌گیری)	۶۸۹۳۴۶۷٫۸۴۶	۲٫۸۶۹۴۷۳۵۶۴×۱۰^۱۰	۱۶۹۳۹۵٫۲۰۵۵

مقادیر مربوط به هزینهٔ غیرخوراکی سالیانهٔ خانوارها

روش	میانگین	برآورد واریانس	خطای استاندارد
برآورد معمولی (بدون خطای اندازه‌گیری)	۱۵۸۹۴۷۳۴٫۶	۷٫۵۳۶۴۰۹۷۸۳×۱۰^۱۱	۸۶۸۱۲۴٫۹۷۸۵
پیشگوی کلاسیک (با خطای اندازه‌گیری)	۱۷۱۳۹۴۵۸٫۶۳	۵٫۳۸۹۷۰۴۸۶۴×۱۰^۱۰	۲۳۲۱۵۷٫۳۷۹

همان‌طورکه از جدول‌های فوق ملاحظه می‌شود برآورد واریانس برآوردگر میانگین هزینهٔ سالیانهٔ خانوارها بدون در نظرگرفتن خطای اندازه‌گیری از برآورد واریانس پیشگوی میانگین هزینهٔ سالیانهٔ خانوارها با در نظرگرفتن خطای اندازه‌گیری بیشتر است. این به دلیل استفاده از نمونه‌ای است که به شیوهٔ نمونه‌گیری دومرحله‌ای از جامعه انتخاب شده است و با ملاحظهٔ فرمول برآورد واریانس برآوردگر میانگین جامعه مشخص می‌شود که آن تابعی از واریانس درون بلوک‌ها و واریانس بین بلوک‌هاست و به دلیل ناهمگنی بین بلوک‌ها، برآورد واریانس برآوردگر میانگین جامعه نسبت به برآورد واریانس پیشگوی میانگین جامعه بزرگتر شده است.

در انتها پیشنهاد می‌شود در مواردی که خطای اندازه‌گیری مشخصه‌ها قابل اغماض نیستند برای دسترسی به اطلاعات واقعی‌تر از جامعهٔ مورد نظر با در نظرگرفتن یک مدل خطای اندازه‌گیری، پارامترهای جامعه با استفاده از این رهیافت پیشگویی شوند.

مراجع

[۱] گزارش طرح تحقیقاتی مقایسه دو شیوه اطلاع‌گیری از هزینه و درآمد خانوار (۱۳۷۸)، مدیریت پژوهش‌های آماری، مرکز آمار ایران.

[2] Bhattacharyya, S. (1997). Some studies on estimation of mean and variances in finite population sampling. Unpublished Ph.D. Thesis submitted to Indian Statistical Institute, Calcutta.

[3] Bolfarine, H. (1991). Finite-population prediction under error-in-variables super-population models. Canadian Journal of Statistics, 19(2), 191-207.

[4] Bolfarine, H., and Zacks, S. (1991). Bayes and minimax prediction in finite population. Journal of Statistical Planning and Inference, 28, 139-151.

[5] Fuller, W. A. (1987). Measurement Error Models. Wiley, New York.

[6] Fuller, W. A. (1989). Prediction of the true values for the measurement error models, In P. J. Brown and W. A. Fuller, eds., Statistical Analysis of Measurement Error Models and Applications Am. Math. Soc., Providence, RI.

[7] Mukhopadhyay, P. (1992). On prediction in finite population under error-in-variables superpopulation models., Indian Statistical Institute Technical Report No. ASC/92/11.

[8] Mukhopadhyay, P. (1994). Prediction in finite population under error-in-variables superpopulation models. Journal of Statistical Planning and Inference, 41, 151-161.

[9] Mukhopadhyay, P. (1995). Bayes and minimax estimator for two-stage sampling from a finite population under measurement error models. Communications in statistics, Theory and Methods, 24(3),663-674.

[10] Mukhopadhyay, P. (2000). Topics in Survey Sampling. Springer-Verlag, New York.

[11] Rao, C. R. (1973). Linear Statistical Inference. Second edition, John Wiley and Sons, New York.

[12] Sprent, P. (1966). A generalized least squares approach to linear function relationships. J. Roy. Statist. Soc. Ser. B, 28, 279-297.

استفاده از برآوردگرهای عادی و رگرسیونی نمونه‌گیری مجموعه رتبه‌دار در برآورد مقدار کل محصول گندم ایران

نادر نعمت‌الهی[۱]، لطیف سعادتی[۲]

[۱] گروه آمار دانشگاه علامه طباطبایی
[۲] کارشناسی ارشد آمار دانشگاه علامه طباطبایی

چکیده: در بعضی از جوامع مورد بررسی ساختار جامعه معلوم نیست و معمولاً از روش نمونه‌گیری تصادفی ساده برای برآورد میانگین متغیر مورد نظر استفاده می‌شود. برآوردگر حاصل از روش نمونه‌گیری تصادفی ساده به علت اینکه کنترلی روی واحدهای انتخاب شده نیست از درجه اعتماد بالایی برخوردار نمی‌باشد. در این مواقع می‌توان از روش نمونه‌گیری مجموعه رتبه‌دار استفاده کرد. روش‌های نمونه‌گیری مجموعه رتبه‌دار یک برآوردگر نااریب برای میانگین جامعه مورد بررسی بدست می‌دهد که این برآوردگر کاراتر از برآوردگر نمونه‌گیری تصادفی ساده با اندازه نمونه برابر می‌باشد.

در این مقاله برآوردگرهای عادی و رگرسیونی روش‌های نمونه‌گیری مجموعه رتبه‌دار ساده، میانه‌ای و حدی مورد بررسی قرار می گیرند. سپس مقدار کل محصول گندم ایران را توسط این روشها برآورد کرده و نشان می‌دهیم که این برآوردگرها نسبت به برآوردگر تصادفی ساده از دقت بیشتری برخوردار می‌باشند.

واژه‌های کلیدی: آمارهٔ مرتب، نمونه‌گیری مجموعهٔ رتبه‌دار، برآوردگر رگرسیونی

۱ مقدمه

روش‌های نمونه‌گیری مجموعه رتبه‌دار یک برآوردگر نااریب برای میانگین جامعه مورد بررسی بدست می‌دهند که این برآوردگر دارای واریانس کوچکتر از برآوردگر نمونه تصادفی ساده با اندازه نمونه برابر می‌باشد. علاوه بر آن در بسیاری از کارهای عملی اندازه‌گیری تک تک واحدهای نمونه مشکل یا پرهزینه می‌باشد، اما واحدها می‌توانند به آسانی و بی هیچ هزینه‌ای رتبه‌بندی شوند و واحدهای رتبه‌بندی شدهٔ خاصی اندازه‌گیری گردند. در مسائل زیست محیطی و کشاورزی رتبه‌بندی واحدها بدون اینکه مقدار واقعی آنها اندازه‌گیری شود امکان‌پذیر است. در بعضی موارد (مثل اندازه‌گیری میزان آلودگی خاک یک منطقه) بخش اعظم هزینه مربوط به تجزیه و تحلیل و کارهای آزمایشگاهی میباشد، درحالیکه تشخیص قابلیت (پتانسیل) واحدهای نمونه کار ساده‌ای است. در این حالت می‌توانیم از روش نمونه‌گیری مجموعه رتبه‌دار استفاده کنیم، یعنی می‌توانیم

تعداد زیادی از واحدهای جامعه را انتخاب کنیم و فقط زیر نمونه‌ای از واحدهای انتخاب شده را از نظر کمیت مورد نظر، اندازه‌گیری کنیم. همچنین استفاده از روش نمونه‌گیری مجموعه رتبه‌دار باعث طبقه‌بندی جامعه به طبقات همگن می‌شود. پس با روش نمونه‌گیری مجموعه رتبه‌دار علاوه بر افزایش دقت و کارایی می‌توان جامعه را بدون اینکه از ساختار جامعه اطلاعی در دست باشد، طبقه‌بندی کرد (و از تمام طبقات نمونه انتخاب کرد) که همین امر باعث افزایش دقت برآوردگر میانگین جامعه می‌شود.

روش نمونه‌گیری مجموعه رتبه‌دار اولین بار توسط مک‌انتایر[1] (۱۹۵۲) ارائه شد. در سال ۱۹۶۷ تاکاهاسی[2] و واکیماتو[3] نشان دادند که میانگین حسابی حاصل از روش نمونه‌گیری مجموعه رتبه‌دار یک برآوردگر ناااریب برای میانگین جامعه و کاراتر از روش نمونه‌گیری تصادفی ساده با اندازه نمونه برابر می‌باشد. در سال ۱۹۷۷ استوکس[4] کاربرد متغیر همراه در روش نمونه‌گیری مجموعه رتبه‌دار را مطرح نمود. در سال ۱۹۹۳ پتیل[5]، سینها[6] و تایلی[7] برآوردگر رگرسیونی نمونه‌گیری تصادفی ساده براساس متغیر همراه را با برآوردگر عادی نمونه‌گیری مجموعه رتبه‌دار مورد مقایسه قرار دادند. در سال ۱۹۹۷ یو[8] و لام[9] برآوردگر رگرسیونی براساس نمونه‌گیری مجموعه رتبه‌دار را در نظر گرفته و آن را با برآوردگر عادی نمونه‌گیری مجموعه رتبه‌دار و برآوردگر رگرسیونی نمونه‌گیری تصادفی ساده مقایسه نمودند. در سال ۱۹۹۶ سماوی[10]، ابودایه[11] و احمد[12] روش نمونه‌گیری مجموعه رتبه‌دار حدی را معرفی کردند. در سال ۱۹۹۷ و ۱۹۹۸ مطلق[13] روش نمونه‌گیری مجموعه رتبه‌دار میانه‌ای را با و بدون استفاده از متغیر همراه ارائه کرد و در سال ۲۰۰۱ برآوردگرهای رگرسیونی چندین روش نمونه‌گیری مجموعه رتبه‌دار را برای برآورد میانگین جامعه بکار برد و آنها را با برآوردگرهای عادی روش نمونه‌گیری مجموعه رتبه‌دار مورد مقایسه قرار داد.

در این مقاله ضمن بیان روش‌های نمونه‌گیری مجموعه رتبه‌دار ساده، میانه‌ای و حدی و ارایه برآوردگرهای عادی و رگرسیونی در این روشها، مقدار کل محصول گندم ایران را با استفاده از آنها برآورد و با یکدیگر مقایسه می‌کنیم. بر این اساس در بخش ۲ برآوردگرهای عادی و در بخش ۳ برآوردگرهای رگرسیونی را در روش‌های نمونه‌گیری مجموعه رتبه‌دار ساده، میانه‌ای و حدی را مورد بررسی قرار داده و خواص آنها را بیان می‌کنیم (برای اثبات این خواص به مراجع فوق مراجعه شود). در بخش ۴ مقدار کل محصول گندم ایران را با استفاده از برآوردگرهای فوق برآورد و با یکدیگر مقایسه می‌کنیم.

---

1) McIntyre 2) Takahasi 3) Wakimoto 4) Stokes 5) Patil 6) Sinha 7) Taili 8) Yu 9) Lam 10) Samawi 11) Abu Dayyeh 12) Ahmad 13) Muttlak

۲ برآوردگرهای عادی در روش‌های نمونه‌گیری مجموعه رتبه‌دار

۱.۲ روش نمونه‌گیری مجموعه رتبه‌دار[14] ساده

در روش نمونه‌گیری مجموعه رتبه‌دار ساده برای انتخاب یک نمونه nتایی، ابتدا یک نمونه $n^۲$ انتخاب شده و به n زیر مجموعه nتایی تقسیم می‌شوند و اعضای هر مجموعه بر اساس صفت مورد نظر رتبه‌بندی می‌شوند. سپس از نمونه nتایی اول واحد دارای کوچکترین رتبه، از نمونه nتایی دوم واحد دارای دومین رتبه، ... و از نمونه nتایی nام واحد دارای بزرگترین رتبه اندازه‌گیری می‌شود.

فرض کنید $n^۲$ ،$Y_{۱۱}, Y_{۱۲}, \ldots, Y_{۱n}; Y_{۲۱}, Y_{۲۲}, \ldots, Y_{۲n}; \ldots; Y_{n۱}, Y_{n۲}, \ldots, Y_{nn}$ واحد انتخاب شده از جامعه باشند و فرض می‌شود که متغیرهای تصادفی مستقل با تابع توزیع یکسان $F(Y)$ و تابع چگالی یکسان $f(y)$ بوده و $Y_{i[۱]}, Y_{i[۲]}, \ldots, Y_{i[n]}$ آماره‌های مرتب نمونه $Y_{i۱}, Y_{i۲}, \ldots, Y_{in}$ برای $i = ۱, ۲, \ldots, n$ باشند. نمونه اندازه‌گیری شده nتایی از این $n^۲$ واحد انتخابی عبارتند از $Y_{۱[۱]}, Y_{۲[۲]}, \ldots, Y_{n[n]}$ که متغیرهای تصادفی مستقل می‌باشند. میانگین واحدهای اندازه‌گیری شده عبارت است از

$$\overline{Y}_{RSS} = \overline{Y}_{[n]} = \frac{۱}{n} \sum_{i=۱}^{n} Y_{i[i]} \qquad (۱)$$

هنگامیکه تعداد نمونه زیاد باشد n (بزرگتر از ۳ یا ۴) رتبه‌بندی عناصر به راحتی امکان‌پذیر نیست. در چنین حالتی برای رتبه‌بندی راحتتر و ایجاد خطای کمتر، عموماً روش نمونه‌گیری مجموعهٔ رتبه‌دار با اندازه نمونه کوچکتر انجام گرفته و سپس روش نمونه‌گیری چند بار تکرار می‌شود. مثلاً برای انتخاب یک نمونه $n_۰$تایی که $n_۰ = nm$، روش نمونه‌گیری مجموعهٔ رتبه‌دار nتایی، m بار تکرار می‌شود، که در آن $n_۰$ تعداد کل نمونه انتخابی مورد نظر می‌باشد. اگر m تکرار وجود داشته باشد در این صورت m مشاهد برای $\overline{Y}$ بدست می‌آید و میانگین کل بصورت زیر خواهد بود.

$$\overline{Y}_{RSS} = \overline{Y}_{[n]} = \frac{۱}{nm} \sum_{i=۱}^{n} \sum_{j=۱}^{m} Y_{i[i]j} \qquad (۲)$$

که به آن برآوردگر روش نمونه‌گیری مجموعه رتبه‌دار ساده گویند. خواص این برآوردگر در قضیه زیر بیان شده است.

قضیه ۱.۲ برآوردگر $\overline{Y}_{RSS}$ یک برآوردگر نااریب برای میانگین جامعه است و واریانس آن با m بار تکرار عبارت است از

$$\sigma^۲(\overline{Y}_{[n]m}) = Var(\overline{Y}_{[n]m} = \frac{\sigma^۲_{[n]}}{nm} = \frac{\sigma^۲}{nm} - \frac{۱}{n^۲ m} \sum_{k=۱}^{n} (\mu_{n,k[k]} - \mu)^۲$$

14) Ranked Set Sampling

و اگر $\overline{Y}_{SRS}$ میانگین نمونه تصادفی ساده nتایی باشد آنگاه همواره $Var(\overline{Y}_{RSS}) \leq Var(\overline{Y}_{SRS})$ است.

دلیل اینکه میانگین مجموعه رتبه‌دار با میانگین نمونه تصادفی ساده حاصل از یک نمونه nتایی و نه n^2تایی مقایسه می‌شود این است که در وضعیت مورد مطالعه ما، هزینه رتبه‌بندی عناصر انتخابی مورد نظر نیست و تنها هزینه عناصر اندازه‌گیری شده مورد نظر است.

۲.۲ روش نمونه‌گیری مجموعهٔ رتبه‌دار میانه‌ای

در روش نمونه‌گیری مجموعهٔ رتبه‌دار میانه‌ای n نمونه nتایی (نمونه n^2تایی) به‌طور تصادفی از جامعه انتخاب می‌شود. واحدهای داخل هر نمونه بر اساس متغیر مورد نظر رتبه‌بندی می‌شود. اگر تعداد نمونه فرد باشد، از هر نمونه nتایی رتبه‌بندی شده، $\frac{n+1}{2}$امین واحد (میانه نمونه) انتخاب می‌شود. اگر اندازه نمونه زوج باشد از $\frac{n}{2}$ نمونه اول $\frac{n}{2}$امین واحد رتبه‌بندی شده و از $\frac{n}{2}$ نمونه بعدی $\frac{n+2}{2}$امین واحد رتبه‌بندی شده، انتخاب می‌شود.

فرض کنید $Y_{11}, Y_{12}, \ldots, Y_{1n}; Y_{21}, Y_{22}, \ldots, Y_{2n}; \ldots; Y_{n1}, Y_{n2}, \ldots, Y_{nn}$ متغیرهای تصادفی مستقل باشند که دارای تابع توزیع $F(y)$ و تابع چگالی $f(y)$ باشند و $Y_{i(1)}, Y_{i(2)}, \ldots, Y_{i(n)}$، $Y_{i1}, Y_{i2}, \ldots, Y_{in}$ آماره های مرتب $i = 1, 2, \ldots, n$ باشند. اگر تعداد نمونه فرد باشد از هر نمونه $Y_{i[\frac{n+1}{2}]}$ یعنی میانهٔ iامین نمونه $(i = 1, 2, \ldots, n)$ انتخاب می‌شود، به عبارت دیگر $\frac{n+1}{2}$امین آماره مرتب از iامین نمونه انتخاب می‌شود. اگر اندازه نمونه زوج باشد از $\frac{n}{2}$ نمونه‌ها، $\frac{n}{2}$امین آماره مرتب یعنی $Y_{i[\frac{n}{2}]}$ به ازای $(i = 1, 2, \ldots, \frac{n}{2})$ و از $\frac{n}{2}$ نمونه بعدی $\frac{n+2}{2}$امین آماره مرتب یعنی $Y_{i[\frac{n+2}{2}]}$، به ازای $i = \frac{n}{2} + 1, \ldots, n$ انتخاب می‌شود. به عبارت دیگر $Y_{i[\frac{n}{2}]}$ و $i = 1, 2, \ldots, \frac{n}{2}$, $Y_{i[\frac{n+2}{2}]}$, $i = \frac{n}{2} + 1, \ldots, n$ نشان دهنده نمونه‌های اندازه‌گیری شدهٔ روش نمونه‌گیری مجموعهٔ رتبه‌دار میانه‌ای خواهند بود.

اگر تعداد نمونه فرد باشد میانگین مجموعهٔ رتبه‌دار میانه‌ای به صورت زیر خواهد بود

$$\overline{Y}_{MRSS1} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{i[\frac{n+1}{2}]} \qquad (۳)$$

و در حالتی که تعداد نمونه زوج باشد میانگین عبارت است از

$$\overline{Y}_{MRSS2} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n}{2}} Y_{i[\frac{n}{2}]} + \sum_{i=\frac{n}{2}+1}^{n} Y_{i[\frac{n+1}{2}]} \right) \qquad (۴)$$

در قضیه زیر خواص این دو برآوردگر بیان شده‌اند.

قضیه ۲.۲ اگر $\overline{Y}_{MRSS۱}$ $\overline{Y}_{MRSS۲}$ میانگینهای مجموعهٔ رتبه‌دار میانه‌ای باشند آنگاه
الف) $\overline{Y}_{MRSS۱}$ و $\overline{Y}_{MRSS۲}$ یک برآوردگر نااریب برای میانگین جامعه (μ) می‌باشند اگر توزیع جامعه حول μ متقارن باشد.
ب) $Var(\overline{Y}_{MRSS۱})$ و $Var(\overline{Y}_{MRSS۱})$ از واریانس نمونه‌گیری تصادفی ساده $Var(\overline{Y})$ کوچکتر است.
ج) اگر توزیع حول میانگین μ متقارن نباشد، میانگین مربع خطا (MSE)، $\overline{Y}_{MRSS۱}$ و $\overline{Y}_{MRSS۲}$ از واریانس نمونه‌گیری تصادف ($Var(\overline{Y}_{srs})$) کوچکتر خواهد بود.

۳.۲ روش نمونه‌گیری مجموعهٔ رتبه‌دار حدی

روش نمونه‌گیری مجموعهٔ رتبه‌دار حدی به این صورت است که برای انتخاب یک نمونه nتایی، n نمونهٔ nتایی (نمونه $n^۲$تایی) از جامعه‌ای که باید میانگین آن برآورد شود، انتخاب می‌شود. فرض می‌شود که بزرگترین و کوچکترین واحدهای این مجموعه می‌تواند با یک نگاه اجمالی تعیین شوند. این روش ساده و یک فرآیند عملی می‌باشد. از اولین مجموع nتایی کوچکترین آماره مرتب واحدها، از دومین مجموع nتایی دوم بزرگترین واحدها رتبه‌بندی شده و از مجموع nتایی سوم، کوچکترین واحد رتبه‌بندی شده و به همین ترتیب از $(n-۱)$امین مجموعهٔ بزرگترین واحد رتبه‌بندی شده اندازه‌گیری می‌شوند. برای انتخاب nامین واحد از nامین مجموعهٔ (مجموعهٔ آخری) بسته به این که تعداد نمونه فرد باشد یا زوج به صورت زیر انتخاب می‌شود.
الف) اگر تعداد نمونه n زوج باشد، بزرگترین واحد رتبه‌بندی شده، اندازه‌گیری می‌شود. این روش نمونه‌گیری را با $ERSS_a$ نشان می‌دهند.
اگر تعداد نمونه فرد باشد به دو صورت زیر می‌توان نمونه را انتخاب کرد.
ب) برای اندازه‌گیری nامین واحد، میانگین اندازه‌گیری شده کوچکترین و بزرگترین واحد nامین مجموعهٔ را در نظر می‌گیرند. این روش نمونه‌گیری را با $ERSS_b$ نشان می‌دهند.
ج) برای اندازه‌گیری nامین واحد، مقدار میانه nامین مجموعهٔ در نظر گرفته می‌شود. این روش نمونه‌گیری با $ERSS_c$ نشان می‌دهند.
برآوردگرهای حاصل از روش نمونه‌گیری مجموعه رتبه‌دار حدی به صورت زیر خواهند بود.

$$\overline{Y}_{ERSS_a} = \overline{Y}_a = \frac{۱}{۲}\left(\frac{۱}{۲}\sum_{i=۱}^{\frac{n}{۲}} Y_{۲i-۱[۱]} + \frac{۱}{۲}\sum_{i=۱}^{\frac{n}{۲}} Y_{۲i[n]}\right) \qquad (۵)$$

$$\overline{Y}_{ERSS_b} = \overline{Y}_b = \frac{Y_{۱[۱]} + Y_{۲[۲]} + \cdots + Y_{n-۱[n]} + \frac{۱}{۲}[Y_{n[۱]} + Y_{n[n]}]}{n} \qquad (۶)$$

$$\overline{Y}_{ERSS_c} = \overline{Y}_c = \frac{Y_{۱[۱]} + Y_{۲[۲]} + \cdots + Y_{n-۱[n]} + Y_{n[\frac{n+۱}{۲}]}}{n} \qquad (۷)$$

قضیه ۳.۲ وقتی تعداد نمونه n زوج و توزیع جامعه متقارن باشد، $\overline{Y}_{ERSS_a}$ یک برآوردگر ناااریب برای میانگین جامعه با واریانس زیر است.

$$Var(\overline{Y}_{ERSS_a}) = \frac{۱}{۲n}(\sigma^{۲}_{n,۱[۱]} + \sigma^{۲}_{n,n[n]})$$

که در آن $\sigma^{۲}_{n,n[n]} = Var(\overline{Y}_{n[n]})$ و $\sigma^{۲}_{n,۱[۱]} = Var(\overline{Y}_{۱[۱]})$.

همچنین اگر تعداد نمونه فرد باشد $ERSS_b$ و $ERSS_c$ یک برآوردگر ناااریب برای میانگین جامعه است و واریانس این برآوردگرهای به ترتیب عبارت است از

$$Var(\overline{Y}_{ERSSb}) = \frac{۱}{۴n^{۲}}[(۲n-۱)(\sigma^{۲}_{n,۱[۱]} + \sigma^{۲}_{n,n[n])} + ۲\sigma_{(۱,n)}] \qquad (۸)$$

$$Var(\overline{Y}_{ERSSc}) = \frac{n-۱}{۲n^{۲}}[(\sigma^{۲}_{n,۱[۱]} + \sigma^{۲}_{n,n[n])}] + \frac{۱}{n^{۲}}\sigma^{۲}_{n,n[\frac{n+۱}{۲}]} \qquad (۹)$$

که $\sigma_{(۱,n)} = Cov(Y_{n[۱]}, Y_{n[n]})$ است.

۳ برآوردگرهای رگرسیونی در روش‌های نمونه‌گیری مجموعه رتبه‌دار

هنگامیکه رتبه‌بندی متغیر اصلی Y واحدهای نمونه مشکل یا پر هزینه است اما یک متغیر همراه مانند X وجود دارد که با متغیر اصلی رابطه قویی دارد و به آسانی قابل رتبه‌بندی است، می‌توان از متغیر همراه X برای رتبه‌بندی استفاده کرد. در این روش همه عناصر انتخاب شده بر اساس متغیر همراه رتبه‌بندی می‌شوند و فقط متغیر اصلی Y متناظر با بعضی از آنها اندازه‌گیری خواهند شد. عناصر متناظر اندازه‌گیری شده، نمونه نهایی را تشکیل می‌دهند. حال به وسیله این عناصر اندازه‌گیری شده می‌توان برآوردگر عادی و یا رگرسیونی را بدست آورد که در این مقاله تنها به بررسی برآوردگرهای رگرسیونی در دو حالت معلوم یا نامعلوم بودن میانگین جامعه متغیر همراه خواهیم پرداخت.

۱.۳ برآوردگرهای رگرسیونی در نمونه‌گیری مجموعه رتبه‌دار با معلوم بودن میانگین جامعه متغیر همراه

برای انتخاب یک نمونه nتایی از متغیر اصل Y با استفاده از متغیر همراه X به صورت زیر عمل می‌کنیم. n نمونه دو متغیره (زوجی) nتایی (نمونه $n^{۲}$تایی) از جامعه مورد مطالعه انتخاب شده، نمونه‌های زوجی بر اساس متغیر همراه X رتبه‌بندی می‌شوند. سپس بوسیله یکی

از روش‌های نمونه‌گیری مجموعه رتبه‌دار متغیر اصلی متناظر با متغیر همراه انتخاب و اندازه‌گیری می‌شوند. این عمل m بار تکرار می‌گردد تا یک نمونه nmتایی حاصل گردد.

فرض می‌شود که رگرسیون Y روی X خطی بوده $\frac{X-\mu_X}{\sigma_X}$ و $\frac{Y-\mu_Y}{\sigma_Y}$ و دارای توزیع یکسان باشند. برای مثال هر دو دارای توزیع نرمال دو متغیر باشند. در این صورت رگرسیون Y روی X بصورت

$$Y = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(X - \mu_X) + \epsilon \qquad (۱۰)$$

است که در آن X و ϵ مستقل ϵ دارای میانگین صفر و واریانس $\sigma_Y^۲(۱ - \rho^۲)$ میباشد.

فرض کنید $X_{k[k]j}$ و $Y_{k(k)j}$ نمونه‌های انتخابی مجموعه رتبه‌دار ساده با استفاده از متغیر همراه باشند که در آن $X_{k[k]j}$، kامین واحد رتبه‌بندی شده متغیر همراه در kامین نمونه و در jامین تکرار و $Y_{k(k)j}$ متغیر اصلی اندازه‌گیری شده متناظر با آن است. همچنین فرض کنید رگرسیون Y روی X خطی باشد. در این صورت رابطه (۱۰) را می‌توان بصورت زیر نوشت

$$Y_{k(k)j} = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(X_{k[k]j} - \mu_X) + \epsilon_{kj} \quad k = ۱, ۲, \ldots, n \; j = ۱, ۲, \ldots, m \qquad (۱۱)$$

میانگین نمونه متغیرهای اصلی انتخابی با nm تکرار برابر خواهد بود با $\overline{Y}_{RSSC} = \frac{۱}{nm}\sum_{i=۱}^{n}\sum_{j=۱}^{m} Y_{k(k)j}$.

هنگامیکه میانگین متغیر همراه μ_X معلوم باشد، برآوردگر تفاضلی میانگین جامعه متغیر اصلی Y با به کارگیری روش مجموعه رتبه‌دار (ساده) بصورت زیر خواهد بود

$$\overline{Y}_{RSSD} = \overline{Y}_{RSSC} + \beta_۱(\mu_X - \overline{X}_{RSS}) \qquad (۱۲)$$

که در آن $\beta_۱$ نامعلوم است و یک برآوردگر برای آن عبارت است از

$$\hat{\beta}_۱ = \frac{\sum_{k=۱}^{n}\sum_{j=۱}^{n}(X_{k[k]j} - \overline{X}_{RSS})(Y_{k(k)j} - \overline{Y}_{RSSC})}{\sum_{k=۱}^{n}\sum_{j=۱}^{n}(X_{k[k]j} - \overline{X}_{RSS})^۲} \qquad (۱۳)$$

در نتیجه برآوردگر رگرسیونی در نمونه‌گیری مجموعه رتبه‌دار ساده برای میانگین جامعه متغیر اصلی (μ_Y) به صورت زیر خواهد بود

$$\overline{Y}_{RSSLR} = \overline{Y}_{RSSC} + \hat{\beta}_۱(\mu_X - \overline{X}_{RSS}) \qquad (۱۴)$$

که خواص آن در قضیه زیر بیان شده است.

قضیه ۱.۳ اگر رگرسیون Y روی X خطی و بفرم(۱۱) باشد آنگاه برآوردگر رگرسیونی در نمونه‌گیری مجموعه رتبه‌دار (ساده) یک برآوردگر نااریب برای میانگین جامعه است و واریانس آن عبارت است از

$$Var(\overline{Y}_{RSSLR}) = \frac{\sigma_y^۲}{nm}(۱ - \rho^۲)[۱ + E(\frac{\overline{Z^۲}_{RSS}}{s_{z_۱}^۲})]$$

که در آن $Z_{k[k]j} = \frac{X_{k[k]j}-\mu_x}{\sigma_x}$ ، $\overline{Z}_{RSS} = \frac{۱}{nm}\sum_{k=۱}^{n}\sum_{j=۱}^{n} Z_{k[k]j}$ و

$$S^{۲}_{z_۱} = \frac{۱}{nm}\sum_{k=۱}^{n}\sum_{j=۱}^{n}\left((Z_{k[k]j} - \overline{Z}_{RSS}\right)^{۲}$$

حال فرض کنید $X_{k[m]j}$ و $Y_{k(m)j}$ به ترتیب نشان دهندهٔ میانه X و متغیر اصلی متناظر بدست آمده از kامین نمونه $k = ۱, ۲, \ldots, n$ در jامین تکرار در صورت فرد بودن اندازه نمونه باشد و همچنین $\frac{n}{۲}$امین آماره مرتب از kامین نمونه به ازای $k = ۱, \ldots, \frac{n}{۲}$ و $\frac{n+۲}{۲}$امین آماره مرتب از kامین نمونه به ازای $k = \frac{n}{۲} + ۱, \ldots, n$ در jامین تکرار در صورت زوج بودن اندازه نمونه باشد. اگر رگرسیون Y روی X خطی باشد در این صورت رگرسیون خطی $Y_{k(m)j}$ روی $X_{k[m]j}$ با توجه به رابطه (۱۰) عبارت است از

$$\begin{aligned} Y_{k(m)j} &= \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(\mu_X - X_{k[m]j}) + \epsilon_{kj} \\ k &= ۱, ۲, \ldots, n \qquad\qquad j = ۱, ۲, \ldots, m \end{aligned} \tag{۱۵}$$

اگر میانگین متغیر همراه معلوم باشد، برآوردگر تفاضلی میانگین جامعه متغیر اصلی Y با به کارگیری روش نمونه‌گیری مجموعه رتبه‌دار میانه‌ای به صورت زیر خواهد بود

$$\overline{Y}_{MRSSD} = \overline{Y}_{MRSSC} + \beta_۲(\overline{X}_{MRSS} - \mu_X) \tag{۱۶}$$

که در آن

$$\overline{Y}_{MRSSC} = \frac{۱}{nm}\sum_{k=۱}^{n}\sum_{j=۱}^{m} Y_{k(m)j},$$

$$\overline{X}_{MRSS} = \frac{۱}{nm}\sum_{k=۱}^{n}\sum_{j=۱}^{m} X_{k[m]j}$$

و $\beta_۲$ یک مقدار ثابت و نامعلوم است و یک برآوردگر برای آن عبارت است از

$$\hat{\beta}_۲ = \frac{\sum_{k=۱}^{n}\sum_{j=۱}^{n}(X_{k[m]j} - \overline{X}_{MRSS})(Y_{k(m)j} - \overline{Y}_{MRSSC})}{\sum_{k=۱}^{n}\sum_{j=۱}^{n}(X_{k[m]j} - \overline{X}_{MRSS})^{۲}} \tag{۱۷}$$

بنابراین برآوردگر رگرسیونی در نمونه‌گیری مجموعه رتبه‌دار میانه‌ای به صورت زیر خواهد بود

$$\overline{Y}_{MRSSD} = \overline{Y}_{MRSSC} + \hat{\beta}_۲(\mu_X - \overline{X}_{MRSS}) \tag{۱۸}$$

که خواص آن در قضیهٔ زیر بیان شده است.

قضیه ۲.۳ اگر رگرسیون Y روی X بفرم (۱۱) باشد برآوردگر رگرسیونی در نمونه‌گیری مجموعه رتبه‌دار میانه‌ای یک برآوردگر نااریب برای میانگین جامعه و دارای واریانس بصورت زیر است

$$Var(\overline{Y}_{MRSSLR}) = \frac{\sigma_y^2}{nm}(1-\rho^2)[1+E(\frac{\overline{Z^2}_{MRSS}}{s_{z_2}^2})]$$

که در آن $Z_{k[m]j} = \frac{X_{k[m]j}-\mu_x}{\sigma_x}$ ، $\overline{Z}_{MRSS} = \frac{1}{nm}\sum_{k=1}^{n}\sum_{j=1}^{n} Z_{k[m]j}$ و

$$S_{z_2}^2 = \frac{1}{nm}\sum_{k=1}^{n}\sum_{j=1}^{n}\left((Z_{k[k]j}-\overline{Z}_{MRSS}\right)^2$$

حال فرض کنید $X_{k[e]j}$ و $Y_{k(e)j}$ به ترتیب نشان دهندهٔ کوچکترین مقدار X و متغیر اصلی متناظر بدست آمده از kامین نمونه $k = 1, 2, \ldots, \frac{n}{2}$ و بزرگترین واحد رتبه‌بندی شده X و متغیر اصلی متناظر بدست آمده از $k = \frac{n}{2}+1, \ldots, n$ در jامین تکرار در صورت زوج بودن اندازه نمونه باشد. همچنین این نماد را برای نشان دادن کوچکترین واحد رتبه‌بندی شده و مقدار متغیر اصلی متناظر با آن در kامین نمونه $k = 1, 2, \ldots, \frac{n-1}{2}$ و میانه X و متغیر اصلی متناظر با آن در kامین نمونه $k = \frac{n+1}{2}$ و بزرگترین واحد رتبه‌بندی شده X و متغیر اصلی متناظر با آن در kامین نمونه $k = \frac{n+1}{2}+1, \ldots, n$ و jامین تکرار، در صورتی که اندازه نمونه فرد باشد را بکار می‌بریم.

در این صورت برآوردگر رگرسیونی نمونه‌گیری مجموعه رتبه‌دار حدی برای μ_Y عبارت است از

$$\overline{Y}_{ERRSSLR} = \overline{Y}_{ERSSC} + \hat{\beta}_2(\mu_X - \overline{X}_{ERSS}) \qquad (19)$$

که $\overline{Y}_{ERSSC} = \frac{1}{nm}\sum_{k=1}^{n}\sum_{j=1}^{m} Y_{k(e)j}$ ، $\overline{X}_{ERSS} = \frac{1}{nm}\sum_{k=1}^{n}\sum_{j=1}^{m} X_{k[e]j}$ و $\hat{\beta}_2$ برابر است با

$$\hat{\beta}_2 = \frac{\sum_{k=1}^{n}\sum_{j=1}^{n}(X_{k[e]j}-\overline{X}_{ERSS})(Y_{k(e)j}-\overline{Y}_{ERSSC})}{\sum_{k=1}^{n}\sum_{j=1}^{n}(X_{k[e]j}-\overline{X}_{ERSS})^2} \qquad (20)$$

خواص $\overline{Y}_{ERSSLR}$ در قضیه زیر بیان شده است.

قضیه ۳.۳ اگر رگرسیون Y روی X خطی باشد آنگاه برآوردگر رگرسیونی در نمونه‌گیری مجموعه رتبه‌دار حدی یک برآوردگر نااریب برای میانگین جامعه است و واریانس آن عبارت است از

$$Var(\overline{Y}_{ERSSLR}) = \frac{\sigma_y^2}{nm}(1-\rho^2)[1+E(\frac{\overline{Z^2}_{ERSS}}{s_{z_3}^2})]$$

که در آن $Z_{k[e]j} = \frac{X_{k[e]j}-\mu_x}{\sigma_x}$ ، $\overline{Z}_{ERSS} = \frac{۱}{nm}\sum_{k=۱}^{n}\sum_{j=۱}^{n} Z_{k[e]j}$ و

$$S^{۲}_{z_۳} = \frac{۱}{nm}\sum_{k=۱}^{n}\sum_{j=۱}^{n}\left((Z_{k[e]j} - \overline{Z}_{ERSS}\right)^{۲}$$

۲.۳ برآوردگرهای رگرسیونی در نمونه‌گیری مجموعه رتبه‌دار با نامعلوم بودن میانگین متغیر کمکی

در بیشتر کاربردها میانگین جامعه متغیر کمکی نامعلوم است و با استفاده از روش نمونه‌گیری دوگانه یا دو فازی[۱۵] برآورد می‌شود. برای این کار در فاز اول یک نمونه $n^{۲}m$تایی به صورت تصادفی انتخاب می‌شود و از آن برای برآورد میانگین جامعه متغیر کمکی μ_X استفاده می‌شود. سپس از این نمونه $n^{۲}m$تایی یک زیر نمونه nmتایی در فاز دوم به صورت تصادفی ساده یا نمونه‌گیری مجموعه رتبه‌دار انتخاب می‌شود که برای مطالعه خصیصه Y استفاده می‌شود. فرض کنید $\overline{X}_d$ میانگین نمونه X بر اساس مشاهدات $n^{۲}m$تایی X در فاز اول باشد. در این صورت $\overline{X}_d$ یک برآوردگر ناااریب برای μ_X است. اگر نمونه‌گیری تصادفی ساده در فاز دوم نمونه‌گیری استفاده شود، برآوردگر رگرسیونی با استفاده از نمونه‌گیری دوگانه به صورت زیر خواهد بود

$$\overline{Y}_{LRD} = \overline{y} + \hat{\beta}(\overline{X}_d - \overline{x}) \qquad (۲۱)$$

قضیه ۴.۳ $\overline{Y}_{LRD}$ یک برآوردگر ناااریب برای میانگین جامعه است و اگر (X,Y) دارای توزیع نرمال دو متغیره باشند آنگاه واریانس آن عبارت است از

$$Var(\overline{Y}_{LRD}) = \frac{\sigma^{۲}_y}{nm}(۱-\rho^{۲})[۱ + \frac{n-۱}{n}\frac{۱}{nm-۳}] + \frac{\rho^{۲}\sigma^{۲}_y}{n^{۲}m} \qquad (۲۲)$$

فرض کنید نمونه‌گیری مجموعه رتبه‌دار ساده در دومین فاز از نمونه‌گیری دوگانه استفاده شود. یو و لام (۱۹۹۷) برآوردگر رگرسیونی میانگین جامعه μ_Y را به صورت زیر ارائه کردند

$$\overline{Y}_{RSSLRD} = \overline{Y}_{RSSC} + \hat{\beta}_۱(\overline{X}_d - \overline{X}_{RSS}) \qquad (۲۳)$$

که $\hat{\beta}_۱$ توسط رابطه (۱۳) بدست می‌آید.

قضیه ۵.۳ برآوردگر $\overline{Y}_{RSSLRD}$ یک برآوردگر ناااریب برای میانگین جامعه μ_Y است و واریانس آن عبارت است از

$$Var(\overline{Y}_{RSSLRD}) = \frac{\sigma^{۲}_y}{nm}(۱-\rho^{۲})[۱ + E(\frac{(\overline{Z}_{RSS} - \overline{Z})^{۲}}{s^{۲}_{z_۱}})] + \frac{\rho^{۲}\sigma^{۲}_y}{n^{۲}m} \qquad (۲۴)$$

15) Two-Phase Sampling

که $\overline{Z} = \frac{\overline{X}_d - \mu_X}{\sigma_X}$، $\overline{Z}_{RSS}$ و $S^2_{Z_1}$ در قضیه ۱.۳ معرفی شده‌اند.

اگر نمونه‌گیری مجموعه رتبه‌دار میانه‌ای در دومین فاز از نمونه‌گیری دوگانه استفاده شود، برآوردگر رگرسیونی میانگین جامع μ_Y به صورت زیر خواهد بود.

$$\overline{Y}_{MRSSLRD} = \overline{Y}_{MRSSC} + \hat{\beta}_2(\overline{X}_d - \overline{X}_{MRSS}) \qquad (۲۵)$$

که $\hat{\beta}_2$ توسط رابطه (۱۷) به دست می‌آید. همانند قبل برآوردگر $\overline{Y}_{MRSSLRD}$ یک برآوردگر ناااریب برای میانگین جامعه μ_Y با واریانس زیر است

$$Var(\overline{Y}_{MRSSLRD}) = \frac{\sigma_y^2}{nm}(1-\rho^2)[1 + E(\frac{(\overline{Z}_{MRSS} - \overline{Z})^2}{s_{z_2}^2})] + \frac{\rho^2\sigma_y^2}{n^2 m} \qquad (۲۶)$$

که $\overline{Z}_{MRSS}$ و $S^2_{Z_2}$ در قضیه ۲.۳ معرفی شده‌اند. اگر نمونه‌گیری مجموعه رتبه‌دار حدی در دومین فاز از نمونه‌گیری دوگانه استفاده شود، برآوردگر رگرسیونی میانگین جامعه μ_Y به صورت زیر خواهد بود

$$\overline{Y}_{ERSSLRD} = \overline{Y}_{ERSSC} + \hat{\beta}_3(\overline{X}_d - \overline{X}_{ERSS}) \qquad (۲۷)$$

که $\hat{\beta}_3$ توسط رابطه (۲۰) به دست می‌آید. همانند قبل برآوردگر $\overline{Y}_{ERSSLRD}$ یک برآوردگر ناااریب برای میانگین جامعه μ_Y با واریانس زیر است

$$Var(\overline{Y}_{ERSSLRD}) = \frac{\sigma_y^2}{nm}(1-\rho^2)[1 + E(\frac{(\overline{Z}_{ERSS} - \overline{Z})^2}{s_{z_3}^2})] + \frac{\rho^2\sigma_y^2}{n^2 m} \qquad (۲۸)$$

که $\overline{Z}_{ERSS}$ و $S^2_{Z_3}$ در قضیه ۳.۳ معرفی شده‌اند.

۴ کاربرد روش‌های مختلف نمونه‌گیری مجموعه رتبه‌دار در برآورد مقدار کل محصول گندم ایران

برای نشان دادن چگونگی اجرای روش‌های نمونه‌گیری مجموعه رتبه‌دار، به عنوان کاربرد، میزان تولید گندم کشور را برآورد می‌کنیم. با استفاده از اطلاعات موجود در سالنامه آماری ۱۳۸۰ مرکز آمار، ابتدا کشور ایران را به شهرستانهای مختلف تقسیم می‌کنیم. این شهرستانها به عنوان واحدهای جامعه در نظر گرفته می‌شوند. متغیر مورد نظر در هر واحد (شهرستان) میزان تولید محصول گندم در آن واحد (شهرستان) است و در جاهایی که متغیر همراه لازم است سطح زیر کشت گندم آن شهرستان به عنوان متغیر همراه در نظر گرفته می‌شود. میزان تولید محصول گندم بر

حسب تن و سطح زیرکشت آن بر حسب هکتار می‌باشد. جامعه آماری مورد نظر شهرستانهای ایران است و حجم جامعه ۲۷۶ می‌باشد. همچنین تعداد نمونه انتخابی برای برآورد مقدار کل گندم ایران ۲۰ نمونه می‌باشد.

با به کارگیری روش‌های مختلف نمونه‌گیری مجموعه رتبه‌دار برآوردهای مقدار کل محصول گندم ایران بدست آمده که نتایج آن در جداول زیر ارائه شده است.

جدول ۱: دقت نسبی برآوردهای عادی روش نمونه‌گیری مجموعه رتبه‌دار نسبت به نمونه‌گیری تصادفی ساده

روش نمونه‌گیری	خطای استاندارد برآوردها	برآورد مقدار کل گندم ایران	دقت نسبی
رتبه‌دار ساده	۱,۶۹۷,۰۸۱	۸,۸۶۴,۵۲۷	۵٫۲۸
رتبه‌دار میانه‌ای	۱,۱۳۶,۰۰۴	۷,۳۸۶,۹۲۱	۱۱٫۷۹
رتبه‌دار حدی	۱,۰۵۸,۲۵۲	۸,۸۹۱,۲۵۹	۱۳٫۵۹

جدول ۲: دقت نسبی برآوردهای رگرسیونی مجموعه رتبه‌دار با μ_X معلوم نسبت به نمونه‌گیری تصادفی ساده

برآورد رگرسیونی	خطای استاندارد برآوردها	برآورد مقدار کل گندم	دقت نسبی
رتبه‌دار ساده μ_X معلوم	۱,۵۱۰,۴۸۷	۸,۴۴۷,۴۹۸	۶٫۶۷
رتبه‌دار میانه‌ای μ_X معلوم	۱,۶۷۵,۷۴۸	۱۰,۴۷۹,۰۰۳	۵٫۴۲
رتبه‌دار حدی μ_X معلوم	۱,۵۵۱,۱۳۳	۸,۶۲۷,۹۹۹	۶٫۳۲
ساده μ_X معلوم	۳,۷۷۲,۹۳۱	۱۳,۴۵۳,۲۶۷	۱٫۰۷

از جدول ۱ معلوم می‌شود که برآورد روش نمونه‌گیری مجموعه رتبه‌دار حدی در بین برآوردهای عادی روش نمونه‌گیری مجموعه رتبه‌دار کاراتر از بقیه برآوردهای دیگر است.

از جدول ۲ معلوم می‌شود که برآورد رگرسیونی مجموعه رتبه‌دار ساده با معلوم بودن میانگین متغیر کمکی در بین برآوردهای رگرسیونی روش نمونه‌گیری مجموعه رتبه‌دار با معلوم بودن میانگین متغیر کمکی کاراتر از بقیه برآوردهای دیگر است.

از جدول ۳ معلوم می‌شود که برآورد رگرسیونی مجموعه رتبه‌دار ساده با معلوم نبودن میانگین متغیر کمکی در بین برآوردهای رگرسیونی روش نمونه‌گیری مجموعه رتبه‌دار با معلوم نبودن میانگین متغیر کمکی کاراتر از بقیه برآوردهای دیگر است

جدول ۳: دقت نسبی برآوردهای رگرسیونی مجموعه رتبه‌دار با معلوم نبودن μ_X نسبت به نمونه‌گیری تصادفی ساده

برآورد رگرسیونی	خطای استاندارد برآوردها	برآورد مقدار کل گندم ایران	دقت نسبی
رتبه‌دار ساده μ_X نامعلوم	۱,۵۲۰,۱۴۴	۸,۶۲۸,۸۵۶	۶/۵۸
رتبه‌دار میانه‌ای μ_Xنامعلوم	۱,۶۷۵,۷۴۸	۱۰,۲۸۱,۱۱۶	۵/۴۲
رتبه‌دار حدی μ_Xنامعلوم	۱,۵۶۳,۶۷۱	۹,۱۰۲,۷۱۰	۶/۲۲
تصادفی سادهμ_Xنامعلوم	۳,۷۷۹,۴۲۸	۱۴,۴۳۵,۶۴۹	۱/۰۷

جدول ۴: مقایسه دقت نسبی برآوردهای کارای روش‌های نمونه‌گیری مجموعه رتبه‌دار

روش نمونه‌گیری	خطای استاندارد برآوردها	برآورد مقدار کل گندم ایران	دقت نسبی
رتبه‌دار حدی	۱,۰۵۸,۲۵۲	۸,۸۹۱,۲۵۹	۵۹/۱۳
رتبه‌دار میانه‌ای با استفاده از متغیر همراه	۱,۶۵۷,۱۲۷	۹,۸۷۵,۴۲۹	۵۴/۵
رتبه‌دار ساده μ_X معلوم	۱,۵۱۰,۴۸۷	۸,۴۴۷,۴۹۸	۶۷/۶
رتبه‌دار ساده μ_X نامعلوم	۱,۵۲۰,۱۴۴	۸,۶۲۸,۸۵۶	۵۸/۶

در آخر به مقایسه برآوردهای کارای بدست آمده از جداول بالا می‌پردازیم.

از جدول ۴ معلوم می‌شود که برآورد روش نمونه‌گیری مجموعه رتبه‌دار حدی کاراتر از بقیه برآوردهای کارا می‌باشد.

از جداول ۱ تا ۴ معلوم می‌شود که روش‌های مختلف نمونه‌گیری مجموعه رتبه‌دار همیشه کاراتر از روش نمونه‌گیری تصادفی ساده با تعداد نمونه برابر می‌باشند.

مراجع

[1] McIntyre, G. A. (1952). A method of unbiased selective sampling, using ranked set. Australian Journal of Agricultural Research, 3, 385-390.

[2] Muttlak, H. A. (1997). Median ranked set sampling, Journal of Applied Statistical Sciences, 6, 245-255.

[3] Muttlak, H. A. (1998). Median ranked set sampling with concomitant variables and a comparison with ranked set sampling and regression estimators, Environmetrics, 9, 225-267.

[4] Muttlak, H. A. (2001). Regression estimators in extreme and median ranked set samples, Journal of Applied Statistical Sciences, 28(8) 1003-1017.

[5] Patil , G. P., Sinha, A. K. & Taillie, C. (1993). Relative precision of ranked set sampling: a comparison with ranked set sampling and regression estimators, Environmetrics, 4, 399-412.

[6] Samawi, H., Abu-Dayyeh, W. & Ahmad, S. (1996). Extreme ranked set sampling, The Biometrical Journal, 30, 557-586.

[7] Stokes, S. L. (1977). Ranked set sampling with concomitant variables, Communication in Statistics, AG, 1207-1211.

[8] Takahasi, K. & Wakimoto, K. (1967). On unbiased estimates of the population mean based on the sample stratified by mean of ordering, Annals of Statistical Mathematics, 20, 1-31.

[9] Yu, P.L. H. & Lam, K. (1997). Regression estimator in ranked set sampling, Biometrics, 53, 1070-1080.

[۱۰] سالنامه آماری ۱۳۸۰، مرکز آمار ایران.

انتشار داده‌های آماری و کنترل افشای اطلاعات فردی

حمیدرضا نواب‌پور[۱]، محمد بردبار عشرت‌آبادی[۲]

[۱] گروه آمار دانشگاه علامه طباطبایی
[۲] محقق پژوهشکدهٔ آمار

چکیده: با افزایش تقاضا برای آمارها در عرصه‌های جدید، مانند آمارهای زیست محیطی، آمارهای فقر، آمارهای جنسیتی، آمارهای فرهنگی و . . . ، سازمانهای ملی آمار اقدام به برنامه‌ریزی برای تولید و انتشار اینگونه آمارها علاوه بر آمارهایی که به صورت ادواری تولید و منتشر می‌کنند، کرده‌اند. برنامه‌ریزان، اقتصاددانان، و پژوهشگران همواره به داشتن داده‌های خرد تمایل داشته‌اند، زیرا این داده‌ها مانور آنها را در تحلیل پدیده‌های اجتماعی، اقتصادی و فرهنگی بیشتر می‌کند. اما طبق قانون، انتشار داده‌های جمع‌آوری شده نباید منجر به افشای هویت واحدهای اطلاع‌گیری شده، گردد. این امر همواره بهانه‌ای برای عدم انتشار داده‌های حاصل از آمارگیریها به نحوی که مورد تقاضای کاربران آنهاست، بوده است و یا در برخی موارد داده‌ها در جدولهای انتشاراتی ایمن نبوده و امکان افشای هویت واحدهای اطلاع‌گیری شده وجود داشته است. هدف این مقاله ارایه روشهایی است که به توسط آنها می‌توان هم تقاضای کاربران آمارهای تولید شده را اجابت کرد و هم از افشای اطلاعات فردی که خلاف قوانین عمومی آمار کشورها است جلوگیری کرد.

واژه‌های کلیدی: آمارهای رسمی، محرمانگی، داده‌های خرد، رکوردهای ناایمن، متغیر سلسله مراتبی، جدول پیشایندی، فراداده

۱ مقدمه

امروزه مدیریت جامعه‌ها مبتنی بر اطلاعات است، لذا برنامه‌ریزان، اقتصاددانان، و پژوهشگران برای تصمیم‌سازی نیازمند اطلاعات آماری دقیق، روزآمد، و به موقع هستند. از این رو کشورها با تصویب قانون عمومی آمار، اقدام به تأسیس سازمانهای ملی آمار با وظیفه اصلی تولید و انتشار آمارهای رسمی نموده‌اند. معمولا در قانونهای عمومی آمار آحاد ملت ملزم به ارایه اطلاعات صحیح به پرسشگران هستند. در همین حال این قانونها به آنها اطمینان می‌دهد که اطلاعات آنها حفاظت شده و جز در تهیه آمارهای کلی به‌کار برده نمی‌شود. به عبارت دیگر این قانونها به پاسخگویان اطمینان می‌دهد که نحوه انتشار اطلاعات مربوط به افراد موجب افشای هویت آنها نخواهد شد. اصل محرمانگی اطلاعات فردی در همه نظامهای آماری پذیرفته شده است، لذا سازمانهای ملی آمار برای رعایت این اصل محدودیتهایی را برای انتشار داده‌های حاصل از

آمارگیریها اعمال می‌کنند به طوری که بعضا نیازهای پژوهشگران به برخی اطلاعات آماری تامین نمی‌شود و یا گاهی امکان افشای هویتها از جدولهای انتشاراتی سازمانهای ملی آمار وجود دارد. بنابراین در طول سه دهه گذشته ایجاد تعادل بین تأمین نیازهای آماری کاربران و جلوگیری از افشای هویت واحدهای اطلاع‌گیری شده، یکی از دغدغه‌های مهم سازمانهای ملی آمار کشورها بوده است.

یکی از هدفهای مهم یک سازمان ملی آمار، تأمین نیازهای اطلاعات آماری جامعه تا حد امکان است. این هدف، سازمان ملی آمار را موظف می‌کند تا اطلاعات آماری مورد نیاز کاربران را با کمک ابزارهای قانونی جمع‌آوری کرده و به صورت مناسبی در اختیار آنان قرار دهد. اما نکته مهم در این فرآیند، وجود خطر افشای هویت فردی توسط فرد یا افراد متخلف است. شخص متخلف فرضی ممکن است علاقمند به دستیابی به اطلاعات در مورد افرادی خاص باشد، یا اینکه تلاش کند تا با افشای اطلاعات، گردآورنده‌ی آنها را بی‌اعتبار کند.

عموما دو نوع افشای اطلاعات در نظر گرفته می‌شود: افشای هویت و افشای صفت. در نوع اول که در واقع مهمترین نوع افشا است، ابتدا فرد مورد نظر شناسائی شده و سپس بر اساس هویت فرد، اطلاعات مربوط به او از داده‌ها استخراج می‌گردد. اما همواره تعیین هویت شرط لازم برای افشای اطلاعات حساس در مورد یک پاسخگو نیست. در برخی موارد تنها دانستن این که یک پاسخگو عضوی از یک گروه است، بدون این که معلوم شود که کدام یک از آنها است، برای افشای اطلاعات در مورد وی کفایت می‌کند، این نوع افشاء را افشای صفت می‌نامند. عمل افشاء به هر دلیل که صورت پذیرد موجب بی‌اعتمادی عمومی نسبت به سازمان ملی آمار به عنوان حافظ اطلاعات شخصی پاسخگویان شده و کاهش همکاری یا عدم تمایل به همکاری پاسخگویان را در طرحهای آمارگیری به همراه خواهد داشت. نتیجه این امر عدم تولید آمار دقیق، روزآمد، به موقع، ارزان و گسترده خواهد بود که طبیعتا امکان تأمین آمارهای مورد نیاز کاربران به صورت مناسب و مطلوب توسط سازمان ملی آمار وجود نخواهد داشت.

۲ مرور نوشتگان

موضوع کنترل افشای آماری[1] اولین بار توسط دالینوس در سال ۱۹۷۷ مطرح شد و سپس سایر آمارشناسان در کشورهای امریکایی و اروپایی بحث و تحقیق در این موضوع را شروع کردند. کاکس (۱۹۸۰) روشهای پنهان‌سازی مقدماتی و مکمل را بحث کرده است. روشهای پرشیده[2] شامل روشهای گرد کردن تصادفی توسط کاکس (۱۹۸۷)، روش پاسخ پس تصادفی شده[3] توسط گوویلیو و همکاران (۱۹۸۸) بررسی شده‌اند. ویلنبورگ و دوال (۱۹۹۸) روی روشهای بازکدگذاری

1) Statistical Disclosure Control 2) Perturbative Methods 3) Post Randomized Response Method (PRAM) 4) Global Recoding

عام[۴] و پنهان‌سازی موضعی[۵] کار کرده‌اند. رهیافت‌های پرشیده نیز در شکلهای مختلف توسط فاینبرگ و همکاران (۱۹۹۸) ارایه شده‌اند. پیشرفتهای حاصل در کنترل افشای داده‌های آماری را می‌توان در دو گزارش کمیته فدرال روش شناختی آماری، در سالهای ۱۹۷۸ و ۱۹۹۴ و نیز در مقاله‌های شماره‌های ویژه سالهای ۱۹۹۳ و ۱۹۹۸ مجله آمار رسمی[۶] مشاهده نمود. در ایران تا بحال هیچگونه مطالعه و پژوهشی در موضوع کنترل افشای داده‌های آماری به جز پایان‌نامه کارشناسی ارشد آقای محمد بردبار عشرت آبادی (۱۳۸۲) صورت نگرفته است و نیز هیچ تجربه‌ای در ایمن‌سازی اطلاعات آماری منتشر شده وجود ندارد.

۳ روشهای محدودسازی افشاء در داده‌های خرد

هرگاه یک سازمان آماری بخواهد یک مجموعه داده‌های خرد را منتشر کند، متغیرهای شناسایی مستقیم را از فایل داده‌ها حذف می‌کند. بنابراین شیوه مستقیمی برای تعیین اینکه یک رکورد معین مربوط به کدام پاسخگو است وجود ندارد. ظاهرا می‌توان گفت که در صورت انتشار چنین مجموعه داده‌های خردی خطر افشایی وجود ندارد اما این مطلب درست نیست. به مثال زیر توجه کنید. فرض کنید یک سازمان آماری بخواهد مجموعه داده‌های خردی شامل اطلاعاتی در مورد محل سکونت، شغل و سابقه جنایی پاسخگویان را منتشر نماید. علاوه بر این فرض کنید یک رکورد با ترکیب مقادیر زیر در مجموعه داده‌های خرد وجود داشته باشد: «محل سکونت: تهران، شغل: شهردار، سابقه جنایی: یک مورد».

اگر چه فرض بر این است که نام یا آدرس پاسخگو منتشر نمی‌شود، اما افراد زیادی در ایران پی می‌برند که پاسخگو چه کسی است. به ویژه می‌توانند نتیجه بگیرند که این پاسخگو یعنی شهردار تهران دارای یک سابقه جنایی است. به طور کلی هر پاسخگویی که در یک ترکیب از مقادیری که به تعداد کم در جامعه اتفاق می‌افتد قرار بگیرد، ممکن است در معرض خطر شناسایی باشد.

تشخیص اینکه در فایل داده‌های خرد کدام رکوردها و ترکیبهای متغیرها در معرض خطر افشاء هستند، بستگی به چگونگی تعریف سناریوی افشاء دارد. یک چنین سناریویی به تشریح مدلی می‌پردازد که یک شخص متخلف در جهت دستیابی به اطلاعات محرمانه یک یا چند پاسخگو در یک مجموعه داده‌های خرد، به چه صورتی ممکن است عمل کند و چه اطلاعات پیشینی در مورد یک جمعیت خاص می‌تواند داشته باشد. پس از مشخص شدن رکوردها و ترکیبهای ناایمن برای انتشار در بخش ارزیابی خطر افشاء، مرحله حفاظت از فایل داده‌های خرد آغاز می‌شود. در این مرحله توسط روشهای محدودسازی خطر افشای مربوط به داده‌های خرد، یک فایل داده‌های خرد ناایمن به یک فایل داده‌های خرد ایمن قابل انتشار با خطر افشائی که در حد قابل قبولی پائین است، تبدیل می‌شود. این روشها به دو گروه عمده تقسیم می‌شوند.

5) Local Suppression 6) Journal of Official Statistics

گروه اول شامل روشهای ناپرشیده[7] مانند روشهای بازکدگذاری عام و پنهان‌سازی موضعی است که به سبب اعمال آنها روی فایل داده‌های خرد، خطائی به داده‌ها افزوده نمی‌شود و فقط اندکی از میزان اطلاعات موجود در داده‌ها کاسته می‌شود. گروه دوم روشهای پرشیده نام دارند، مانند روش پس تصادفی سازی که به واسطه اعمال این روشها، مقادیر اصلی مشاهده شده، با مقادیر دیگری تعویض می‌شوند. این عمل به گونه‌ای صورت می‌گیرد که امکان انجام تحلیل‌های آماری رایج روی فایل پرشیده شده امکانپذیر است.

۱.۳ روش بازکدگذاری عام و روش پنهان سازی موضعی

این دو روش از رایجترین روشهای محدودسازی افشاء در داده‌های خرد می‌باشند و در گروه روشهای ناپرشیده قرار می‌گیرند. از این دو روش اغلب به صورت توأم استفاده می‌شود. هر دو روش بازکدگذاری عام و پنهان سازی موضعی برای متغیرهای رسته‌ای به کار می‌روند. در روش بازکدگذاری عام چندین رسته از یک متغیر مانند A با یکدیگر ترکیب شده و رسته جدیدی را پدید می‌آورند. مقادیر A متناظر با رسته‌های جدید، مجددا کد گذاری می‌شوند. تعویض رسته‌ها تنها مختص بخش ناایمن فایل داده‌های خرد نیست بلکه به منظور دستیابی به یک گروهبندی یکنواخت، روی تمام فایل داده‌های خرد اعمال می‌شود.

در روش پنهان‌سازی موضعی، مقدار یک متغیر مانند A در یک رکورد K با یک مقدار گمشده تعویض می‌گردد. در حالی که بازکدگذاری عام روی کل فایل داده‌های خرد تأثیر می‌گذارد، پنهان‌سازی موضعی تنها برای یک مقدار به خصوص در یک رکورد ناایمن به کار می‌رود.

هر دو روش موجب از دست رفتن مقداری از اطلاعات موجود در داده‌های خرد می‌شوند. علاوه بر این پنهان‌سازی موضعی ممکن است در صورتی که مقادیر گمشده نادیده گرفته شوند موجب برآوردگرهای اریب گردد. به همین دلیل سعی می‌شود تا از آن در مقیاسی کوچک استفاده شود. هنگامی که باید تعداد زیادی ترکیبهای ناایمن حذف شوند، روش بازکدگذاری عام ترجیح داده می‌شود. وقتی که از تعداد اندکی پنهان‌سازی موضعی استفاده شود، غالبا اریبی تولید شده در برآوردگرها ناچیز است. در عمل یک توازن بین استفاده از این دو روش به وجود می‌آید. اغلب در ابتدا بعضی از متغیرها بازکدگذاری عام می‌شوند و در مورد باقیمانده رکوردهای ناایمن از پنهان‌سازی موضعی برای حداقل مقادیر ممکن، استفاده می‌شود. مینیمم‌سازی تعداد پنهان‌سازی‌ها یک موضوع اساسی است که باید مورد توجه قرار گیرد.

۲.۳ روش پس تصادفی شده (PRAM)

پرام (PRAM) یک روش پرشیده است که برای کنترل افشای متغیرهای رسته‌ای در فایل داده‌های خرد به کار می‌رود. در این روش برای هر رکورد در فایل داده‌های خرد، امتیاز تعدادی از متغیرها متناظر با یک مکانیسم احتمالاتی معین، تغییر می‌کند، وقتی که فایل داده‌های خرد

7) Non-Perturbative Methods

پرشیده شد، شناسایی رکوردهای متناظر با اشخاصی معین در جامعه، دشوار می‌گردد. بنابراین رکوردها در فایل اصلی در برابر خطر شناسائی محافظت شده‌اند که این امر، هدف پرام است. از طرف دیگر تا وقتی که مکانیسم احتمالاتی استفاده شده معلوم باشد، مشخصه‌های داده‌های اصلی که پنهان شده‌اند می‌تواند از فایل داده‌های پرشیده شده برآورد شوند. بنابراین به کار بردن تمام تحلیل‌های آماری امکانپذیر است. فرض کنید ξ متغیری رسته‌ای در فایل اصلی باشد، هدف به کار بردن پرام روی این متغیر است. همچنین فرض کنید X همان متغیر رسته‌ای در فایل پرشیده شده است و ξ و X دارای k رسته، $k = ۱, ۲, \ldots, K$ باشند. اگر $P_{kl} = P(X = l|\xi = k)\ (l, k = ۱, ۲, \ldots, K)$ احتمال این باشد که امتیاز $\xi = k$ در فایل اصلی به $X = l$ در فایل پرشیده شده تغییر یابد، آنگاه $P = p_{kl}$ ماتریسی $K \times K$ با اعضای p_{kl} است. P یک ماتریس مارکوف است یعنی در رابطه $\boldsymbol{PI} = \boldsymbol{I}$ که در آن $\boldsymbol{I}$ برداری $K \times ۱$ از یک‌ها است، صدق می‌کند.

در اینجا فرض کلی بر این است که P معکوس‌پذیر است. این ویژگی برای به کار بردن پرام ضروری نیست ولی $P^{-۱}$ می‌تواند در برآورد توزیع فراوانی ξ در فایل اصلی، به خوبی برای برآورد واریانس تولید شده با استفاده از روش پرام، به کار رود.

پرام را می‌توان به صورت مستقل برای متغیرهای مختلف و همچنین هم زمان روی بیش از یک متغیر نیز به کار برد. مثال زیر تأثیر پرام بر محدودسازی افشاء را تشریح می‌کند.

مثال: فرض کنید فایل داده‌های خرد شامل n رکورد است. این فایل یک نمونهٔ تصادفی ساده به اندازه n از جامعه‌ای با اندازه N است. فایل دقیقا شامل یک زن جراح است. پرام، روی متغیر جنسیت و برای هر رکورد مستقل از رکوردهای دیگر به کار رفته است. امتیاز جنسیت با احتمال ۰٫۹ بدون تغییر باقی می‌ماند و با احتمال ۰٫۱ تغییر می‌کند. یعنی:

الف) ξ: متغیر جنسیت دارای دو ردهٔ ۱ = مرد و ۲ = زن (در فایل اصلی)

ب) X: متغیر جنسیت دارای دو ردهٔ ۱ = مرد و ۲ = زن (در فایل پرشیده شده)

ج) $P_{۲۲} = ۰٫۹$، $P_{۲۱} = ۰٫۱$، $P_{۱۲} = ۰٫۱$، و $P_{۱۱} = ۰٫۹$.

حال فرض کنید که شخص متخلف بداند که جامعه دارای ۱ زن جراح و ۹۹ مرد جراح است. احتمال این که وی پی ببرد، زن جراح در فایل پرشیده شده، در واقع زن جراح در جامعه می‌باشد برابر است با:

$$P(\xi = ۲|\, X = ۲) = \frac{P_{۲۲}P(\xi = ۲)}{P_{۱۲}P(\xi = ۱) + P_{۲۲}P(\xi = ۲)} = \frac{۰/۹ \times ۰/۰۱}{۰/۱ \times ۰/۹۹ + ۰/۹ \times ۰/۰۱} \approx ۰/۰۸$$

مشاهده می‌شود که مقدار این احتمال (۰٫۰۸)، بسیار اندک است. بنابراین داده‌های پرشیده شده به اندازهٔ کافی ایمن هستند. حال فرض کنید اگر متغیر جنسیت با احتمال ۰٫۹۹۹۹ بدون تغییر

بماند و با احتمال ۰/۰۰۰۱ تغییر کند، آنگاه احتمال اینکه زن جراح در فایل پرشیده شده متناظر با زن جراح در جامعه باشد برابر ۰/۹۹ است که احتمال بالایی است و داده‌های پرشیده شده ایمن به نظر نمی‌رسند.

این مثال مشخص می‌کند که به یک کمیت مشخص نیاز است تا معلوم شود که آیا فایل نتیجه شده از اعمال پرام در واقع ایمن شده است؟ این ایده منجر به معرفی کمیتی با عنوان نسبت مورد انتظار به عنوان معیاری برای میزان عدم اطمینان ایجاد شده توسط پرام می‌گردد.

فرض کنید $\xi^{(r)}(X^{(r)})$ امتیاز $\xi(X)$ برای rامین رکورد در فایل داده‌های خرد باشد. به کار بردن پرام، بدین معنا است که با $\xi^{(r)}=k$ مفروض، امتیاز $X^{(r)}$ از توزیع احتمال p_{k1}، p_{k2}، p_{kk} استخراج شود.

نسبت مورد انتظار امتیاز k $(k=,\ldots,K)$ یعنی $ER(k)$ به صورت زیر معرفی می‌شود:

$$k=1,\ldots,K$$

$$ER(k)=\frac{P_{kk}T_{\xi}(k)}{\sum P_{lk}T_{\xi}(l)}$$

که در آن $T_{\xi}(k)$ تعداد رکوردهای فایل اصلی برای $\xi^{(k)}=k$ است.

اگر k در اصل یک امتیاز کمیاب باشد، نسبت مورد انتظار یعنی $ER(k)$، نسبت بین تعداد متوسط رکوردهایی که واقعا متعلق به امتیاز کمیاب k هستند و تعداد متوسط رکوردهایی است که در نتیجه به کار بردن پرام امتیاز k را کسب می‌کنند. مقدار کوچک (ER(k)) نشان دهندهٔ این است که بیشتر احتمال می‌رود که یک رکورد برای $X^{(r)}=k$ در اصل متعلق به این امتیاز نبوده و بنابراین ایمن کنندهٔ فایل پرشیده شده است.

هنگامی‌که پرام برای یک فایل داده‌های خرد به کار می‌رود، این سؤال پیش می‌آید که: پرام چه تأثیری روی انواع تحلیلهای آماری دارد؟ روشن است که تحلیلی که روی فایل پرشیده شده صورت گیرد، نتایجی متفاوت از تحلیل روی فایل اصلی دارد. این امکان وجود دارد تا نتایجی که از تحلیل‌های معینی روی فایل پوشیده شده حاصل می‌شود، بتواند به نتایجی که می‌توانست از فایل اصلی حاصل گردد، تبدیل شوند.

برای مثال فرض کنید بخواهیم یک تحلیل رگرسیونی روی متغیر عددی وابسته y و متغیر رسته‌ای مستقل ξ انجام دهیم، و فرض کنید که پرام برای ξ با K رسته استفاده شده است. برای هر رکورد در فایل داده‌ها، متغیرهای ظاهری $\delta_1,\delta_2,\cdots,\delta_k$ به صورت زیر تعریف می‌شوند:

$$\delta_k=\begin{cases}1 \longrightarrow \xi^{(r)}=k\\ 0 \longrightarrow \xi^{(r)}\neq k\end{cases}\qquad k=1,\ldots,K$$

علاوه بر این فرض کنید که:

$$T_{\xi}^{y}(k)=\sum y^{(r)}I_{\{\xi^{(r)}=k\}}$$

و

$$T_\xi^y = (T_\xi^y(۱), \ldots, T_\xi^y(K))$$

که در آن $y^{(r)}$ بر مقدار y برای rامین رکورد در فایل و I بر تابع نشانگر[8] دلالت دارد و T_ξ^y مجموع مقادیر y برای رکوردهایی است که دارای امتیاز k می‌باشند. کویمان (۱۹۹۷) نشان داد که T_ξ^y به طور نااریب توسط $\vec{T}$ برآورد می‌شود.

اکنون رگرسیون y روی ξ با رگرسیون y روی $\delta_۱, \ldots, \delta_K$ یکسان است. ضرایب رگرسیونی توسط $(D^tD)^{-۱}D^ty$ داده می‌شوند که در آن D ماتریسی $n \times K$ است که عضو (r,j)ام آن برابر مقدار δ_j برای rامین رکورد در فایل داده‌های اصلی است. (با توجه به اینکه δ_j فقط اعداد ۰ و ۱ را اختیار می‌کند، اعضای ماتریس D اعداد ۰ و ۱ خواهند بود.)

(D^tD) می‌تواند به صورت نااریب توسط $\vec{T}$ (توجه کنید که (D^tD) ماتریسی قطری است که عضو (k,k)ام آن $T_\xi(k)$ است) و D^ty می‌تواند به صورت نااریب توسط $\vec{T}$، برآورد شوند. بعلت اینکه برآوردها نااریبند، برآوردگر رگرسیونی نتیجه شده سازگار است. این نتیجه برای تمام فنون تحلیل آماری بر پایهٔ گشتاورهای مرتبهٔ دوم داده‌ها از قبیل تحلیل تشخیصی[9] و تحلیل واریانس[10] صادق است. در عمل ابتدا باید تصمیم گرفت که برای کدام متغیرها باید از پرام استفاده کرد. سپس، برای هر یک از متغیرهایی که پرام برای آنها به کار خواهد رفت، باید تصمیم گرفت که کدام رسته می‌تواند به کدام رسته و با چه احتمالی تغییر کند.

۳.۳ روش ریزانبوهش[11]

روش ریزانبوهش در گروه روشهای پرشیده جای می‌گیرد و برای متغیرهای کمی به کار می‌رود. این روش در ساده‌ترین حالت بر روی یک متغیر و در حالت پیچیده‌تر بر روی بیش از یک متغیر اعمال می‌شود. ایدهٔ اصلی در این روش این است که قواعد محرمانگی این اجازه را می‌دهد که اگر در یک مجموعه داده‌ها، پاسخگویان در گروههای k عضوی یا بیشتر قرار گیرند به طوری که k یک مقدار آغازین باشد و هیچ پاسخگویی از روی مقدارش شناسایی نشود، آنگاه آن مجموعه داده‌ها امکان انتشار دارد.

مسأله افرازی که در ریزانبوهش پیش می‌آید با مسأله خوشه‌بندی کلاسیکی که هدف آن تقسیم کردن جامعه به تعداد ثابتی گروههای ناسازگار، بدون توجه به اندازهٔ گروهها است، متفاوت اما در ارتباط است. قیدی که در ریزانبوهش وجود دارد در مورد اندازهٔ گروهها است. افرازهای نتیجه شده در ریزانبوهش نمی‌توانند شامل گروههایی با اندازهٔ کوچکتر از k باشند. چنین افرازی، $k-$افراز نام دارد. طبق تعریف $k-$افراز بهینه، آن افرازی است که همگنی درون گروه، ماکسیمم باشد، همگنی بیشتر درون گروه، فقدان اطلاعات کمتری را به همراه دارد، زیرا در ریزانبوهش مقادیر

8) Indicator Function 9) Discriminant Analysis 10) Analysis of Variance
11) Microaggregation

جدول ۱: میزان فروش مؤسسات تجاری به تفکیک نوع و ناحیه فعالیت

ناحیه / فعالیت	A	B	C	کل
۱	۱۱	۴۷	۵۸	۱۱۶
۲	۱	۱۵	۳۳	۴۹
۳	۲	۳۱	۲۰	۵۳
کل	۱۴	۹۳	۱۱۱	۲۱۸

یک گروه با مقدار متوسط گروه، تعویض می‌شوند. در بحث خوشه‌بندی، استفاده از معیار مجموع توانهای دوم برای اندازه‌گیری همگنی، مساله‌ای رایج است که از آن در ریزانبوهش نیز استفاده می‌شود.

۴ روشهای محدودسازی افشاء در جدول‌های انتشاراتی

۱.۴ داده‌های جدولی

جدول‌ها رایجترین محصولات سازمان‌های آماری می‌باشند که شامل داده‌های جمع‌بندی شده به عنوان مقادیر خانه‌های جدول می‌باشند و به دو صورت جدول‌های مقداری و جدول‌های فراوانی منتشر می‌شوند. به عنوان مثال میزان سرمایه‌گذاری کارخانجات بر اساس ناحیه و نوع فعالیت، یک جدول مقداری و تعداد کارخانجات برحسب ناحیه و نوع فعالیت یک جدول فراوانی را تشکیل می‌دهند.

۲.۴ تشخیص خانه‌های حساس

به‌علت اینکه داده‌ها مربوط به افراد پاسخگو به صورت تکی نیستند، به نظر می‌رسد که خطر افشای اطلاعات فردی وجود نداشته باشد، در صورتی که همیشه اینگونه نیست. برای روشن شدن مطلب، جدول زیر را که نشان دهنده‌ی میزان فروش تعدادی از مؤسسات تجاری به تفکیک نوع و ناحیه فعالیت است، در نظر بگیرید. این جدول در نگاه اول برای انتشار، پذیرفتنی به‌نظر می‌رسد زیرا تنها شامل داده‌های خلاصه شده است و داده‌های از پاسخگویان تکی ندارد. به هر حال این نتیجه‌گیری عجولانه است.

برای مثال، فرض کنید که این جدول از یک مشاهدهٔ کامل حاصل شده است و تنها یک مؤسسه در ناحیه B با فعالیت ۲ وجود دارد. در نتیجه می‌توان فهمید که میزان فروش این مؤسسه تجاری برابر ۱۵ است. پس اگر قرار باشد که میزان فروش این مؤسسه تجاری حفاظت شود، جدول نباید منتشر شود. بنابراین جدولهایی که دارای خانه‌های شامل داده‌های مربوط به تنها یک پاسخگو هستند، نباید منتشر شوند. ولی همیشه موضوع به این سادگی نیست. فرض

کنید در ناحیه B و فعالیت ۲ به جای یک مؤسسه، دو مؤسسه تجاری وجود داشته باشد. در این حالت هرکدام از این دو مؤسسه به راحتی می‌تواند به میزان فروش دیگری پی ببرد. ممکن است منتشرکننده‌ی داده‌ها به جای وجود حداقل دو عضو در هر خانه‌ی جدول، خواستار وجود سه عضو یا بیشتر در هر خانه‌ی جدول باشد. متاسفانه این حالت نیز همیشه رضایتبخش نیست.

فرض کنید در ناحیه B، ده مؤسسه تجاری دارای فعالیت ۲ باشند و یکی از آنها ۹۵ درصد مقدارکل خانه را ارایه نماید. در این حالت اگر شخص متخلف بداند که میزان فروش این مؤسسه خیلی بالا است، می‌تواند یک برآورد تقریبا دقیق از میزان فروش آن به‌دست آورد. در چنین حالتی گفته می‌شود که این خانهٔ جدول، تحت تسلط این عضو است.

بنابراین برای کنترل خطر افشاء در جدول‌ها، ابتدا باید دنبال راهی گشت که توسط آن بتوان خانه‌هایی را که مناسب انتشار نیستند و امکان افشای اطلاعات تکی پاسخگویان توسط آنها وجود دارد (خانه‌های حساس) شناسایی نمود که برای این موضوع معمولا از قواعدی استفاده می‌شود که در ادامه بیان می‌شوند.

در مورد جدول‌های مقداری، عمومی‌ترین ملاکی که به‌کار می‌رود، قاعدهٔ تسلط (n,k)[۱۲] است که به صورت زیر تعریف می‌شود: یک خانه جدول حساس است اگر مجموع مقادیر عرضه شده توسط n عدد از بزرگترین پاسخگویان به مقدارکل خانه، بیشتر از k درصد مقدارکل خانه را شامل شود.

معمولا مقدارکوچکی (ماکسیمم ۵) برای n و مقدار بزرگی (مثلا ۸۰) برای k منظور می‌شود. ایده‌ی اصلی این قاعده این است که، حفاظت‌کننده‌ی داده‌ها از این که $n-1$ پاسخگو بتوانند به وسیله‌ی جمع کردن مقادیرشان و مقایسه‌ی آن با مقدارکل خانه، یک برآورد دقیق از مقدار پاسخگوی nام به‌دست آورند، جلوگیری می‌کند.

قاعده‌ی دیگری که در مورد جدول‌های مقداری به‌کار می‌رود قاعده‌ی پیشین-پسین $((p,q))$[۱۳] است. فرض می‌کنیم که هر پاسخگو قبل از انتشار جدول می‌تواند مقدار عرضه شده توسط پاسخگوی دیگر به مقدار یک خانه‌ی جدول را با اختلاف کمتر یا برابر q درصد از مقدار واقعی آن، برآورد نماید. اگر پس از انتشار جدول، این امکان برای برخی پاسخگویان به وجود آید که مقدار پاسخگویی دیگر به خانه‌ی جدول را با اختلاف کمتر یا برابر p درصد $(p<q)$ مقدار اصلی آن، برآورد نماید، این خانه از جدول، حساس شناخته می‌شود.

تعیین حساسیت یک خانه می‌تواند از راههای گوناگون تعیین شود. در این قسمت نشان می‌دهیم که قاعدهی تسلط (n,k) و قاعده‌ی پیشین-پسین $((p,q))$ در حقیقت اعضایی از یک رده‌ی بزرگتر از اندازه‌های حساسیت‌اند که رده‌ی معیارهای حساسیت خطی[۱۴] نامیده می‌شوند.

در مورد جدول‌های فراوانی که در خانه‌های آن، تعداد پاسخگویان دارای شرایط آن خانه‌ها آمده است، قاعده‌ای که غالبا برای تشخیص خانه‌های حساس به کار می‌رود بدین صورت است که نباید تعداد پاسخگویان در هر خانه‌ی جدول از عدد مشخصی (مقدار آغازین) مثلا سه کمتر

12) (n,k) Dominance Rule 13) (p,q) Prior- Postrior rule 14) Linear Sensitivity Measures

جدول ۲: قسمتی از یک جدول بزرگ در مورد تخلف‌های زیست محیطی

تخلف / ناحیه و فعالیت	بله	خیر	کل
-	-	-	-
ناحیهٔ B و فعالیت ۳	۴	۱	۵
-	-	-	-

باشد. در غیر اینصورت، امکان شناسایی پاسخگویان این خانه‌ها به صورت تکی وجود دارد. به هر حال حالتهایی وجود دارد که چنین قاعده‌ی تسلطی برای تعیین خانه‌های حساس مناسب نیست. به‌عنوان مثال جدول ۲ که قسمتی از یک جدول بزرگتر در مورد تخلفهای زیست محیطی مؤسسات است را در نظر بگیرید. این جدول مشخص می‌کند که تنها یک مؤسسه دارای تخلف نبوده است. یک شخص بیرونی می‌تواند با احتمال ۸۰ درصد چهار مؤسسه دارای تخلف را تشخیص دهد در حالی که هرکدام از مؤسسات با احتمال ۷۵ درصد می‌تواند تخلف را تشخیص دهد.

پس از تشخیص خانه‌های حساس، باید قبل از انتشار جدول، روش‌های محدودسازی افشاء را روی آنها به کار برد تا در برابر خطر افشای اطلاعات مربوط به پاسخگویان تکی، ایمن گردند.

۳.۴ روش‌های باز طراحی جدول و پنهان‌سازی خانه‌ای

این روش‌ها که از گروه روش‌های غیر پرشیده هستند، زیاد مورد استفاده قرار می‌گیرند و از قدیمی‌ترین روش‌های محدودسازی خطر افشاء در جدول‌ها می‌باشند.

ممکن است در یک جدول، در یک سطر یا ستون، خانه‌های زیادی، حساس تشخیص داده شوند. در این حالت توصیه می‌شود ابتدا جدول دوباره طراحی شود. یعنی در طبقه‌بندی متغیرهای تبیینی[15] تجدیدنظر شود. در این صورت جزئیات اطلاعات آماری ارایه شده در جدول کاهش می‌یابد اما از تعداد خانه‌های حساس به مقدار زیادی کاسته خواهد شد. اگر هنوز تعداد خانه‌های حساس زیادی باقی مانده باشد می‌توان باز هم جزئیات جدول راکاهش داد. هنگامی‌که تعداد خانه‌های حساس اندک باشد، می‌توان از روش‌های کنترل افشای دیگری که به صورت موضعی عمل می‌کنند استفاده نمود. نمونه‌ای از این روشها، روش پنهان‌سازی خانه‌ای است.

یکی از رایج‌ترین روش‌هایی که برای کنترل خطر افشاء در جدول‌ها به کار می‌رود، روش پنهان‌سازی خانه‌ای است. در این روش مقدار خانه‌ی حساس از جدول حذف شده و به جای آن از یک علامت مثلا حرف X استفاده می‌شود. برای توضیح بیشتر، جدول ۳ را در نظر بگیرید. فرض کنید که مقدار خانهٔ متناظر با فعالیت دو و ناحیهٔ C طبق قاعدهٔ تسلط به کار رفته، حساس تشخیص داده شده و نمی‌تواند منتشر شود. مقدار این خانه پنهان می‌شود. به طور کلی، پنهان کردن خانه حساس، به تنهایی کافی نیست. همانطورکه در جدول ۳ مشخص است، خانهٔ پنهان

15) Explanatory Variables

جدول ۳: میزان سرمایه‌گذاری کارخانجات به تفکیک ناحیه و نوع فعالیت (پس از پنهان‌سازی مقدماتی)

ناحیه / فعالیت	A	B	C	کل
۱	۲۰	۵۰	۱۰	۸۰
۲	۸	۱۹	X	۴۹
۳	۱۷	۳۲	۱۲	۶۱
کل	۴۵	۱۰۱	۴۴	۱۹۰

شده می‌تواند با استفاده از جمع‌های حاشیه‌ای به آسانی محاسبه شود. گزینهٔ دیگر، پنهان کردن خانه‌های اضافه‌ی دیگری در جدول که حساس نیستند، است. این عمل، پنهان‌سازی خانه‌ای مکمل و خانه‌ها، پنهان شده‌های مکمل نامیده می‌شوند. یک الگوی پنهان‌سازی خانه‌ای مکمل در جدول ۳، پنهان کردن مقادیر متناظر با خانه‌های فعالیت دو و سه در ناحیهٔ A و فعالیت سه در ناحیهٔ C است.

۴.۴ افزودن نوفه به داده‌های خرد قبل از جدول‌سازی

یک روش دیگر حفاظت پاسخگویان تکی افزودن نوفه به داده‌های آنهاست. این شیوه برای داده‌های جدول‌های مقداری به کار می‌رود. فرض می‌شود که به داده‌ی متناظر با یک مؤسسه، عدم اطمینانی به میزان اندک مثلا ده درصد افزوده شود (درصد استفاده شده، نزد حفاظت‌کننده‌ی داده‌ها محرمانه باقی می‌ماند). در نتیجه اگر یک خانه شامل تنها یک مؤسسه باشد یا اینکه اگر تنها یک مؤسسه، خانه را تحت تسلط داشته باشد، آنگاه مقدار این خانه نمی‌تواند به‌عنوان تقریبی نزدیک از مقدار مؤسسهٔ مسلط استفاده شود، زیرا این مقدار دارای نوفه‌ی اضافه شده است (در این حالت با ۱۰ درصد تغییر می‌کند) با افزودن نوفه‌ی، از افشای مقدار صحیح مؤسسهٔ مسلط جلوگیری می‌شود.

۵.۴ روش گرد کردن

در این روش مقادیر خانه‌های جدول به مضرب‌های صحیح یک عدد پایه‌ی ثابت، گرد می‌شوند. گرد کردن، یک حالت ویژه‌ی داده‌های پرشیده است که در آن برای حفظ اطلاعات، داده‌ها تغییر می‌یابند.

چندین روش گرد کردن وجود دارد. ساده‌ترین روش، روش گردکردن استاندارد است که در آن هر مقدار از خانه‌های جدول به نزدیکترین مضرب پایه‌ی گردکردن، گرد می‌شود. گردکردن استاندارد دو محدودیت عمده دارد. اول اینکه، معمولا ساختار جمع‌پذیری جدول حفظ نمی‌شود، برای مثال، جدول ۴ را که جدولی یک بعدی و نشان دهنده‌ی تعداد مؤسسات تجاری دارای فعالیت

جدول ۴: تعداد مؤسسات تجاری با فعالیت دو به تفکیک ناحیه

ناحیه / فعالیت	ناحیهA	ناحیه B	کل
فعالیت ۲	۲	۲	۴

دو به تفکیک ناحیه است را در نظر بگیرید. فرض کنید بخواهیم این جدول را با پایه‌ی گردکردن، ۵، گرد کنیم. هر دو مقدار خانه‌های داخلی جدول به عدد صفر گرد می‌شوند، در حالی که جمع حاشیه‌ای جدول به عدد ۵ گرد می‌شود. این امر موجب ناسازگاری در جدول می‌گردد. دومین محدودیت که از نظر کنترل افشاء مهم است، این است که گاهی اوقات امکان دارد که مقادیر خانه‌های اصلی آشکار گردند یا اینکه به تعداد معدودی از مقادیر اصلی محدود گردد. این نقص گردکردن استاندارد، انگیزهٔ جستجو برای روشهای دیگر گردکردن که جمع‌پذیری را حفظ نمایند، به وجود آورد. ایده‌ی این روشها این است که هر مقدار (به جز مقادیری که مضربی از مقدار پایه‌اند) به یکی از دو نزدیکترین مضرب مقدار پایه به طوری که جدول جمع‌پذیر باشد، گرد می‌شوند.

روشهای گردکردن قطعی نیستند بلکه احتمالاتی‌اند. علاوه بر این دارای این خاصیت هستند که خطای گردکردن دارای امید ریاضی صفر است. این امر در صورتی تحقق می‌یابد که یک عدد مانند a که می‌تواند به صورت $a = kb + r$ نوشته شود که در آن $0 \leq r < b$ و k عددی صحیح و b پایه‌ی صحیح گردکردن است، به a^* که مقادیر kb و $(k+1)b$ را با احتمالهای زیر اختیار می‌کند گرد شود.

$$p(a^* = (k+1)b) = \frac{r}{b} \quad p(a^* = kb) = 1 - \frac{r}{b}$$

به سادگی دیده می‌شود که:

$$E(a^*) = (k+1)b \times \frac{r}{b} + kb \times (1 - \frac{r}{b}) = kr + r + kb - kr = kb + r = a$$

می‌توان نشان داد تنها aای که در ۱ تعریف شده است در ۲ صدق می‌کند. در جدولهای یک و دو بعدی مسأله گردکردن با مشکل خاصی روبرو نمی‌شود اما در ابعاد بالاتر گاهی اوقات غیرممکن است. ذکر این نکته ضروری است که در حال حاضر در کشور ما با وجود ارائه اطلاعات آماری به صورت بسیار کلی باز هم می‌توان موارد زیادی از افشای دقیق یا تقریبی اطلاعات شخصی را به دست آورد. در حالی که در صورت اعمال روشهای ذکر شده در این مقاله بر روی داده‌های آماری می‌توان در عین ارایه مفصلتر اطلاعات آماری، جنبهٔ محرمانگی اطلاعات شخصی نیز را رعایت نمود.

۵ کاربرد

بخش صنعت در برنامه‌های توسعهٔ اقتصادی، اجتماعی و فرهنگی کشور به عنوان تأمین کننده‌ی بسیاری از نیازهای برنامه‌ی توسعه، جایگاه ویژه‌ای را به خود اختصاص داده است. بنابراین ضرورت شناخت دقیق، رفع مشکلات و نارسائی‌ها و انجام برنامه‌ریزی‌های آتی این بخش موجب شده است تا اطلاعات تفصیلی و به هنگام کارگاههای صنعتی در قالب زیر جامعه‌های ویژه‌ای جمع‌آوری و منتشر شوند.

طرح آمارگیری از کارگاههای صنعتی ده نفر کارکن و بیشتر[۱۶] هر ساله توسط مرکز آمار ایران و به صورت سرشماری در قالب همین سیاست صورت می‌پذیرد. جمع‌آوری اطلاعات دقیق در این بخش از اهمیت ویژه‌ای برخوردار است، لذا پاسخگویان باید مطمئن باشند که اطلاعاتی که توسط آنها ارایه می‌شود محفوظ بوده و امکان افشاء و استفاده‌های غیرمجاز از آنها وجود ندارد. تاکنون اطلاعات مربوط به کارگاههای صنعتی ده نفر کارکن و بیشتر برای حفظ محرمانگی اطلاعات فقط در سطح استان منتشر شده‌اند، در حالی که می‌توان هم زمان با انتشار اطلاعات در سطوح جغرافیائی کوچکتر، جنبه‌ی محرمانگی آنها را نیز رعایت کرد. در این بخش نشان داده خواهد شد که برخی اطلاعات در جدولهای انتشاراتی این آمارگیری قابل افشاء است. سپس با استفاده از نرم‌افزارهای ذیربط چگونگی ساختن جدول‌های ایمن را شرح خواهیم داد.

۱.۵ چند نمونه از اطلاعات قابل افشاء از جدولهای انتشاراتی

در این قسمت نمونه‌هایی از جدولهای انتشاراتی مرکز آمار ایران به‌عنوان بزرگترین تولید کننده‌ی آمار در کشور که دارای خطر افشای اطلاعات فردی هستند را به عنوان مثال ذکر می‌کنیم.

در ایران کمتر انتشار اطلاعات آماری به صورت داده‌های خرد صورت گرفته است. انتشارات جدولی نیز به صورت بسیار کلی صورت می‌گیرند به طوری که به عنوان مثال در انتشار اطلاعات مربوط به طرح آمارگیری از کارگاههای صنعتی ده نفر کارکن و بیشتر، متغیر ناحیه تنها در سطح کشور و استان ارایه می‌شود. با این حال، در همین سطح کلی نیز می‌توان موارد قابل افشایی را مشاهده کرد. در زیر دو نمونه از جدولهای ناایمن انتشار یافته توسط مرکز آمار ایران آمده است. در مطالعه مثالهای زیر توجه به دو نکته ضروری است:

الف) برای حفظ محرمانگی از بردن نام واحدی که هویت آن قابل افشاء است، خودداری شده است، و

ب) برای اختصار از انعکاس جدولهای انتشاراتی اجتناب شده و تنها به شماره جدولها از نشریه «نتایج آمارگیری از کارگاههای صنعتی ده نفر کارکن و بیشتر ۱۳۷۹» ارجاع داده شده است.

---

۱۶) مشخصات این طرح را می‌توان در نشریه «نتایج آمارگیری از کارگاههای صنعتی ۱۰ نفر کارکن و بیشتر ۱۳۷۹» از انتشارات مرکز آمار ایران یافت.

جدول ۵: قسمتی از جدول شماره ۱ نشریه نتایج آمارگیری از کارگاههای صنعتی ده نفر کارکن و بیشتر ۱۳۷۹

تعداد شاغلان	جمع	۱	۲	۳	۴	۵	۶	۷	۸
فعالیت									
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
تولید محصولات از توتون، تنباکو و سیگار	1	0	0	0	0	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
تولید وسایل نقلیه موتوری	14	0	0	0	1	0	4	4	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

مثال ۱. قسمتی از جدول شماره ۵ از نشریه نتایج آمارگیری از کارگاههای صنعتی ده نفر کارکن و بیشتر سال ۱۳۷۹ جمع‌آوری شده در سال ۱۳۸۰ که با عنوان «تعداد کارگاههای صنعتی ده نفر کارکن و بیشتر برحسب تعداد شاغلان و نوع فعالیت ۱۳۷۹:» آمده است، به صورت زیر است: در این جدول کدهای متغیر تعداد شاغلان به صورت زیر تعریف شده است: ۱:۱۹_۱۰ ۲:۲۹_۲۰ ۳:۳۹_۳۰ ۴:۴۹_۴۰ ۵:۹۹_۵۰ ۶:۴۹۹_۱۰۰ ۷:۹۹۹_۵۰۰ ۸:۱۰۰۰ و بیشتر.

این جدول یک جدول فراوانی است و همانطورکه از آن مشخص است تنها یک کارگاه تولید محصولات از توتون و تنباکو و سیگار در کل کشور وجود دارد که تعداد کارکنانش بیشتر از ۱۰۰۰ نفر است. به علت یکتا بودن این کارگاه در جامعه، به راحتی می‌توان از جدول شماره (۲) نشریه، عمومی بودن نحوه مدیریت و دولتی بودن وضعیت حقوقی کارگاه، از جدول شماره (۵) همان نشریه، تعداد دقیق شاغلان به تفکیک تولیدی یا غیرتولیدی بودن و مرد یا زن بودن آنها، از جدولهای شماره (۶) و (۷)، تعداد دقیق شاغلان با مزد و حقوق و بدون مزد و حقوق کارگاه به تفکیک تولیدی یا غیرتولیدی بودن و مرد یا زن بودن آنها، از جدول شماره (۸)، تعداد دقیق شاغلان به تفکیک سطح مهارت آنها (کارگر ساده، ماهر، تکنسین یا مهندس)، از جدول شماره (۱۱)، میزان مزد و حقوق و سایر پرداختیها به شاغلان، از جدول شماره (۱۷)، تعداد دقیق شاغلان به تفکیک سطح سواد (بیسواد، کمتر از دیپلم، دیپلم، لیسانس، فوق لیسانس یا دکترا)، از جدول شماره (۲۳)، میزان پرداختیهای کارگاه، از جدول شماره (۲۴)، ارزش داده‌های فعالیت صنعتی، از جدول شماره (۲۵)، ارزش ستانده‌های فعالیت صنعتی، از جدول شماره (۲۸)، ارزش سرمایه‌گذاریهای کارگاه، از جدول شماره (۲۹)، ارزش انواع موجودی انبار کارگاه و خلاصه تمام اطلاعات مربوط به این کارگاه خاص قابل دسترسی است. اگر رضایت این کارگاه برای انتشار اطلاعات خاص آن کسب نشده باشد، یک نمونه از افشای کامل اطلاعات فردی اتفاق افتاده است.

مثال ۲. این مثال در مورد کارگاههای تولید وسایل نقلیه موتوری است. در نگاه اول خطر افشانی احساس نمی‌شود، در حالی که این چنین نیست. به عنوان مثال در جدول شماره (۴)، تعداد شاغلان این ۱۴ کارگاه تولید وسایل نقلیه موتوری ۳۲۵۱۰ نفر ذکر شده است. منتشرکننده اطلاعات می‌داند که تعداد شاغلان بزرگترین کارگاه ۱۴۷۱۷ نفر، دومین کارگاه بزرگ ۵۱۱۳ نفر

و سومین کارگاه بزرگ ۲۷۶۱ نفر است که جمعا حدود ۷۷ درصد کل شاغلان این بخش را به خود اختصاص داده‌اند. در صورت اتحاد، این سه کارگاه می‌توانند تعداد شاغلان دیگر کارگاهها را برآورد نمایند. برای توضیح بیشتر، فرض کنید $x_{۱۴}, \ldots, x_۱$ مقادیر عرضه شده توسط ۱۴ کارگاه باشد. طبق قاعده پیشین-پسین (p, q) این خانه از جدول (۴) حساس است اگر دومین کارگاه بزرگ بتواند پس از انتشار جدول، مقدار صفت بزرگترین کارگاه را با اختلاف کمتر یا برابر p درصد از مقدار واقعی آن برآورد کند $(p < q)$. فرض کنید دومین پاسخگوی بزرگ قبل از انتشار جدول بتواند هر یک از مقادیر دیگر پاسخگویان را با اختلاف کمتر یا برابر ۵۰ درصد مقدار واقعی آنها برآورد کند $(q = ۵۰)$. در این صورت دومین کارگاه بزرگ پس از انتشار جدول می‌تواند پی ببرد که $x_۱$ حداکثر برابر

$$x_۱ + (\frac{q}{۱۰۰}) \sum_{i=۳}^{۱۴} x_i = ۱۷۴۱۷ + (\frac{۵۰}{۱۰۰}) \times ۹۹۸۰ = ۲۲۴۰۷$$

و حداقل برابر

$$x_۱ + (\frac{q}{۱۰۰}) \sum_{i=۳}^{۱۴} x_i = ۱۷۴۱۷ + (\frac{۵۰}{۱۰۰}) \times ۹۹۸۰ = ۱۲۴۲۷$$

است یعنی تعداد شاغلان بزرگترین کارگاه را در بازهٔ {۲۲۴۰۷ و ۱۲۴۲۷} برآورد نماید که اختلاف آن با مقدار واقعی کمتر از ۲۹ درصد است $(p = ۲۹)$.

به همین صورت در جدول شماره (۲۴) این نشریه، ارزش کل مواد خام و اولیه، لوازم بسته‌بندی، ابزار و وسایل کار کم‌دوام مصرف شده توسط این کارگاهها در خانه مربوط به آنها ۱۳۲۰۴۱۷۶ میلیون ریال ذکر شده است. با توجه به این که مقدار این متغیر برای بزرگترین کارگاه ۷۰۸۰۰۰۰ میلیون ریال، برای دومین کارگاه بزرگ ۳۳۰۰۰۰۰ میلیون ریال و برای سومین کارگاه بزرگ برابر ۶۲۶۰۰۰ میلیون ریال است، این سه کارگاه جمعا حدود ۸۳ درصد مقدار کل خانه را ارایه کرده‌اند. حال فرض کنید دومین کارگاه بزرگ قبل از انتشار جدول، میزان این متغیر برای هر کارگاه را بتواند با اختلاف کمتر یا برابر ۵۰ درصد مقدار واقعی آنها برآورد کند، در اینصورت پس از انتشار جدول طبق مطالب مطرح شده در بالا، این کارگاه مقدار این متغیر برای بزرگترین کارگاه را می‌تواند در بازه {۸۲۸۲۰۸۸ و ۵۸۷۷۹۱۲} یعنی با تفاوتی کمتر از ۱۷ درصد مقدار واقعی برآورد نماید.

۲.۵ ایمن‌سازی اطلاعات آماری انتشاراتی

سازمان‌های ملی آمار کشورها داده‌های حاصل از آمارگیریها را یا به صورت داده‌های خرد و یا به صورت داده‌های جدولی منتشر می‌کنند. همانگونه که اشاره شد در هر صورت امکان افشاء وجود دارد. به منظور ایمن‌سازی داده‌های خرد از نرم‌افزار $\mu ARGUS$ و برای ایمن‌سازی داده‌های

جدولی از نرم‌افزار $\tau_A RGUS$ که حاصل پروژه $CASC$[۱۷] است استفاده می‌شود. این پروژه با هدف تهیه ابزارهای عملی برای کنترل افشای آماری توسط تعدادی از کشورهای اتحادیه اروپایی، سازماندهی و اجرا شده است.

با توجه به اینکه داده‌های سرشماری از کارگاههای صنعتی دارای ده نفر کارکن و بیشتر به صورت جدولهای پیشایندی منتشر می‌شوند، لذا در این بخش به نحوه ایمن‌سازی جدولهای انتشاراتی طرح مذکور در استان تهران در سال ۱۳۷۹ با استفاده از نرم‌افزار $\tau_A RGUS$ پرداخته می‌شود.

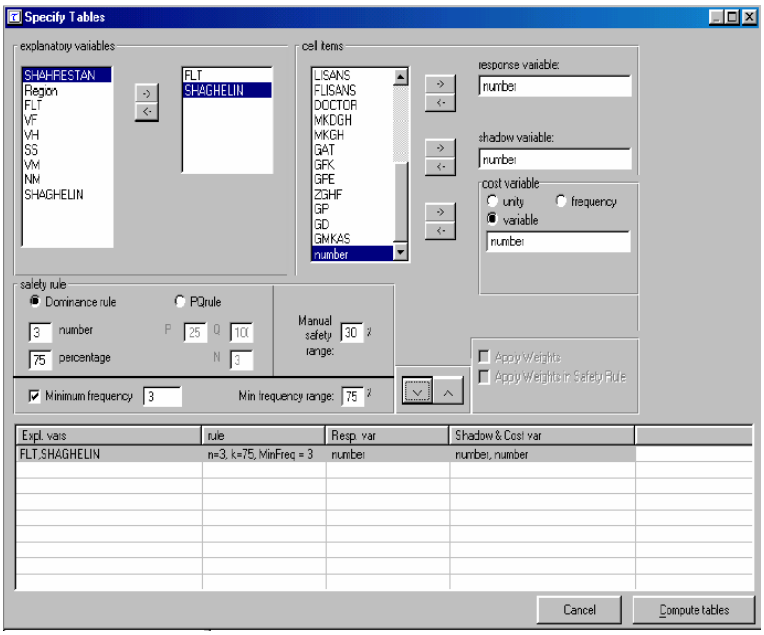
رایجترین شکل انتشار داده‌های آماری، جدولهای پیشایندی هستند، که خود به دو نوع جدولهای فراوانی و جدولهای مقداری تقسیم می‌شوند. اگر چه اطلاعات جدولها، خلاصه شده هستند ولی با توجه به نوع جدول و اعداد خانه‌ها، امکان افشاء وجود دارد. به منظور ساختن جدولهای انتشاراتی ایمن، به فایل داده‌های خرد و فایل فراداده حاوی مشخصات متغیرهای موجود در فایل داده‌های خرد ـ نیاز است. هر دو فایل به صورت متن هستند. در فایل داده‌های خرد هر سطر اصلی شامل نام یک متغیر، نقطه شروع، طول و مقدار گمشدهی متغیر مشخص می‌شود. در سطرهای بعد مشخصات متغیر وجود دارند. در صورتی که فایل فراداده وجود نداشته باشد، نرم‌افزار امکان ایجاد آن را دارد. پس از آن که فایل فراداده توسط نرم‌افزار $\tau_A RGUS$ خوانده شد، می‌توان جدولهایی را که قصد حفاظت آنها در برابر خطر افشاء وجود دارد، مشخص کرد. بدین منظور نرم‌افزار پنجره زیر را برای مشخص کردن جدولهای درخواستی باز می‌کند. در قسمت بالای پنجره شکل ۱، دو قسمت قابل مشاهده است. در یک قسمت، متغیرهای تبیینی (explanatory variables) که برای ساختن جدول به کار می‌روند، مشخص می‌شوند و در قسمت دیگر متغیر پاسخ (response variable) که برای محاسبه جمع‌های خانه‌ای مورد استفاده قرار می‌گیرد، تعیین می‌شود. گزینه‌ی Shadow Variable متغیری است که برای به کار بردن قاعده‌ی تسلط استفاده می‌شود. این متغیر در بیشتر اوقات همان متغیر پاسخ است اما می‌توان متغیر دیگری را نیز انتخاب کرد.

گزینه‌ی Cost Variable نشان دهنده متغیر مورد استفاده برای محاسبه‌ی مینیمم فقدان اطلاعات ناشی از اعمال روشهای کنترل افشاء است. این متغیر لزوما، همان متغیر پاسخ نیست بلکه می‌تواند متغیر تعداد پاسخگویان عرضه کننده‌ی مقدار خانه‌های جدول باشد. در قسمت پائین پنجره پارامترهای قاعدهی تسلط موردنظر را می‌توان وارد نمود. در اینجا دو قاعده‌ی تسلط آورده شده است. قاعده‌ی تسلط (n,k) و قاعده‌ی پیشین-پسین (p,q) که یکی از آن دو را می‌توان انتخاب کرد. باید توجه کرد که قاعده‌ی pـ درصد همان قاعده‌ی پیشن-پسین (p,q) درصد با در نظر گرفتن $q = ۱۰۰$ است. در هر صورت نرم‌افزار بازه‌ی حفاظت را تعیین می‌کند.

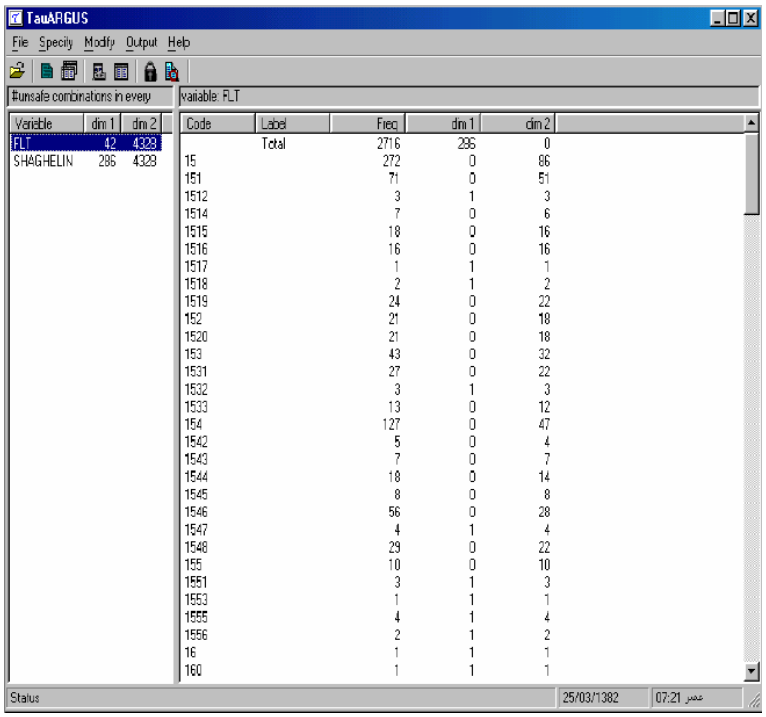
در حالتی که جدولهای انتشاراتی فراوانی هستند، حفاظت‌کننده‌ی جدول ممکن است به سادگی خانه‌هایی که دارای فراوانی‌ای کمتر از یک مقدار آغازین معین هستند را به عنوان خانه‌های ناایمن در نظر بگیرد.

در اینجا به عنوان نمونه جدول تعداد کارگاهها به تفکیک نوع فعالیت و تعداد شاغلان درخواست

17) Computational Aspects of Statistical Confidentiality



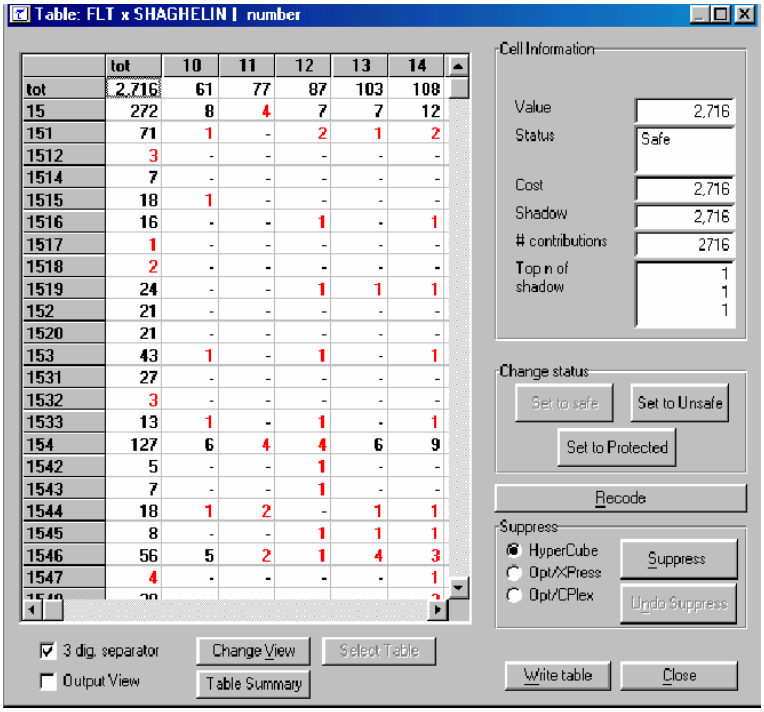
شکل ۱: پنجره‌ی انتخاب جدول



شکل ۲: پنجره نشان‌دهنده‌ی تعداد خانه‌های ناایمن به تفکیک رسته‌های متغیرهای تبیینی

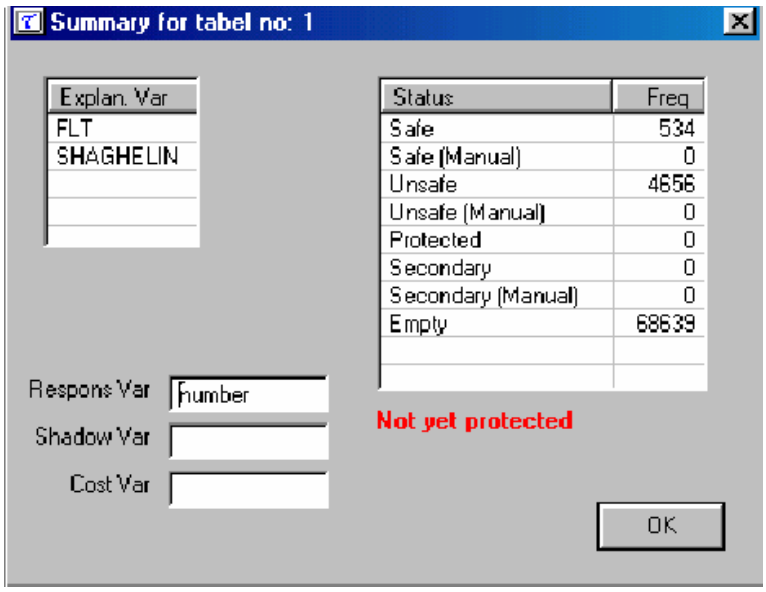
شده است. خانه‌های دارای چهار پاسخگو و کمتر ناایمن در نظر گرفته شده‌اند. پس از اینکه جدول و قواعد حفاظت آن مشخص شدند، نرم‌افزار مشخصات جدول مورد نظر را نشان می‌دهد که در آن، تعداد خانه‌های ناایمن ساخته شده توسط رسته‌های متغیرهای تبیینی انتخاب شده در ابعاد جدول و ابعاد پائینتر نمایش داده می‌شود.(شکل ۲) در نمونه‌ی بالا، جدول درخواستی دارای ۴۳۲۸ خانه‌ی ناایمن است (این تعداد از خانه‌ها دارای فراوانی کمتر از چهار هستند). تعداد زیادی از این خانه‌های ناایمن به دلیل تفصیل زیاد متغیرها حاصل شده‌اند.

با توجه به تعداد خانه‌های ناایمن هر متغیر در ابعاد مختلف می‌توان در مورد متغیر نامزد برای انجام بازکدگذاری تصمیم گرفت. بازکدگذاری یک متغیر موجب جزئیات کمتر و در عین حال تعداد ترکیبها ناایمن کمتر می‌گردد. نرم‌افزار جدول درخواستی را به‌صورت زیر نمایش می‌دهد.(شکل ۳) در سمت چپ پنجره، جدول درخواست شده مشاهده می‌شود. خانه‌های ناایمن با رنگ قرمز، خانه‌های ایمن با رنگ سیاه، خانه‌های پنهان‌سازی مکمل با رنگ آبی و خانه‌های خالی با علامت – مشخص شده‌اند (در اینجا به دلیل سیاه و سفید بودن چاپگر، رنگها مشخص نیستند). در پائین پنجره چند گزینه مشاهده می‌شود. با انتخاب گزینه‌ی Change View می‌توان جای



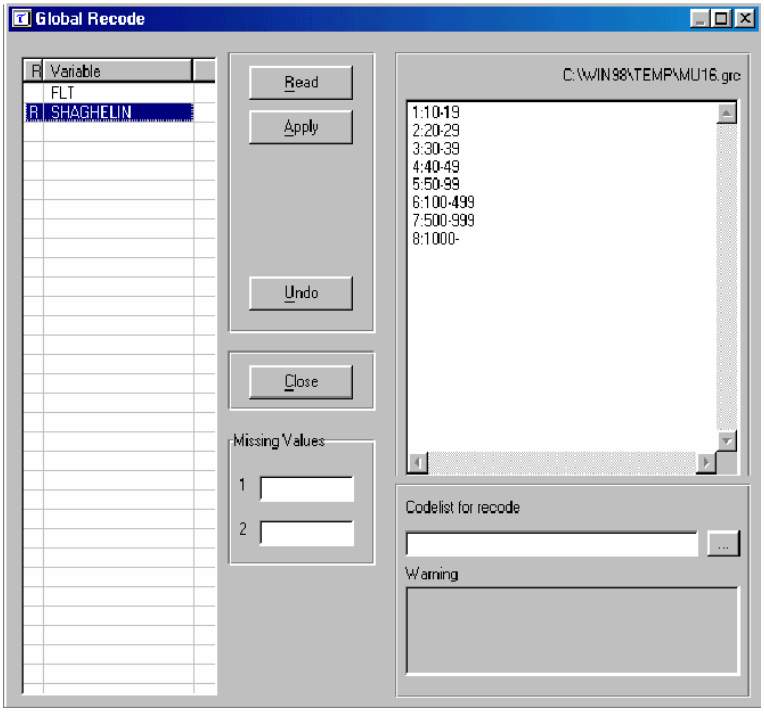
	tot	10	11	12	13	14
tot	2,716	61	77	87	103	108
15	272	8	4	7	7	12
151	71	1	-	2	1	2
1512	3	-	-	-	-	-
1514	7	-	-	-	-	-
1515	18	1	-	-	-	-
1516	16	-	-	1	-	1
1517	1	-	-	-	-	-
1518	2	-	-	-	-	-
1519	24	-	-	1	1	1
152	21	-	-	-	-	-
1520	21	-	-	-	-	-
153	43	1	-	1	-	1
1531	27	-	-	-	-	-
1532	3	-	-	-	-	-
1533	13	1	-	1	-	1
154	127	6	4	4	6	9
1542	5	-	-	1	-	-
1543	7	-	-	1	-	-
1544	18	1	2	-	1	1
1545	8	-	-	1	1	1
1546	56	5	2	1	4	3
1547	4	-	-	-	-	1

شکل ۳: پنجره‌ی نشان دهنده‌ی جدول درخواست شده

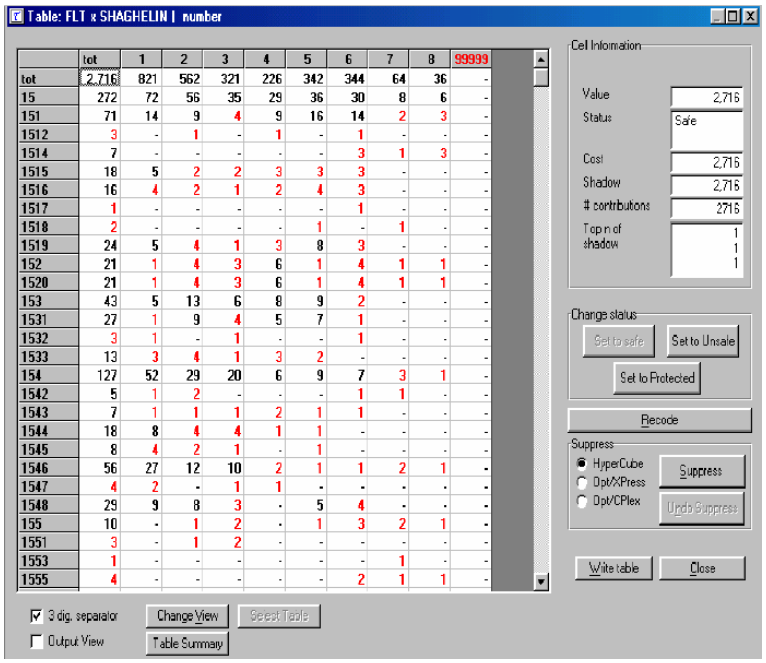


شکل ۴: پنجره‌ی نشان دهنده‌ی وضعیت جدول درخواست شده (قبل از تعدیل)

متغیرهای سطری و ستونی جدول را عوض نمود. اگر جدول دارای بیشتر از دو بعد باشد، متغیر سوم به صورت لایه‌ای نشان داده می‌شود. گزینه‌ی Summary Table چشم‌اندازی از تعداد خانه‌های ایمن و ناایمن بدست می‌دهد. در این حالت تعداد ۵۳۴ خانه ایمن، ۴۶۵۶ خانه ناایمن اولیه و ۶۸۶۳۹ خانه خالی وجود دارد (شکل ۴). در قسمت بالای سمت راست شکل ۴ می‌توان اطلاعات مربوط به خانه‌ی انتخاب شده‌ی جدول را مشاهده نمود. در قسمت پائین‌تر گزینه‌هایی برای تغییر وضعیت خانه یا خانه‌هایی معین وجود دارد. اگر بنا به دلایلی مایل باشیم علاوه بر قاعده‌ی ایمنی به کار رفته، خانه‌ای را از حالت ایمن به ناایمن یا برعکس تبدیل کنیم، می‌توان از این قسمت استفاده کرد. اکنون می‌توان این جدول را به لحاظ محرمانگی حفاظت کرد. یکی از روشهای قابل به کارگیری توسط این نرم‌افزار، روش بازکدگذاری است. با انتخاب گزینه‌ی Recode پنجره‌ای مانند پنجره‌ی زیر باز خواهد شد.(شکل ۵) در سمت چپ نام متغیرهای تبیینی آمده است که می‌توان متغیر مورد نظر را انتخاب کرد. به عنوان مثال در شکل زیر متغیر تعداد شاغلان انتخاب شده است. اگر متغیر مورد نظر مانند این متغیر انتخاب شده، ناسلسله مراتبی باشد، قسمت سمت راست پنجره حالت ویرایش کدها را خواهد داشت. بدین صورت که رسته‌های جدید متغیر که شامل یک یا چند رسته قبلی‌اند، نوشته می‌شوند. می‌توان این بازکدگذاری را در فایل جداگانه‌ای نوشته و در اینجا توسط گزینه‌ی Read آن را بازخوانی نمود. با انتخاب گزینه‌ی Apply کدهای جدید روی متغیر به کار می‌روند. به عنوان مثال



شکل ۵: پنجره‌ی انتخاب روش بازکدگذاری برای متغیر تعداد شاغلان



شکل ۶: پنجره‌ی نشاندهنده‌ی جدول درخواست شده (پس از بازکدگذاری متغیر تعداد شاغلان)

نتیجه‌ی بازکدگذاری روی متغیر تعداد شاغلان را در شکل بعد می‌بینیم.(شکل ۶) با این عمل تعداد خانه‌های ایمن به تعداد ۶۱۱ خانه افزایش می‌یابد و تعداد خانه‌های ناایمن اولیه به ۶۲۴ و خانه‌های خالی به ۷۵۵ خانه کاهش می‌یابند.

اگر متغیر انتخابی از نوع سلسله مراتبی باشد در پنجره باز شده، کدها به صورت درختی در سمت راست نشان داده می‌شوند.

در این پنجره توسط گزینه‌ی Maximum level می‌توان میزان ماکسیمم جزئیات این متغیر را تعیین نمود. مثلا برای متغیر نوع فعالیت ابتدا سطح دو در نظر گرفته می‌شود. نتیجه، در جدولی ظاهر می‌شود. در این حالت تعداد خانه‌های ایمن به تعداد ۳۴۰، خانه‌های ناایمن اولیه به ۲۰۵ و تعداد خانه‌های خالی به ۲۳۵ خانه تغییر می‌یابند.

در صورتیکه نتیجه رضایتبخش نباشد، می‌توان دوباره به مرحله‌ی بازکدگذاری برگشت و کدگذاری جدیدی را انتخاب نمود. به عنوان مثال دوباره روی متغیر نوع فعالیت بازکدگذاری انجام می‌دهیم و یک رقم دیگر از کد آنرا حذف می‌کنیم که در نتیجه تعداد خانه‌های ایمن، خانه‌های ناایمن اولیه و خانه‌های خالی به ترتیب به تعداد ۱۳۹، ۴۶ و ۶۵ خانه تغییر می‌یابند. به علت اینکه هنوز تعداد خانه‌های ناایمن زیاد است در صورت لزوم می‌توان باز هم از مقدار جزئیات

کاست. به عنوان مثال اگر رسته‌های ۷ و ۸ متغیر تعداد شاغلان با هم ترکیب می‌شوند، تعداد خانه‌های ایمن به ۱۳۶، خانه‌های ناایمن اولیه به ۳۵ و خانه‌های خالی به ۵۴ خانه تغییر می‌یابند.

جدول ۶ نتیجه اعمال مراحل بالا است. در این جدول، با توجه به قاعده‌ی تشخیص خانه‌های ناایمن استفاده شده، خانه‌های ناایمن اولیه به صورت خطدار مشخص شده‌اند.
پس از پایان بازکدگذاری متغیرها، باقیمانده‌ی خانه‌های ناایمن را می‌توان توسط روش پنهان‌سازی محافظت نمود. در این مرحله به منظور حفاظت خانه‌های ناایمن باید تعدادی خانه‌ی اضافه به عنوان پنهان‌سازی مکمل انتخاب شوند.

۶ نتیجه‌گیری

یکی از مهمترین وظایف نظام‌های آماری کشورها تولید آمارهای مورد نیازکاربران آنها با مشخصات دقیق بودن، به موقع بودن، پوشش کامل به لحاظ موضوعی و جغرافیائی داشتن، ارزان بودن، و به نحو مناسب اطلاع رسانی شدن، است. به منظور حفظ محرمانگی اطلاعات فردی، نظام‌های آماری موظفند که به شیوهای اطلاع رسانی آماری کنند که هویت افراد افشاء نشود. این امر که غالبا جنبه قانونی نیز دارد، بهانه‌ای شده است که برخی از سازمانهای ملی آمار که وظیفه تولید و اطلاع رسانی آماری دارند از انتشار داده‌های خرد جلوگیری کنند و یا داده‌های حاصل از آمارگیریها را به گونه‌ای منتشر کنند که از آنها هویت افراد قابل افشاء باشد.

به منظور انتشار ایمن داده‌ها چه به صورت خرد و چه به صورت جدول می‌توان از روش‌های کنترل افشاء داده‌های آماری استفاده نمود. این روش‌ها هم از افشای هویتها جلوگیری می‌کند و هم به سازمان‌های ملی آمار اجازه می‌دهد تا به صورت گسترده به اطلاع رسانی آماری مبادرت ورزند. برای این منظور لازم است از نرم‌افزارهای موجود همانند $\mu - ARGUS$ و $\tau - ARGUS$ استفاده شود.

جدول ۶: تعداد کارگاهها به تفکیک تعداد شاغلان و نوع فعالیت

فعالیت / شاغل	«کل»	«۱»	«۲»	«۳»	«۴»	«۵»	«۶»	«۷»	«۹۹۹۹»
«کل»	۲۷۱۶	۸۲۱	۵۶۲	۳۲۱	۲۲۶	۳۴۲	۳۴۴	۱۰۰	-
«۱۵»	۲۷۲	۷۲	۵۶	۳۵	۲۹	۳۶	۳۰	۱۴	-
«۱۶»	۱	-	-	-	-	-	-	۱	-
«۱۷»	۱۶۹	۷۲	۳۰	۸	۱۴	۱۵	۱۹	۱۱	-
«۱۸»	۶۹	۳۶	۱۴	۵	۵	۴	۵	-	-
«۱۹»	۷۲	۲۵	۱۱	۴	۷	۵	۱۵	۵	-
«۲۰»	۲۱	۱۴	۲	۱	۱	۱	۲	-	-
«۲۱»	۵۲	۱۲	۱۱	۹	۴	۸	۸	-	-
«۲۲»	۱۲۴	۵۹	۲۴	۹	۷	۱۱	۱۰	۴	-
«۲۳»	۱۴	۴	۳	۳	-	۲	-	۳	-
«۲۴»	۲۲۲	۵۱	۳۸	۳۱	۱۴	۳۵	۴۷	۶	-
«۲۵»	۱۶۰	۶۶	۳۳	۱۴	۱۰	۲۲	۱۱	۴	-
«۲۶»	۳۶۹	۸۰	۱۲۲	۷۵	۲۹	۲۳	۳۲	۸	-
«۲۷»	۷۳	۱۷	۱۳	۱۳	۳	۸	۱۳	۶	-
«۲۸»	۲۶۱	۹۶	۴۶	۱۹	۲۴	۳۸	۳۱	۷	-
«۲۹»	۳۴۶	۸۵	۷۵	۴۲	۴۵	۴۹	۴۴	۱۱	-
«۳۰»	۳۲	۹	۴	۶	۴	۶	۳	-	-
«۳۱»	۱۱۷	۲۹	۱۸	۱۳	۱۱	۲۵	۱۸	۳	-
«۳۲»	۳۹	۱۲	۴	۲	-	۵	۱۴	۲	-
«۳۳»	۵۰	۱۴	۹	۶	۴	۱۰	۵	۲	-
«۳۴»	۱۲۲	۱۷	۲۱	۱۲	۹	۲۵	۲۵	۱۳	-
«۳۵»	۱۴	۲	۲	-	۳	۲	۵	-	-
«۳۶»	۱۱۵	۵۳	۲۷	۱۳	۳	۱۲	۷	-	-
«۳۷»	۲	۱	-	۱	-	-	-	-	-
«۹۹۹۹»	-	-	-	-	-	-	-	-	-

مراجع

[۱] محمد بردبار عشرت‌آبادی (۱۳۸۲)، رساله‌ی پایان‌نامه کارشناسی ارشد، فنون محدودسازی افشای داده‌های آماری و کاربرد آنها، دانشگاه علامه طباطبایی.

[2] Cox, L. H. (1980), "Suppression Methodology and Statistical Disclosure Control," Journal of the American Statistical Association, Vol. 75, pp. 337-385

[3] Cox L. H. (1987), "A Constructive Procedure for Unbiased Controlled Rounding, " Journal of the American Statistical Association, Vol. 82, pp. 520-524.

[4] Fienberg, S. E., Markov, U. G., and Steel, R. J. (1998), "Disclosure Limitation using Perturbation and related Methods for Categorical Data," Journal of official statistics, Vol. 14, pp. 485-502.

[5] Gouweleeuw, J. M., Kooiman, p., Willenborg, L. C. R. J., and De Wolf, P. P. (1998), "Post Randomisation for Statistical Disclosure Control: Theory and Implementation," Journal of Official Statistics, Vol. 14, pp. 463-474.

[6] Kooiman, P., Willenborg, L.C.R.J. and Gouweleeuw, J.M. (1997), PRAM: A Method for Disclosure Limitation of Microdata, Statistics Netherlands, The Netherlands, Research Paper no. 97 ∘ 5

[7] Willenborg, L. C. R. J. and De Waal, T. (1998), Statistical Disclosure Control in Practice, Springer- Verlag, New York.

برآورد واریانس به روش جک نایف[1] در آمارگیریهای چندچارچوبی[2]

مرجان نورینی

مرکز آمار ایران

چکیده: روشهای معمول برآورد واریانس از قبیل روش بسط سری تیلور(روش دلتا[3]) در آمارگیریهای چندچارچوبی عموماً مستلزم محاسبه مشتقات جزئی بوده و این محاسبات با افزایش تعداد چارچوبها پیچیده‌تر می‌شود. برآورد واریانس به روش جک نایف روش دیگری است که ضمن سهولت در محاسبه، موجب کاهش چشمگیری در اریبی می‌شود. در این مقاله به معرفی برآوردگرهای چندچارچوبی، استفاده از روش جک نایف در برآورد واریانس برآوردگرهای چندچارچوبی و مقایسه آن با روش بسط سری تیلور طی یک مطالعه شبیه‌سازی می‌پردازیم.

واژه‌های کلیدی: بررسی‌های پیچیده[4]، برآورد خطی واریانس[5]، روشهای باز نمونه‌گیری[6]

۱ مقدمه

چارچوب آماری یا فهرست واحدهای آمارگیری، اساس و مبنای یک طرح آمارگیری نمونه‌ای را تشکیل می‌دهد. گاه ممکن است چارچوبی که کلیهٔ واحدهای جامعهٔ مورد مطالعه را پوشش دهد در دسترس نباشد، اما امکان دستیابی به پوشش کامل، با تلفیقی از دو یا چند چارچوب فراهم شود. در چنین حالتی به منظور دسترسی به پوشش مناسب، از دو چارچوب یا بیشتر بطور همزمان استفاده می‌شود. گاهی نیز ممکن است یک چارچوب، پوشش کامل را برای جامعهٔ مورد مطالعه فراهم کند، اما چارچوب ناقص دیگری موجود باشد که هزینهٔ آمارگیری از آن کمتر از هزینهٔ آمارگیری از چارچوب کامل باشد. در این شرایط به دلیل پایین‌تر بودن هزینهٔ آمارگیری از این چارچوب، می‌توان با هزینه‌ای مشخص و ثابت، از دو چارچوب استفاده کرده و حجم نمونه را بزرگتر و کارایی را افزایش داد. گاهی نیز ممکن است یک چارچوب فهرستی کامل در دسترس باشد اما عملاً با گذشت زمانی نسبتاً طولانی بدلیل بروز تغییرات فراوان در آن، منبعی برای بروز خطاهای غیر نمونه‌گیری شود. از آنجا که یک فهرست ناحیه‌ای کمتر در معرض تغییرات می‌باشد ترکیب آن با یک چارچوب از اعضای جامعه که احتمالاً ناقص باشد می‌تواند نتایج مفیدی را حاصل نماید. چنین آمارگیریهایی تحت عنوان آمارگیریهای چندچارچوبی بکار می‌روند.

1) Jackknife 2) Multiple Frame Surveys 3) Delta Method 4) Complex Surveys 5) Linearization Varince Estimation 6) Resampling Mthods

شاید بتوان گفت اولین شالوده آمارگیریهای چندچارچوبی در سال (۱۹۴۹) با آمارگیری از فروشگاههای خرده فروشی که مجری آن دفتر سرشماری آمریکا بود گذاشته شد. سپس هارتلی[7] (۱۹۶۲) نظریه مقدماتی چارچوبهای چندگانه را توسعه داد. او با فرض اینکه اجتماع چارچوبها، جامعه را پوشش می‌دهد، واحدهای جامعه را به زیرمجموعه‌های دو به دو ناسازگار شامل اجتماعها و اشتراکهای چارچوبهای مختلف تقسیم کرد. بعد از وی لاند[8] (۱۹۶۸)، فولر و بورمیستر[9] (۱۹۷۲)، وگل[10] (۱۹۷۵)، فوردوبوسکر[11] (۱۹۷۶)، بانکیر[12] (۱۹۸۶)، اسکینر[13] (۱۹۹۱)، اسکینر و رائو[14] (۱۹۹۶) و کالتون و اندرسون[15] (۱۹۸۶) مقالاتی را در این زمینه ارائه نموده‌اند.

در آمارگیریهای چندچارچوبی، برآوردگرهای مختلفی برای برآورد مجموع جامعه پیشنهاد می‌شود. این برآوردگرها عموماً توابعی غیر خطی از مقادیر نمونه‌ای می‌باشند. برای برآورد واریانس این گونه برآوردگرها، روشهای مختلفی از جمله بسط سری تیلور(خطی سازی) پیشنهاد شده است.

اسکینر و رائو (۱۹۹۶) روش بسط سری تیلور را پیشنهاد دادند. استفاده از این روش مستلزم محاسبه مشتقات جزئی بوده و این محاسبات با افزایش تعداد چارچوبها پیچیده‌تر می‌شود. لیکن برآورد واریانس به روش جک نایف ضمن سهولت در محاسبه، موجب کاهش چشمگیری در اریبی می‌شود. در این مقاله ضمن استفاده از روش جک نایف در برآورد واریانس به مقایسه آن با روش بسط سری تیلور می‌پردازیم. بدین منظور در بخش ۲ برآوردگرهای چندچارچوبی را معرفی می‌کنیم. در بخش ۳ ابتدا شرح مختصری از روش جک نایف را ارائه داده و سپس به استفاده از آن در آمارگیریهای چندچارچوبی می‌پردازیم. در بخش ۴ نتایج حاصل از مقایسه و بررسی برآوردگرها و برآوردگرهای واریانس از طریق شبیه‌سازی ارائه می‌گردد.

۲ برآوردگرهای مجموع جامعه

برای سادگی فرض کنید دو چارچوب A و B موجود است که هر دو ناقص و دارای واحدهای مشترک با یکدیگر هستند بطوریکه مجموع آنها رویهم، کل جامعه مورد مطالعه را پوشش می‌دهد. از چارچوبهای A و B، سه $(۱-۲^۲)$حوزه[16] دو به دو ناسازگار به دست می‌آید.

حوزه a : شامل واحدهایی است که فقط در چارچوب A می‌باشند. $a = A \cap B^c$

حوزه b : شامل واحدهایی است که فقط در چارچوب B می‌باشند. $b = A^c \cap B$

حوزه ab : شامل واحدهایی است که در هر دو چارچوب می باشند. $ab = A \cap B$

(c نشان دهندهٔ مکمل مجموعه می‌باشد)

N_A و N_B تعداد واحدها در چارچوبهای A و B، N_a و N_b و N_{ab} تعداد واحدهای موجود در حوزه‌های a، b و ab می‌باشند.

---

7) Hartley 8) Lund 9) Fuller and Burmeister 10) Vogel 11) Ford and Boseker 12) Bankier 13) skinner 14) Rao 15) Kalton and Anderson 16) domain

دو نمونه مستقل s_A و s_B بر اساس طرحهای نمونه‌گیری احتمالی $p_A(s_A)$ و $p_B(s_B)$ به اندازه n_A و n_B از دو چارچوب فوق گرفته می‌شود. احتمال شمول نمونه حاصل از چارچوب A، $\pi_i^A = p\{i \in s_A\}$ و احتمال شمول نمونه حاصل از چارچوب B، $\pi_i^B = p\{i \in s_B\}$ می‌باشد. براساس نمونه‌های مستقل فوق، n_a و n_{ab}^A، تعداد واحدهای نمونه‌گیری حاصل از چارچوب A می‌باشند که به ترتیب در حوزه‌های a و ab ، قرار دارند. به همین ترتیب n_b و n_{ab}^B نیز تعداد واحدهای نمونه‌گیری حاصل از چارچوب B می‌باشند که به ترتیب در حوزه‌های b و ab هستند. با فرض اینکه Y_a، Y_b، Y_{ab}، μ_a، μ_b و μ_{ab} به ترتیب مجموع و میانگین جامعه در حوزه‌های a، b، ab باشند، داریم:

$$Y = Y_a + Y_{ab} + Y_b \qquad (۱)$$

هدف برآورد Y است.

چندین برآوردگر نقطه‌ای تحت عنوان برآوردگر دو چارچوبی برای برآورد Y پیشنهاد شده که همگی آنها به فرم $\hat{Y} = \hat{Y}_a + \hat{Y}_{ab} + \hat{Y}_b$ می‌باشند. هر یک از این برآوردگرها بسته به اینکه اطلاعات حاصل از دو نمونه برای برآورد Y چگونه با هم ترکیب می‌شوند با هم تفاوتهایی دارند. با فرض اینکه N_A و N_B معلوم بوده و $N_a > ۰$ و $N_b > ۰$، وزنهای نمونه‌گیری w_i^A و w_i^B عبارتند از:

$$\begin{aligned} w_i^A &= N_A\,[\pi_i^A \sum_{j \in s_A} (\frac{۱}{\pi_j^A})]^{-۱} \\ w_i^B &= N_B\,[\pi_i^B \sum_{j \in s_B} (\frac{۱}{\pi_j^B})]^{-۱} \end{aligned}$$

دو متغیر نشانگر زیر را برای هر یک از چارچوبهای A و B در نظر می‌گیریم:

$$\delta_A(i) = \begin{cases} ۱ & \text{واحد } i \text{ متعلق به چارچوب } A \text{ باشد} \\ ۰ & \text{در غیر این صورت} \end{cases}$$

$$\delta_B(i) = \begin{cases} ۱ & \text{واحد } i \text{ متعلق به چارچوب } B \text{ باشد} \\ ۰ & \text{در غیر این صورت} \end{cases}$$

در نتیجه برآوردگرهای مجموع در سه حوزهa و b و ab بقرار زیر می‌باشند:

$$\begin{aligned} \hat{Y}_a^A &= \textstyle\sum_{i \in s_A} w_i^A (۱ - \delta_B(i)) y_i \\ \hat{Y}_{ab}^A &= \textstyle\sum_{i \in s_A} w_i^A \delta_B(i) y_i \\ \hat{Y}_b^B &= \textstyle\sum_{i \in s_B} w_i^B (۱ - \delta_A(i)) y_i \\ \hat{Y}_{ab}^B &= \textstyle\sum_{i \in s_B} w_i^B \delta_A(i) y_i \end{aligned}$$

برآوردگر اندازهٔ هر یک از حوزه‌ها نیز بصورتی مشابه با قراردادن $y_i = ۱$ در تعاریف $\hat{Y}_a^A$ و $\hat{Y}_{ab}^A$ و $\hat{Y}_b^B$ و $\hat{Y}_{ab}^B$ بدست می‌آیند. همچنین دو برآوردگر دیگر را نیز بصورت زیر تعریف می‌کنیم:

$$\begin{aligned} \hat{Y}_{ab}(\theta) &= \theta\hat{Y}_{ab}^A + (۱-\theta)\hat{Y}_{ab}^B \\ \hat{N}_{ab}(\theta) &= \theta\hat{N}_{ab}^A + (۱-\theta)\hat{N}_{ab}^B \end{aligned}$$

۱.۲ برآوردگرهای هارتلی و فولر ـ بورمیستر

هارتلی در سال (۱۹۶۲) برآوردگر دوچارچوبی زیر را برای برآورد مجموع جامعه پیشنهاد کرد:

$$\hat{Y}_H(\theta) = \hat{Y}_a^A + \hat{Y}_b^B + \hat{Y}_{ab}(\theta) \qquad (۲)$$

فولرو بورمیستر هم در سال (۱۹۷۲) برآوردگر زیر را پیشنهاد دادند:

$$\hat{Y}_{FB}(\beta_۱, \beta_۲) = \hat{Y}_a^A + \hat{Y}_b^B + \hat{Y}_{ab}(\beta_۱) + \beta_۲(\hat{N}_{ab}^A - \hat{N}_{ab}^B) \qquad (۳)$$

مقادیر بهینه پارامترهای θ، $\beta_۱$ و $\beta_۲$ به ترتیب واریانس برآوردگرهای $\hat{Y}_H(\theta)$ و $\hat{Y}_{FB}(\beta_۱, \beta_۲)$ را می‌نیمم کرده و بنابراین به کواریانسهای $\hat{Y}_{ab}^A$ و $\hat{Y}_{ab}^B$ بستگی دارند. اما در عمل این کواریانسها مجهول بوده و باید از روی نمونه برآورد شوند. در نتیجه $\hat{Y}_H(\hat{\theta}_H)$ و $\hat{Y}_{FB}(\hat{\beta}_{FB})$ بطور کلی توابعی خطی از y با وزنهای یکسان برای همه متغیرها نخواهند بود و برای هر متغیر پاسخ وزنها باید جداگانه محاسبه شوند. برای مثال اگر $\hat{Y}_{H_۱}$ و $\hat{Y}_{H_۲}$ و $\hat{Y}_{H_۳}$ برآوردگرهای هارتلی برای تعداد کل بیمارانی که دچار تنگی نفس در گروههای سنی ۰ـ۱۶ و ۱۷ـ۴۵ و بالای ۴۵ سال باشد، آن گاه $\hat{Y}_{H_۱} + \hat{Y}_{H_۲} + \hat{Y}_{H_۳}$ لزوماً مساوی برآورد هارتلی تعداد کل بیماران دچار تنگی نفس در جامعه نخواهد بود. در نتیجه فولرو بورمیستر برآن شدند برآوردگری ارائه دهند که از یک مجموعه وزنهای یکسان برای همه متغیرهای پاسخ استفاده کند. اما برآوردگری که ارائه دادند تحت طرح نمونه‌گیری تصادفی ساده طراحی شده بود و مستقیماً با طرحهای نمونه‌گیری پیچیده بکار نمی‌رفت. چون عموماً بصورتی ناسازگار با این نوع نمونه‌گیریها طراحی شده بود. اسکینر (۱۹۹۱) از این برآوردگر تفسیری بصورت یک برآوردگر حداکثر درستنمایی داشت.

۲.۲ برآوردگر شبه حداکثر درستنمایی

اسکینر و رائو (۱۹۹۶) برآوردگری ارائه دادند که علاوه بر استفاده از یک مجموعه یکسانی از وزنها، اصلاح شدهٔ یک برآوردگر حداکثر درستنمایی تحت طرح نمونه‌گیری تصادفی ساده برای رسیدن به برآوردگری سازگار با طرحهای نمونه‌گیری پیچیده بود. آنها برآوردگر شبه حداکثر درستنمایی به

صورت زیر را ارائه دادند:

$$\hat{Y}_{PML}(\theta) = \frac{N_A - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_a^A}\hat{Y}_a^A + \hat{N}_{ab}^{PML}(\theta)\frac{\hat{Y}_{ab}(\theta)}{\hat{N}_{ab}(\theta)} + \frac{N_B - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_b^B}\hat{Y}_b^B \quad (۴)$$

بطوریکه $\hat{N}_{ab}^{PML}(\theta)$ تابعی از $\hat{N}_{ab}^A$ و $\hat{N}_{ab}^B$ و θ بوده وکوچکترین ریشه معادله درجه دو به فرم زیر است:

$$(\frac{\theta}{N_B} + \frac{۱-\theta}{N_A})x^۲ - (۱ + \frac{\theta}{N_B}\hat{N}_{ab}^A + \frac{۱-\theta}{N_A}\hat{N}_{ab}^B)x + \hat{N}_{ab}(\theta) = ۰$$

این برآوردگر بر خلاف برآوردگرهای هارتلی و فولر ـ بورمیستر، تابعی خطی از y است. اسکینر و رائو (۱۹۹۶) انتخاب $\theta = \theta_P$ را برای می‌نیمم کردن واریانس مجانبی $\hat{N}_{ab}^{PML}(\theta)$ با

$$\theta_P = \frac{N_a N_B V(\hat{N}_{ab}^B)}{N_a N_B V(\hat{N}_{ab}^B) + N_b N_A V(\hat{N}_{ab}^A)} \quad (۵)$$

پیشنهاد کردند.

در عمل N_a و N_b و واریانسها در رابطه (۵) مجهول بوده و باید از روی داده‌ها برآورد شود. در نتیجه برآوردگر حاصل عبارت از $\hat{Y}_{PML}(\hat{\theta}_P)$ است.

۳.۲ برآوردگرهای چارچوب منفرد

بانکیر (۱۹۸۶)، کالتون و اندرسون (۱۹۸۶) و اسکینر (۱۹۹۱) برای برآورد مجموع جامعه با در نظر گرفتن مشاهدات بگونه‌ای که آنها از یک چارچوب منفرد با وزنهای اصلاح شده برای مشاهدات در حوزه متداخل ab نمونه‌گیری شده‌اند، برآوردی ارائه دادند. وزنهای اصلاح شده برای برآوردگرهای تک چارچوبی اسکینر(۱۹۹۱) و کالتون و اندرسون (۱۹۸۶) به شناسایی واحدهای مشابه در نمونه‌ها نیازی ندارد. این وزنها عبارتند از:

$$w_i = \begin{cases} (\pi_i^A + \pi_i^B)^{-۱} & i \in ab \\ (\pi_i^A)^{-۱} & i \in a \\ (\pi_i^B)^{-۱} & i \in b \end{cases}$$

آنها برآوردگر زیر را برای برآورد مجموع جامعه پیشنهاد کردند:

$$\hat{Y}_{SF} = \sum_{i \in s_A} w_i y_i + \sum_{i \in s_B} w_i y_i \quad (۶)$$

همانطور که بانکیر (۱۹۸۶) اشاره کرد، این برآوردگر قابل تعمیم به بیش از دو چارچوب می‌باشد. این برآوردگر از هیچ گونه اطلاعات کمکی در ارتباط با N_A و N_B استفاده نمی کند. از آنجائیکه استفاده از اطلاعات کمکی باعث کاهش واریانس و افزایش دقت برآوردگر [بانکیر (۱۹۸۶)] می‌شود، اسکینر و رائو (۱۹۹۶) دو روش برای تصحیح این برآوردگر ارائه دادند که عبارتند از:
الف) روش تعدیل نسبتی[17] [بانکیر(۱۹۸۶)]
ب) روش رگرسیونی
اسکینر و رائو پیرو اثبات قضیه (۱) اسکینر (۱۹۹۱) نشان دادند که فرآیند تعدیل به برآوردگر تجربی زیر همگراست:

$$\hat{Y}_{SFrake} = \frac{N_A - \hat{N}_{ab}^{rake}}{\hat{N}_a}\hat{Y}_a + \frac{\hat{N}_{ab}^{rake}}{\hat{N}_{abS}}\hat{Y}_{abS} + \frac{N_B - \hat{N}_{ab}^{rake}}{\hat{N}_b}\hat{Y}_b \qquad (۷)$$

که در آن:

$$\begin{aligned} \hat{Y}_{abS} &= \textstyle\sum_{s_A} w_i \delta_i^B y_i + \sum_{s_B} w_i \delta_i^A y_i \\ \hat{N}_{abS} &= \textstyle\sum_{s_A} w_i \delta_i^B + \sum_{s_B} w_i \delta_i^A \end{aligned}$$

و $\hat{N}_{ab}^{rake}$ کوچکترین ریشه معادله درجه دو بفرم زیر می‌باشد:

$$\hat{N}_{abS}x^2 - [\hat{N}_{abS}(N_A + N_B) + \hat{N}_{aS}^A \hat{N}_{bS}^B]\; x \; + \; \hat{N}_{abS} N_A N_B = 0 \qquad (۸)$$

تصحیح با N_A و N_B معلوم از طریق روش رگرسیونی برآوردگر زیر را نتیجه می‌دهد:

$$\hat{Y}_{SFreg} = \hat{Y}_S + \hat{\boldsymbol{\beta}}^T [N_A - \hat{N}_S^A, N_B - \hat{N}_S^B] \qquad (۹)$$

بطوریکه مقدار بهینه‌ی $\hat{\boldsymbol{\beta}}$ عبارتست از:

$$\hat{\boldsymbol{\beta}}_S^T = -Cov\{[\hat{N}_{AS}/V(\hat{N}_{AS}), \hat{N}_{BS}/V(\hat{N}_{BS})], \hat{Y}_S\}$$

۳ برآورد واریانس

اسکینر و رائو (۱۹۹۶) با استفاده از بسط سری تیلور، روشی را برای برآورد واریانس $\hat{Y}_{PML}$ ارائه دادند. در این بخش ابتدا برآورد واریانس به روش جک نایف برای برآوردگرهای دو چارچوبی را ارائه داده و سپس نشان می‌دهیم که برآوردگر جک نایف واریانس بطور مجانبی هم ارز با برآوردگر خطی واریانس می‌باشد. برای سادگی حالت دوچارچوبی را در نظر می‌گیریم اما این نتایج براحتی قابل تعمیم به آمارگیریهای چندچارچوبی می‌باشد.

17) Raking Ratio

نخستین بار کوئنولی[18] (۱۹۵۹) جک نایف را به عنوان روشی بر اساس حذف هر بار یک مشاهده از مجموعه داده‌های اولیه و محاسبه مجدد برآوردگر با استفاده از بقیهٔ داده‌ها برای کاهش اریبی برآوردگر ضریب همبستگی پیاپی[19] معرفی نمود. در مقاله‌ای در سال (۱۹۵۶) این روش را تعمیم داده و خواص کلی کاهش اریبی آن را در یک جامعه متناهی بیان کرد. سپس توکی[20] (۱۹۵۸) پیشنهاد کرد که این روش علاوه بر کاهش اریبی، می‌تواند برای برآورد واریانس نیز استفاده شود.

فرض کنید چارچوب‌های A و B به ترتیب دارای H و L طبقه باشند بطوریکه طبقات h و l هر یک از آنها به ترتیب شامل N_h^A و N_l^B واحد بوده و در مجموع $\tilde{N}_h^A$ و $\tilde{N}_l^B$ واحد نمونه‌گیری اولیه[21] (psu) دارند که به ترتیب $\tilde{n}_h^A$ و $\tilde{n}_l^B$ واحد از آنها نمونه‌گیری می‌شود. بطوریکه $\tilde{n}_A = \sum_{h=1}^{H} \tilde{n}_h^A$ و $\tilde{n}_B = \sum_{l=1}^{L} \tilde{n}_l^B$ کل نمونه‌ای است که به ترتیب از چارچوبهای A و B انتخاب می‌شود. وزنهای طبقات h و l در چارچوبهای A و B را به ترتیب با $W_h^A = N_h^A/N_A$ و $W_h^B = N_l^B/N_B$ نشان می‌دهیم. لازم به یاددآوری است که اغلب در نمونه‌گیریهای چندمرحله‌ای، نمونهٔ بدون جایگذاری از واحدهای نمونه‌گیری اولیه با احتمال متناسب با اندازه (به دلیل کارایی بالای نمونه‌گیری بدون جایگذاری نسبت به نمونه‌گیری با جایگذاری) بعمل آمده و سپس در مرحله برآورد واریانس برای صرفه‌جویی در زمان و هزینه و اجتناب از محاسبات بالقوه پر زحمت احتمالات شمول توام، نمونه‌گیری را بگونه‌ای در نظر می‌گیرند که گویی با جایگذاری انتخاب شده است. این تقریب عموماً منجر به بیش برآورد واریانس برآوردگر مجموع می‌شود اما در صورتی که کسرهای نمونه‌گیری مرحله اول بزرگ نباشد، اریبی نسبی بزرگ نخواهد بود. برآوردگر واریانس جک نایف و دیگر برآوردگرهای واریانس بازنمونه‌گیری از این روش در برآورد واریانس استفاده می‌کنند.

حال فرض می‌کنیم که نمونه‌هایی به اندازه $\tilde{n}_h^A$ (≥ 2) و $\tilde{n}_l^B$ به ترتیب با احتمالات شمول متناسب با اندازه، $\tilde{\pi}_{hi}^A = \tilde{n}_h^A p_{hi}^A$ و $\tilde{\pi}_{lj}^B = \tilde{n}_l^B p_{lj}^B$ از چارچوبهای A و B انتخاب می شود. بطوریکه p_{hi}^A و p_{lj}^B به ترتیب برابر با احتمال انتخاب واحد iام و jام در طبقات h و l متناسب با اندازه psu ها بوده و $\sum_i p_{hi}^A = 1$ و $\sum_j p_{lj}^B = 1$) ($p_{hi}^A = u_i / \sum_A u_i$ و $p_{lj}^B = u_j / \sum_B u_j$ که u_i اندازه تقریبی i امین psu می باشد). با در نظر گرفتن دو بردار $\boldsymbol{A} = (Y_a, Y_{ab}, N_{ab}, N_A)^T$ و $\boldsymbol{B} = (Y_b, Y_{ab}, N_{ab}, N_B)^T$ برآورد $\boldsymbol{A}$ عبارت خواهد بود

18) Quenoulle 19) Serial Correlation Cofficient 20) Tukey 21) Primary Sampling Units

از:

$$\hat{\boldsymbol{A}} = \sum_{h=1}^{H}\sum_{i=1}^{\tilde{n}_h^A} \hat{\boldsymbol{A}}_{hi}/\tilde{\pi}_{hi}^A = \sum_{h=1}^{H}\sum_{i=1}^{\tilde{n}_h^A} \hat{A}_{hi}/\tilde{n}_h^A p_{hi}^A = \sum_{h=1}^{H}\sum_{i=1}^{\tilde{n}_h^A} N_h^A \boldsymbol{a}_{hi}/\tilde{n}_h^A = \sum_{h=1}^{H} N_h^A \bar{\boldsymbol{a}}_h$$

که $\hat{\boldsymbol{A}}_{hi}$ برآورد ناااریب بردار مجموع جوامع در iامین psu نمونه از طبقه hام بوده و $\boldsymbol{a}_{hi} = \hat{\boldsymbol{A}}_{hi}/(N_h^A p_{hi}^A)$ می‌باشد. برآوردگر $\hat{\boldsymbol{B}} = \sum_{l=1}^{L}\sum_{j=1}^{\tilde{n}_l^B} N_l^B \boldsymbol{b}_{lj}/\tilde{n}_l^B$ و $\boldsymbol{b}_{lj}$ نیز بصورتی مشابه تعریف می‌شوند. تحت فرضیات نمونه‌گیری با جایگذاری، $\boldsymbol{a}_{hi}$ها برآوردگرهای ناااریب مستقل و هم‌توزیع میانگین جامعه در طبقه hام چارچوب A می‌باشند در حالیکه برای $h \neq h'$ ، $\boldsymbol{a}_{hi}$ و $\boldsymbol{a}_{h'i'}$ مستقل بوده اما لزوماً هم توزیع نیستند. بطور مشابه $\boldsymbol{b}_{lj}$ها هم برآوردگرهای ناااریب مستقل و هم توزیع بردار میانگین جامعه در طبقه lام چارچوب B می‌باشند.

پارامتری به فرم $\tau = g(\bar{\boldsymbol{A}}, \bar{\boldsymbol{B}})$ که تابعی از میانگینهای جامعه $\bar{\boldsymbol{A}} = \boldsymbol{A}/N$ و $\bar{\boldsymbol{B}} = \boldsymbol{B}/N$ است را در نظر می‌گیریم. میانگینهای $\bar{\boldsymbol{A}} = \boldsymbol{A}/N$ و $\bar{\boldsymbol{B}} = \boldsymbol{B}/N$ بصورت $\hat{\bar{\boldsymbol{A}}} = \sum_{h=1}^{H} W_h^A \bar{\boldsymbol{a}}_h$ و $\hat{\bar{\boldsymbol{A}}} = \sum_{l=1}^{L} W_l^B \bar{\boldsymbol{b}}_l$ برآورد شده و τ نیز بصورت $\hat{\tau} = g(\hat{\bar{\boldsymbol{A}}}, \hat{\bar{\boldsymbol{B}}})$ برآورد می‌شود.

برآوردگرهای دوچارچوبی می‌توانند در قالب $\hat{\tau}$ با در نظر گرفتن مجموع جامعه به صورت $Y = N\bar{Y} = Ng(\bar{\boldsymbol{A}}, \bar{\boldsymbol{B}})$ بیان شوند که میانگین $\bar{Y}$ عبارتست از:

$$g(\bar{\boldsymbol{A}}, \bar{\boldsymbol{B}}) = \bar{Y} = (N_A/N)(\bar{A}_1 + \theta \bar{A}_2) + (N_B/N)(\bar{B}_1 + (1-\theta)\bar{B}_2) \quad (10)$$

در این صورت برآوردگر هارتلی عبارت از $\hat{Y}_H(\theta_H) = Ng(\hat{\bar{\boldsymbol{A}}}, \hat{\bar{\boldsymbol{B}}})$ خواهد بود. برآوردگر PML نیز به ازای θ ثابت، می‌تواند بصورتی مشابه بعنوان تابعی از $\hat{\bar{\boldsymbol{A}}}$ و $\hat{\bar{\boldsymbol{B}}}$ و با توجه به اینکه $\hat{N}_{ab}^{PML}(\theta)$ تابعی از $\hat{\bar{A}}_2$ و $\hat{\bar{B}}_3$ است، و سپس با استفاده از رابطه (۴) بیان شود.

ماتریس واریانس ـ کواریانس بردار $\hat{\bar{\boldsymbol{A}}}$ برابر با $\sum_A = \sum_{h=1}^{H} (W_h^A)^2 \sum_A^h / \tilde{n}_h^A$ بوده که با $\boldsymbol{S}_A = \sum_{h=1}^{H} (W_h^A)^2 \boldsymbol{S}_h^A / \tilde{n}_h^A$ برآورد می‌شود. بطور مشابه واریانس $\hat{\bar{\boldsymbol{B}}}$ یعنی $\sum_B$ نیز بصورت $\boldsymbol{S}_B = \sum_{l=1}^{L} (W_l^B)^2 \boldsymbol{S}_l^B / \tilde{n}_l^B$ برآورد شده که $\boldsymbol{S}_h^A$ و $\boldsymbol{S}_l^B$ به ترتیب برآوردگرهای

واریانس $\sqrt{\tilde{n}_l^B}\bar{\mathbf{b}}_l$ $\sqrt{\tilde{n}_h^A}\bar{\mathbf{a}}_h$ می‌باشند و عبارتند از:

$$\boldsymbol{S}_h^A = (\tilde{n}_h^A - ۱)\sum_{i=۱}^{\tilde{n}_h^A}(\boldsymbol{a}_{hi} - \bar{\boldsymbol{a}}_h)(\boldsymbol{a}_{hi} - \bar{\boldsymbol{a}}_h)^T$$

$$\boldsymbol{S}_l^B = (\tilde{n}_l^B - ۱)\sum_{j=۱}^{\tilde{n}_l^B}(\boldsymbol{b}_{lj} - \bar{\boldsymbol{b}}_l)(\boldsymbol{b}_{lj} - \bar{\boldsymbol{b}}_l)^T$$

برای اثبات سازگاری برآوردگر جک نایف واریانس و همچنین هم‌ارزی مجانبی آن با برآوردگر خطی واریانس به شروط زیر احتیاج داریم .

الف) فرض می‌کنیم به ازای هر h و l ، $W_h^A\tilde{n}_A/\tilde{n}_h^A = O(۱)$ و $W_l^B\tilde{n}_B/\tilde{n}_l^B = O(۱)$ باشد. در مجموع فرض می‌کنیم $\sum_h W_h^A\sum_h^A = O(۱)$ و $\sum_l W_l^B\sum_l^B = O(۱)$ باشد.

ب) فرض کنید $g_A(\tilde{\mathbf{a}},\tilde{\mathbf{b}})$ برداری q بعدی از مشتقهای اول تابع $g(.)$ با توجه به مولفه های $\mathbf{a}$ و $g_B(\tilde{\mathbf{a}},\tilde{\mathbf{b}})$ نیز برداری r بعدی از مشتقهای اول $g(.)$ با توجه به مولفه‌های $\mathbf{b}$ باشد که مقدار هر یک در $\tilde{\mathbf{a}}$ و $\tilde{\mathbf{b}}$ محاسبه شده‌اند. نیز فرض کنید $g_A''(\tilde{\mathbf{a}},\tilde{\mathbf{b}})$ یک ماتریس $q\times q$ بعدی از مشتقهای دوم $\partial^۲g/(\partial a_j\partial a_k)$ و $g_B''(\tilde{\mathbf{a}},\tilde{\mathbf{b}})$ یک ماتریس $r\times r$ بعدی از مشتقهای دوم $\partial^۲g/(\partial b_j\partial b_k)$ در $\tilde{a}$ و $\tilde{b}$ باشند. g_A'' و g_B'' در یک همسایگی از $(\bar{\mathbf{A}},\bar{\mathbf{B}})$ پیوسته و کراندار باشند.

ج) $\tilde{n} = \tilde{n}_A + \tilde{n}_B$ بطوریکه $\tilde{n}_A/\tilde{n} \to k \in (۰,۱)$

شروط فوق توسط رائو و وو[۲۲] (۱۹۸۵) برای تحقیق و بررسی خواص برآوردگرهای واریانس در طرح نمونه‌گیری طبقه‌بندی چندمرحله‌ای مورد استفاده قرار گرفته است. شرط (ج) نیز تضمین می‌کند که نمونه حاصل از یک چارچوب بطور مجانبی روی نمونهٔ دیگر هیچ تأثیری نمی‌گذارد.

قضیه ۱ فرض می‌کنیم شروط (الف) ــ (ج) برقرار باشند. آن گاه:

$$Var(\hat{\tau}) = g_A^T(\bar{\boldsymbol{A}},\bar{\boldsymbol{B}})\sum\nolimits_{AgA}(\bar{\boldsymbol{A}},\bar{\boldsymbol{B}}) + g_B^T(\bar{\boldsymbol{A}},\bar{\boldsymbol{B}})\sum\nolimits_{BgB}(\bar{\boldsymbol{A}},\bar{\boldsymbol{B}}) + o(\tilde{n}^{-۱}) \quad (۱۱)$$

و برآوردگر خطی واریانس فوق عبارتست از:

$$v_L(\hat{\tau}) = g_A^T(\hat{\bar{\boldsymbol{A}}},\hat{\bar{\boldsymbol{B}}})\boldsymbol{S}_A g_A(\hat{\bar{\boldsymbol{A}}},\hat{\bar{\boldsymbol{B}}}) + g_B^T(\hat{\bar{\boldsymbol{A}}},\hat{\bar{\boldsymbol{B}}})\boldsymbol{S}_B g_B(\hat{\bar{\boldsymbol{A}}},\hat{\bar{\boldsymbol{B}}}) \quad (۱۲)$$

فرض می‌کنیم $\hat{\tau}_{(hi)}^A$ برآوردگری به فرم $\hat{\tau}$ باشد که بعد از حذف مشاهدات iامین psu نمونهٔ طبقهٔ hام به دست آمده و $\hat{\tau}_{(hi)}^A = g(\hat{\bar{\mathbf{A}}}_{(hi)},\hat{\bar{\mathbf{B}}})$ که $\hat{\bar{\mathbf{A}}}_{(hi)}$ نیز برآوردگر $\bar{\mathbf{A}}$ بعد از حذف iامین psu نمونه طبقه hام در چارچوب A می‌باشد. بطور مشابه $\hat{\tau}_{(lj)}^B = g(\hat{\bar{\mathbf{A}}},\hat{\bar{\mathbf{B}}}_{(lj)})$ که

22) Wu

$\hat{\bar{\mathbf{B}}}_{(lj)}$ برآورد $\bar{\boldsymbol{B}}$ بعد از حذف jامین psu نمونهٔ طبقهٔ lام چارچوب B است. برآوردگر جک نایف واریانس $\hat{\tau}$ (با توجه به استقلال نمونه‌ها) عبارتست از:

$$v_J(\hat{\tau}) = \sum_{h=1}^{H} \frac{\tilde{n}_h^A - 1}{\tilde{n}_h^A} \sum_{i=1}^{\tilde{n}_h^A} (\hat{\tau}_{(hi)}^A - \hat{\tau})^2 + \sum_{l=1}^{L} \frac{\tilde{n}_l^B - 1}{\tilde{n}_l^B} \sum_{j=1}^{\tilde{n}_l^B} (\hat{\tau}_{(lj)}^B - \hat{\tau})^2 \quad (۱۳)$$

قضیه ۲ فرض می‌کنیم شروط (الف) ــ (ج) برقرار باشند. آن گاه:

$$v_J(\hat{\tau}) = v_L(\hat{\tau}) + o_p(\tilde{n}^{-1}) \quad (۱۴)$$

برآوردگر جک نایف واریانس فوق در حالتی که $\hat{\tau}$ تابع همواری از $\hat{\bar{\boldsymbol{A}}}$ و $\hat{\bar{\boldsymbol{B}}}$ باشد مورد بررسی قرار گرفته و طی قضایای فوق سازگاریش و نیز هم ارزیش با برآوردگر خطی واریانس ثابت می‌شود. در بین برآوردگرهای دو چارچوبی، برآوردگرهای SF و $SFrake$ بصورت توابع همواری از میانگین جوامع قابل بیان هستند. دیگر برآوردگرها نیز مادامی که پارامترهای θ_p، θ_H، β_{FB} و β_{S2} ثابت بوده و از روی داده‌ها برآورد نشوند، نیز توابع همواری از میانگین جوامع خواهند بود. بنابراین قضیه (۲) بیان می‌دارد که برآوردگر جک نایف واریانس برای شکل بهینهٔ هر برآوردگر سازگار است.

۱.۳ جک نایف کامل و اصلاح شده

برآوردگرهای $\hat{\theta}_p$، $\hat{\theta}_H$، $\hat{\beta}_{FB}$ و $\hat{\beta}_{S2}$ همگی توابعی از واریانس ـ کواریانس میانگین جوامع، S_A و S_B هستند و بطور کلی نمی‌توانند بصورت توابع قابل همواری از میانگین در نمونه‌گیری طبقه‌بندی بیان شوند (واریانسها معمولاً نامعلوم بوده و باید از روی نمونه برآورد شوند) بنابراین قضیه (۲) همیشه نمی‌تواند بطور مستقیم با برآوردگرهایی که توابعی از $\boldsymbol{S}_A$ و $\boldsymbol{S}_B$ هستند بکار رود. مشکل دیگر که در استفاده از جک نایف در خیلی از نمونه‌های طبقه بندی رخ می‌دهد، هنگام محاسبه $\boldsymbol{S}_{(hi)}^A$ و $\boldsymbol{S}_{(lj)}^B$ می‌باشد، چون

$$\boldsymbol{S}_{(hi)}^A = \boldsymbol{S}_A + \frac{(W_h^A)^2}{\tilde{n}_h^A - 2} \left[\frac{2\boldsymbol{S}_h^A}{\tilde{n}_h^A} - \frac{\tilde{n}_h^A}{(\tilde{n}_h^A - 1)^2} (\boldsymbol{a}_{hi} - \bar{\boldsymbol{a}_h})(\boldsymbol{a}_{hi} - \bar{\boldsymbol{a}_h})^T \right] \quad (۱۵)$$

در صورتی که $\tilde{n}_h^A = 2$ باشد قابل محاسبه نیست. بطور مشابه $\boldsymbol{S}_{(lj)}^B$ نیز وقتی که $\tilde{n}_l^B = 2$ قابل محاسبه نیست. برای رفع این مشکل رائو و لهر (۱۹۹۷) یک برآورد جک نایف اصلاح شده در حالتی که $\tilde{n}_h^A = 2$ و یا $\tilde{n}_l^B = 2$ ارائه دادند.

جک نایف (جک نایف اصلاح شده برای طرح نمونه‌گیری با ۲ واحد نمونه‌گیری اولیه در هر طبقه که بعداً به معرفی آن می‌پردازیم) حتی زمانی که برآوردگرها به $\boldsymbol{S}_A$ و $\boldsymbol{S}_B$ بستگی دارند، برآورد سازگاری از واریانس می‌دهد.

در $\hat{\tau}$ فرض بر این است که پارامترهای θ_P، θ_H، β_{FB} و $\beta_{S۲}$ ثابت بوده و از روی نمونه برآورد نمی‌شوند. اما اگر این پارامتر ثابت نبوده و به $\boldsymbol{S}_A$ و $\boldsymbol{S}_B$ بستگی داشته باشند باید اثر این برآوردگرها در $\hat{\tau}$ منظور شود. فرض می‌کنیم که

$$\begin{aligned}\hat{\tau} &= g(\hat{\bar{\boldsymbol{A}}}, \hat{\bar{\boldsymbol{B}}}) = f(\hat{\bar{\boldsymbol{A}}}, \hat{\bar{\boldsymbol{B}}}, \beta)\\ \beta &= [\boldsymbol{E}_A + \boldsymbol{E}_B]^{-۱}[\boldsymbol{e}_A + \boldsymbol{e}_B]\end{aligned}$$

که هر عنصر ماتریس $p \times p$ بعدی $\boldsymbol{E}_A$ و هر عنصر بردار p بعدی $\boldsymbol{e}_A$ ترکیب خطی از عناصر $\sum_A$ هستند و به همین ترتیب برای $\boldsymbol{E}_B$ و $\boldsymbol{e}_B$. آن گاه برآوردگرهای $\hat{Y}_H(\hat{\theta}_H)$ $\hat{Y}_{PML}(\hat{\theta}_p)$، $\hat{Y}_{FB}(\hat{\beta}_{FB})$ و $\hat{Y}_{SFreg}(\hat{\beta}_{S۲})$ همگی به فرم $N\hat{\zeta} = Nf(\hat{\bar{\boldsymbol{A}}}, \hat{\bar{\boldsymbol{B}}}, \hat{\beta})$ می‌باشند بطوریکه $\hat{\beta}$ برآورد کواریانسهای جامعه یعنی $\boldsymbol{S}_A$ و $\boldsymbol{S}_B$ را در β جایگزین می‌کند. برای مثال اگر بردارهای $\boldsymbol{A}$ و $\boldsymbol{B}$ را به صورت زیر در نظر بگیریم:

$$\boldsymbol{A} = (Y_a, Y_{ab}, N_{ab}, N_A)^T \qquad \boldsymbol{B} = (Y_b, Y_{ab}, N_{ab}, N_B)^T$$

آن گاه برآوردگر هارتلی $\hat{Y}_H(\hat{\theta}_H)$ را می‌توان به صورت زیر نوشت:

$$\hat{Y}_H(\hat{\theta}_H) = \hat{Y}_H(\theta_H) + (\hat{\beta}_H - \beta_H)[N_A(\hat{\bar{\boldsymbol{A}}}_{۲} - \bar{\boldsymbol{A}}_{۲}) - N_B(\hat{\bar{\boldsymbol{B}}}_{۲} - \bar{\boldsymbol{B}}_{۲})]$$

بطوریکه مقدار بهینه‌ی β_H برابر است با:

$$\beta_H = -\frac{k^{۲}\sum_A(۱,۲) - \sum_B(۱,۲) - \sum_B(۲,۲)}{k^{۲}\sum_A(۲,۲) + \sum_B(۲,۲)} \quad , \quad k = N_A/N_B$$

$\sum_A(۱,۲)$ عبارت از $(۱,۲)$ امین عنصر $\sum_A$ می‌باشد. برای برآوردگر شبه حداکثر درستنمایی، پارامتر مناسب عبارت از $\beta_p = k^{۳}\sum_A(۳,۳)/\sum_B(۳,۳)$ می‌باشد.

فرض کنید $\hat{\zeta}^A_{(hi)}$ برآوردگری به فرم $\hat{\zeta}$ باشد که بعد از حذف مشاهدات iامین نمونه psu طبقه hام حاصل شده است. آن گاه

$$\hat{\zeta}^A_{(hi)} = f(\hat{\bar{\boldsymbol{A}}}_{(hi)}, \hat{\bar{\boldsymbol{B}}}, \hat{\beta}^A_{(hi)})$$

که $\hat{\beta}^A_{(hi)}$ نیز برآوردگر β با استفاده از $\boldsymbol{S}^A_{(hi)}$ و $\boldsymbol{S}_B$ با فرض $(\tilde{n}^A_h > ۲)$ می‌باشد. به طور مشابه

$$\hat{\zeta}^B_{(lj)} = f(\hat{\bar{\boldsymbol{A}}}, \hat{\bar{\boldsymbol{B}}}_{(lj)}, \hat{\beta}^B_{(lj)})$$

به شرط اینکه $\tilde{n}^B_l > ۲$ باشد. آن گاه برآوردگر جک نایف واریانس $\hat{\zeta}$ عبارتست از:

$$v_J(\hat{\zeta}) = \sum_{h=۱}^{H} \frac{\tilde{n}^A_h - ۱}{\tilde{n}^A_h} \sum_{i=۱}^{\tilde{n}^A_h} (\hat{\zeta}_{(hi)} - \hat{\zeta})^{۲} + \sum_{l-۱}^{L} \frac{\tilde{n}^B_l - ۱}{\tilde{n}^B_l} \sum_{j=۱}^{\tilde{n}^B_l} (\hat{\zeta}_{(lj)} - \hat{\zeta})^{۲} \quad (۱۶)$$

برآوردگر جک نایف اصلاح شده واریانس، $\hat{V}_{MJ}(\hat{\zeta})$ نیز به همان فرم $\hat{V}_J(\hat{\zeta})$ بوده با این تفاوت که به جای $\hat{\beta}$ از $\hat{\beta}^B_{(lj)}$ یا $\hat{\beta}^A_{(hi)}$ استفاده کرده و قابل کاربرد با $\tilde{n}^A_h = ۲$ یا $\tilde{n}^B_l = ۲$ می‌باشد. برآوردگر خطی واریانس $\hat{\zeta}$ عبارتست از:

$$v_L(\hat{\zeta}) = \hat{g}^T_A(\hat{\tilde{\boldsymbol{A}}}, \hat{\tilde{\boldsymbol{B}}}) S_A \hat{g}_A(\hat{\tilde{\boldsymbol{A}}}, \hat{\tilde{\boldsymbol{B}}}) + \hat{g}^T_B(\hat{\tilde{\boldsymbol{A}}}, \hat{\tilde{\boldsymbol{B}}}) S_B \hat{g}_B(\hat{\tilde{\boldsymbol{A}}}, \hat{\tilde{\boldsymbol{B}}}) \qquad (۱۷)$$

که $\hat{g}_A$ و $\hat{g}_B$ از $\hat{\beta}$ بجای β استفاده می‌کنند. هر سه برآوردگر واریانس $v_J(\hat{\zeta})$، $v_{MJ}(\hat{\zeta})$ و $v_L(\hat{\zeta})$ بطور مجانبی در حالتی که اختلاف بین هر جفت از آنها از مرتبه $o_p(\tilde{n}^{-۱})$ باشد هم‌ارزند.[راثو و لهر(۱۹۹۷)]

برآورگرهای جک نایف واریانس مذکور برای حالت نمونه‌گیری طبقه‌بندی بدون جایگذاری از واحدهای نمونه‌گیری اولیه با فرض ناچیز بودن کسر نمونه‌گیری مطرح شده و قابل کاربرد با طرحهای نمونه‌گیری طبقه‌بندی چندمرحله‌ای بدون جایگذاری نیز می‌باشند. در حالتی که نمونه‌گیری با جایگذاری از واحدهای نمونه‌گیری اولیه به عمل آید نیز همین فرمولهای واریانس صادق هستند. یک حالت خاص ممکن است زمانی رخ دهد که در یکی از چارچوبها مثلاً چارچوب فهرستی B، نمونه‌گیری تصادفی ساده طبقه‌بندی انجام شده و کسرهای نمونه‌گیری ناچیز باشد. در این حالت $\boldsymbol{b}_{lj}$ برداری از مقادیر مرتبط با jامین واحد در طبقه lام چارچوب B و $\tilde{n}^B_l = n^B_l$ عبارت از تعداد واحدهای نمونه‌گیری شده از N^B_l واحد در طبقه lام چارچوب B می‌باشد. در صورتی که کسر نمونه‌گیری n^B_l/N^B_l ناچیز نباشد، بطور مشابه با روش جک نایف در حالت چارچوب منفرد(ولتر (۱۹۸۵))، $(n^B_l - ۱)/n^B_l$ با $((n^B_l - ۱)/n^B_l)(۱ - n^B_l/N^B_l)$ جایگزین می‌شود.

۲.۳ محاسبهٔ برآورد جک نایف واریانس برآوردگرهای دوچارچوبی

تمام برآوردگرهایی دو چارچوبی می‌توانند به فرم زیر بیان شوند:

$$\hat{Y} = \sum_{t \in s_A} \tilde{w}^A_t y_t + \sum_{t \in s_B} \tilde{w}^B_t y_t \qquad (۱۸)$$

که از وزنهای اصلاح شده $\tilde{w}^A_t$ و $\tilde{w}^B_t$ استفاده می‌کند. برای مثال $\hat{Y}_{PML}(\theta)$ از وزنهای زیر استفاده می‌کند:

$$\tilde{w}^A_t = \begin{cases} w^A_t [N_A - \hat{N}^{PML}_{ab}(\theta)]/\hat{N}^A_a & t \in a \\ w^A_t \theta \hat{N}^{PML}_{ab}(\theta)/\hat{N}_{ab}(\theta) & t \in ab \end{cases}$$

$$\tilde{w}^B_t = \begin{cases} w^B_t [N_B - \hat{N}^{PML}_{ab}(\theta)]/\hat{N}^B_b & t \in b \\ w^B_t (۱ - \theta) \hat{N}^{PML}_{ab}(\theta)/\hat{N}_{ab}(\theta) & t \in ab \end{cases}$$

اگر $\hat{\theta}_p$ را به عنوان برآوردگر پارامتر θ در نظر بگیریم، کافیست در وزنهای فوق بجای θ برآورد آن را جایگزین کنیم.

برای محاسبه‌ی برآورد جک نایف $\hat{Y}^A_{(hi)}$ بسادگی وزنهای w^A_t را با $w^A_{t(hi)}$ جایگزین می‌کنیم. اگر واحد t در خوشهٔ k ام طبقهٔ g در چارچوب A باشد آن گاه

$$w^A_{t(hi)} = \begin{cases} \circ & hi = gk \\ \dfrac{\tilde{n}^A_h}{(\tilde{n}^A_h - ۱)} w^A_t & h = g \quad i \neq k \\ w^A_t & h \neq g \end{cases}$$

بطور مشابه $\hat{Y}^B_{(lj)}$ را از وزنهای $w^B_{t(lj)}$ متناظراً محاسبه کرده و سپس برآورد جک نایف واریانس $\hat{Y}$ از رابطهٔ (۱۴) بدست می‌آید.

اگر وزنهای اصلاح شدهٔ $\tilde{w}^A_t$ و $\tilde{w}^B_t$ به عناصر $\boldsymbol{S}_A$ و $\boldsymbol{S}_B$ بستگی داشته باشند(همانطور که برای $\hat{Y}^A_{(hi)}$ اتفاق افتاد) برای محاسبهٔ برآورد جک نایف کامل $\hat{Y}_{PML}(\hat{\theta}_p)$ احتیاج به محاسبهٔ $\boldsymbol{S}^A_{(hi)}$ داریم. ماتریس $\boldsymbol{S}^A_{(hi)}$ به ازای $\tilde{n}^A_h \geq ۳$ می‌تواند با استفاده از یک جک نایف جداگانه درون هر تکرار جک نایف، این بار با مجموعه داده‌ای که مشاهدات i امین psu طبقه‌ی h ام چارچوب A حذف شده است.

اگر $\tilde{n}^A - H$ و $\tilde{n}^B - L$ هر دو بزرگ باشند، آنگاه یک فاصله اطمینان تقریبی ۹۵ درصد برای τ می‌تواند به صورت زیر بنا شود:

$$\hat{\zeta} \pm ۱/۹۶ \sqrt{\hat{V}(\hat{\zeta})} \qquad (۱۹)$$

در غیر اینصورت با استفاده از صدک توزیع t با درجهٔ آزادی برابر $min(\tilde{n}^A - H, \tilde{n}^B - L)$ بجای ۱/۹۶ به یک فاصله اطمینان محافظه کاری دست یابیم.

۴ نتایج شبیه سازی

در این بخش با استفاده از شبیه‌سازی به مطالعه خواص تجربی برآوردگرهای واریانس می‌پردازیم. فرض می‌شود که جامعه نامتناهی بوده و طرح نمونه‌گیری برای هر چارچوب یک طبقه دارد. یک نمونه خوشه‌ای دو مرحله‌ای با $\tilde{n}_A$ خوشه و m عنصر از هر خوشه به عنوان نمونه حاصل از چارچوب A $(n_A = \tilde{n}_A.m)$ و یک نمونه تصادفی ساده با n_B مشاهده به عنوان نمونه حاصل از چارچوب B تولید می‌شود. با فرض نامتناهی بودن جامعه، N_a/N و N_b/N را با γ_a و γ_b جایگزین می‌کنیم. نمونهٔ حاصل از چارچوب A شامل مقادیر $\{(y_{ij}, m_{ai})\ ,\ i = ۱, ۲, \cdots, \tilde{n}_A \quad j = ۱, \cdots, m\}$ می‌باشد. m_{ai} عبارت از تعداد عناصر نمونه‌گیری شده از iامین خوشهٔ نمونه‌ای، متعلق به حوزهٔ a و y_{ij} نیز مقدار مرتبط با

jامین عنصر نمونه‌ای در iامین خوشهٔ نمونه می‌باشد. نمونهٔ حاصل از چارچوب B شامل مقادیر نمونه‌ای $\{(y_j, n_b)\ ,\ j = ۱, ۲, \cdots, n_B\}$ است که در آن n_b، تعداد عناصر نمونه‌ای متعلق به حوزهٔ b و y_j، مقدار مرتبط با jامین عنصر نمونه‌ای است.

برای شبیه‌سازی، نمونه‌گیری را R=۱۰۰۰۰ بار تکرار کرده و از هرنمونه، برآورگرهای $\hat{Y}_H(\hat{\theta}_H)$، $\hat{Y}_{FB}(\hat{\beta}_{FB})$ و $\hat{Y}_{PML}(\hat{\theta}_P)$ و برآوردگرهای تک چارچوبی $\hat{Y}_{SF}$ و $\hat{Y}_{SFrake}$ و $\hat{Y}_{SFreg}(\hat{\beta}_{S۲})$ را با استفاده از مقادیر بهینهٔ $\hat{\theta}_H$ و $\hat{\beta}_{FB}$ و $\hat{\theta}_P$ و $\hat{\beta}_{S۲}$ و نیز سه برآوردگر واریانس خطی (L)، جک نایف کامل (J) و جک نایف اصلاح شده (MJ) را به ازای مقادیر مختلف پارامترهای طرح محاسبه می‌کنیم. در محاسبه جک نایف کامل به دلیل اینکه از مقادیر برآورد شدهٔ پارامترها استفاده می‌شود و این مقادیر، خود به ماتریس‌های S_A و S_B وابسته‌اند، $S^A_{(hi)}$ و $S^B_{(lj)}$ را با استفاده از یک جک نایف جداگانه درون هر تکرار جک نایف محاسبه می‌کنیم. در جدول (۱) نتایج میانگین توان دوم خطای تجربی EMSE (متوسط مربع انحراف برآوردگر از مقدار واقعی) برآوردگرهای دوچارچوبی و همچنین اریبی نسبی[۲۳] (RB) و خطای معیار نسبی[۲۴] (RSD) برآوردگرهای خطی، جک نایف و جک نایف اصلاح شده ارائه شده که بصورت زیر محاسبه می‌شوند:

$$EMSE = \frac{۱}{R}\sum_{r=۱}^{R}(\hat{Y}_r - Y)^۲$$

$$RB = ۱۰۰(EV - EMSE)/EMSE$$

$$RSD = (\text{انحراف معیار ۱۰۰۰۰ برآورد واریانس})/\sqrt{EMSE}$$

$\hat{Y}_r$ عبارت از مقدار $\hat{Y}$ برای r امین تکرار شبیه سازی و R تعداد تکرارهای شبیه سازی و نیز EV عبارت از میانگین ۱۰۰۰۰ برآورد واریانس در هر روش است. به دلیل مشابه بودن نتایج اریبی نسبی خطی و جک نایف اصلاح شده فقط نتایج حاصل از روش خطی ارائه شده است. همچنین با استفاده از فاصله اطمینان ۹۵ درصد نرمال

$$\text{برآورد} \pm ۱/۹۶SE(\text{برآورد})$$

و نیز صدک توزیع t با $(\tilde{n}_A - ۱)$ درجه آزادی، احتمال پوشایی تجربی برای روشهای خطی و جک نایف کامل داده شده است. در این شبیه‌سازی، تمام محاسبات با استفاده از نرم افزار S-PLUS 2000 صورت گرفته است.

نتایج شبیه‌سازی در جدول (۱) نشان می‌دهد که اریبی نسبی هر سه برآوردگر واریانس با افزایش حجم نمونه میل به کاهش دارند. همچنین روش جک نایف از پایداری کمتری نسبت به برآوردگر خطی واریانس برخوردار است. بعنوان مثال وقتی $\tilde{n}_A = ۱۰$ خطای معیار نسبی جک نایف تقریباً دو برابر خطای معیار نسبی خطی است با این حال با افزایش حجم نمونه پایداری برآوردگر جک نایف بهبود می‌یابد.

23) Relative Bias 24) Relative standard deviation

نتایج حاصل از فاصله اطمینان نشان می‌دهد که روش جک نایف کامل به وضوح احتمال پوشایی بالاتری از دو روش دیگر دارد. با استفاده از صدک توزیع t با $\tilde{n}_A - 1$ درجه آزادی بجای ۱٫۹۶، جک نایف کامل احتمال پوشایی نزدیک به ۰٫۹۵ دارد در حالیکه دو روش دیگر به وضوح احتمال پوشایی کمتری دارند. با افزایش حجم نمونه این احتمال پوشایی افزایش می‌یابد.

۵ نتیجه‌گیری

برآوردگر جک نایف واریانس را بطور نظری توجیه کرده با استفاده از شبیه‌سازی نشان دادیم که دارای اریبی کوچکتری از برآوردگر خطی واریانس است. همچنین برآوردگر جک نایف به وضوح دارای احتمال پوشایی بالاتری از برآوردگر خطی است خصوصاً وقتی از صدک t استفاده می‌شود احتمال پوشایی جک نایف خیلی نزدیک ۰٫۹۵ است.

جک نایف معمولاً با توابع غیر خطی نظیر نسبت مجموع دو جامعه بکار می‌رود. در این حالت مشتقات جزئی که در محاسبه‌ی برآورد خطی چنین مقادیر غیر خطی استفاده می‌شود، در بررسی‌های دوچارچوبی پیچیده‌تر از تک چارچوبی است. در صورتی که با استفاده از جک نایف می‌توان از این محاسبات پیچیده اجتناب کرد. برآوردگرهای جک نایف اخیراً در اداره آمار کانادا برای محاسبه برآورد واریانس در آمارگیری وسیع ملی کودکان و جوانان[۲۵] استفاده می‌شود. روشهای بازنمونه‌گیری دیگر برآورد واریانس نظیر تکرار متعادل پاسخ[۲۶] (BRR) و بوت استرپ[۲۷] نیز می‌توانند به موازات جک نایف در این گونه طرحها مورد استفاده قرار گیرند. حسن این روشها این است که بر خلاف جک نایف با توابع غیر هموار نظیر میانه نیز قابل کاربرد هستند.

25) National Longitudinal Survey of Children and Youth 26) Balanced Repeated Replication 27) Bootstrap

		اریبی نسبی		احتمال پوشایی			
				نرمال		t	
برآوردگر	$EMSE$	L	J	L	J	L	J
H	۷٫۱۶	-۱۹٫۹۴ (۰٫۷۸)	۱۵٫۴۵ (۱٫۹۰)	۰٫۸۸۷	۰٫۹۱۹	۰٫۹۲۳	۰٫۹۴۳
FB	۶٫۸۳	-۱۲٫۷۵ (۰٫۷۷)	۱۰٫۴۶ (۱٫۶۴)	۰٫۹۰۱	۰٫۹۲۱	۰٫۹۳۱	۰٫۹۴۶
PML	۶٫۵۸	-۹٫۹۱ (۰٫۸۲)	۱۱٫۴۶ (۱٫۸۷)	۰٫۹۱۲	۰٫۹۲۴	۰٫۹۳۹	۰٫۹۵۰
SF	۱۱٫۲۶	۴٫۱۷ (۰٫۴۷)	۴٫۱۷ (۰٫۴۷)	۰٫۹۴۸	۰٫۹۴۸	۰٫۹۷۲	۰٫۹۷۲
$SFreg$	۶٫۹۷	-۱۳٫۷۳ (۰٫۸۲)	۱۱٫۴۲ (۱٫۷۸)	۰٫۸۹۷	۰٫۹۱۸	۰٫۹۲۸	۰٫۹۴۳
$SFrake$	۶٫۵۵	-۰٫۱۱ (۱٫۰۸)	۴٫۸۵ (۱٫۱۹)	۰٫۹۱۸	۰٫۹۲۳	۰٫۹۴۴	۰٫۹۵۰
H	۳٫۵۸	-۱۵٫۹۷ (۰٫۴۱)	۱٫۸۰ (۰٫۷۴)	۰٫۹۰۶	۰٫۹۲۸	۰٫۹۲۳	۰٫۹۴۴
FB	۳٫۵۲	-۱۲٫۶۹ (۰٫۴۰)	-۰٫۶۳ (۰٫۶۷)	۰٫۹۱۳	۰٫۹۲۸	۰٫۹۳۰	۰٫۹۴۳
PML	۳٫۴۴	-۱۱٫۲۶ (۰٫۴۲)	-۱٫۱۰ (۰٫۷۴)	۰٫۹۱۹	۰٫۹۳۰	۰٫۹۳۵	۰٫۹۴۵
SF	۵٫۸۳	-۰٫۶۹ (۰٫۲۳)	-۰٫۶۹ (۰٫۲۳)	۰٫۹۴۴	۰٫۹۴۴	۰٫۹۵۸	۰٫۹۵۸
$SFreg$	۳٫۵۷	-۱۲٫۹۲ (۰٫۴۲)	۰٫۰۸ (۰٫۷۲)	۰٫۹۱۳	۰٫۹۲۸	۰٫۹۲۸	۰٫۹۴۲
$SFrake$	۳٫۴۸	-۵٫۳۳ (۰٫۵۳)	-۳٫۰۷ (۰٫۵۵)	۰٫۹۲۵	۰٫۹۲۹	۰٫۹۴۰	۰٫۹۴۴

جدول (۱). لازم به توجه است که در این جدول نتایج تنها به ازای $\gamma_a = ۰/۱$ و $\gamma_b = ۰/۲$ ارائه شده است. در بخش بالایی، $\tilde{n}_A = ۱۰$ و $n_B = ۱۰۰$ و در بخش پائینی $\tilde{n}_A = ۲۰$ و $n_B = ۲۰۰$ می‌باشد. $EMSE$ عبارت از میانگین مربعات خطای مونت کارلو برای ۱۰۰۰۰ بار تکرار شبیه‌سازی است. اریبی نسبی روش خطی و جک نایف و همچنین خطای معیار نسبی آنها در پرانتز زیر مقادیر اریبی نسبی داده شده است.

مراجع

[1] Hartley, H. O.(1962). Multiple Frame Surveys, in Proceedings of the Social Statistics Section,American Statistical Association, 203-206.

[2] Lund, R. E. (1968) . Estimators in Multiple Frame Surveys, in Proccedings of the Social Statistics Section, American Statistical Association, 282-288.

[3] Fuller, W. A. and Burmeister, L. F. (1972). Estimators for Samples Selected From two Overlapping Frames,in Proccedings of the Social Statistics Section, American Statistical Association, 245-249.

[4] Hartley, H. O.(1974). Multiple Frame Methodology and Selected Applications, Sankhya, Ser. C, 36, 203-206.

[5] Vogel, F. A. (1975). Survey With Overlapping Frame-Problems in Application, in Proccedings of the Social Statistics Section, American Statistical Association, 694-699.

[6] Ford, B. L. and Bosecker, R. R. (1979) Multiple Frame Estimation With Stratified Overlap Domain ,in Procceding of the Social Statistics Section, American Statistical Association, 219-224.

[7] Bankier, M. D. (1986). Estimators Based On Several Stratified Samples With Applications to Multiple Frame Surveys , Journal of the American Statistical Association, 81, 1074-1079.

[8] Skinner, C. J. (1991). On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys , Journal of the American Statistical Association, 86, 779-784.

[9] Skinner, C. J. and Rao, J. N. K. (1996). Estimation in Dual Frame Survey With Complex Designs, Journal of the American Statistical Association , 91, 349-356.

[10] Kalton, G. and Anderson, D. W. (1986). Sampling Rare Populations, Journal of the Royal Statistical Society ,Ser. A, 149, 65-82.

[11] Rao, J. N. K. and Lohr, S. L. (1997), Jackknife Variance Estimation in Dual Frame Surveys, Technical Report, Department of Mathematics and Statistics, Carleton University.

[12] Rao, J. N. K. and Wu, C. F. J. (1985). Inference from Stratified Samples: Second-Order Analysis of Three Methods for Nonlinear Statistics, Journal of the American Statistical Association , 80, 620-630.

[13] Wolter, K. M. (1985). Introduction to Variance Estimation, New York: Springer-Verlag.

[14] Rao, J. N. K. and Lohr, S. L. (2000). Inference From Dual Frame Survey Journal of The American Statistical Association, 95, 271-280.

آزمون برازش توزیع لجستیک در تعیین توانایی و رتبه‌بندی در امتحان‌ها

سید مقتدی هاشمی پرست[۱]، سید حسن مرتضوی نصیری[۲]، کمال عقیق[۳]

[۱] مرکز مطالعات، تحقیقات و ارزشیابی آموزشی
[۲] مرکز تحقیقات سیاست علمی کشور
[۳] دانشگاه صنعتی خواجه نصیرالدین طوسی

چکیده: در این مقاله ابتدا با استفاده از توزیع لجستیک سه پارامتری در پاسخگویی به سوالات چهار گزینه‌ای در امتحانات ورودی، میزان توانایی هر یک از داوطلبان برای هر سوال در نتیجه هر امتحان با روش نظریه سوال-پاسخ (IRT) مورد اندازه‌گیری و بر اساس توانایی نمره او تعیین شده و مبتنی بر این اطلاعات رتبه‌بندی داوطلبان انجام پذیرفته است. با استفاده از روش متداول سوالات چهارگزینه‌ای، توزیع‌های تجربی و نظری نیز محاسبه و رتبه‌بندی انجام پذیرفته است. دو نتیجه حاصل مورد قیاس قرارگرفته و با استفاده از آزمون نکویی برازش تناظر بین رتبه‌بندی‌ها مورد مطالعه و بحث قرار گرفته‌اند. در پایان نتایج حاصل در تعیین رتبه‌بندی دانشجویان در امتحان تولیمو ۱۲ به کارگرفته شده و با نتایج حاصل از رتبه‌بندی متداول مورد مقایسه قرارگرفته است.†

واژه‌های کلیدی: نظریه سوال-پاسخ، IRT، رتبه‌بندی، توزیع لجستیک، توانایی

۱ مقدمه

در این مقاله ابتدا به تحلیل و بررسی آزمون‌هایی که به صورت تستهای چهار جوابی است با استفاده از مدل نظریه سوال-پاسخ[۱] (IRT) مبتنی بر مدل لجستیک سه پارامتری می‌پردازیم. در این تحلیل

۱) برآورد پارامترهای هر یک از سؤالات (دشواری b، تشخیص a و حدس c)،
۲) برآورد احتمال ارایه پاسخ درست در هر یک از سطوح توانایی به آنها توسط نرم‌افزار ترسیم منحنی مشخصه‌های سوالات و نمودار تابع اطلاع آنها،
۳) طبقه‌بندی سوالات،
۴) براورد توانایی هر یک از امتحان دهندگان،
۵) نمره‌گذاری نتایج بر اساس برآورد توانایی هر امتحان دهنده (نمره‌گذاری مبتنی بر توانایی)،
۶) مقایسه نتایج بر اساس نمرات خام و نمرات مبتنی بر توانایی با استفاده از آزمون کای اسکور.

†) با تشکر از سازمان سنجش و آموزش کشورکه با پشتیبانی و حمایت آن سازمان این تحقیق انجام گرفته است.
1) Theory Item-Response

با وجود اینکه سالیان درازی است که آزمون‌گری کلاسیک از رواج و شهرت فراوانی برخوردار است اما همچنان دارای کاستی‌هایی است که سودمندی آن را به عنوان مبنا و اساسی برای آزمون‌گری مدرن محدود می‌سازد. شکوفایی و گسترش فن‌آوری کامپیوتر و توسعه روزافزون کاربرد آن در روانسنجی سبب شده است تا برخی از محدودیتهای نظریه کلاسیک آشکارتر و برجسته‌تر شود. کاربرد نظریه کلاسیک در موقعیتهای سنتی آزمون‌گری، اعم از گروهی یا فردی بسیار مناسب است. در این موقعیت‌ها که عملا نفس رتبه‌بندی اهمیت پیدا می‌کند و از تعیین توانایی افراد غفلت می‌شود، آزمون‌هایی یکسان یا دسته‌هایی از سوالات موازی به تمام اعضای جمعیت هدف (مثلا اشخاصی که درصدد ورود به دانشگاه هستند) داده می‌شود. این مجموعه سوالات را می‌توان به صورت چاپ شده یا از طریق کامپیوتر به امتحان دهندگان ارایه داد. وقتی از توانایی در زمینه‌های علمی صحبت می‌کنیم، می‌توانیم از اصطلاحات توصیف کننده‌ای همچون توانایی خواندن یا توانایی ریاضی استفاده کنیم. هر یک از این اصطلاحات درست به همان چیزی اشاره دارد که متخصصان روان‌سنجی از آن با عنوان صفت نامشهود[2] یا صفت مکنون[3] یاد می‌کنند. با وجود اینکه متخصصان روان‌سنجی چنین اصطلاحی را به راحتی توصیف می‌کنند و افراد مطلع می‌توانند خواص آن را برشمارند، اما اندازه‌گیری این مقولات مفهومی، به آسانی اندازه‌گیری خصوصیاتی فیزیکی و جسمانی نیست. یکی از هدف‌های اولیه اندازه‌گیری‌های روانی و تربیتی این است که تعیین کند افراد از چه مقداری از این صفت‌ها یا توانایی‌ها برخوردارند.

از اوایل دهه ۱۹۵۰ مدل‌هایی از اندازه‌گیری‌های روانی عرضه شدند که بر این تنگناهای نظریه کلاسیک انگشت گذاردند. امروزه رایج‌ترین و پخته‌ترین این مدلها، مجموعه‌ای را تشکیل می‌دهند که به شناخت و تعیین ویژگیها و مشخصه‌های ریاضی پاسخهای امتحان دهندگان به آزمون کمک می‌کند. این مدلها به عنوان مدل‌های نظریه سوال-پاسخ یا مدل‌های IRT شناخته شده‌اند. نظریه سوال-پاسخ (IRT) رابطه مشخصه‌ها یا پارامترهای هر سوال و ویژگیها یا توانایی افراد (صفت مکنون) را با احتمال ارائه پاسخ صحیح بررسی می‌کند. فرض بر آن است که مقدار این صفت (توانایی) همیشه در امتداد پیوستاری تک بعدی[4] تغییر می‌پذیرد که معمولاً آن را با حرف θ نشان می‌دهند. طبق نظریه سوال-پاسخ، هم سوالات آزمون و هم افراد پاسخ دهنده به آنها، روی مقیاس θ از کمترین تا بیشترین مقدار چیده می‌شوند. جای شخص iام که روی محور θ با θ_i نشان داده می‌شود معمولاً به توانایی یا شایستگی وی تعبیر می‌شود. جای سوال j روی محور θ که معمولاً با bj نشان داده می‌شود، به دشواری سوال تعبیر می‌شود. به طور شمی یا شهودی می‌توان انتظار داشت وقتی که $b_j - \theta_i$ افزایش پیدا می‌کند، احتمال وقوع پاسخ صحیح به سوال jام نیز به طور یکنواخت افزایش یابد. تمام مدلهای IRTاحتمال پاسخ صحیح به سوالی از آزمون‌های چند گزینه‌ای با یک پاسخ درست را به عنوان تابعی از θ مشروط به یک یا چند پارامتر سوال نشان می‌دهند. برای هر سوال می‌توان احتمال پاسخ صحیح دادن یا موافقت با طبقه خاصی از پاسخ را روی نمودار نشان داد. این توابع معرف رگرسیون غیرخطی احتمال وقوع پاسخ صحیح هستند. در این توابع احتمال وقوع پاسخ صحیح برحسب وجود مقدار معینی

2) unobservable trait 3) latent trait 4) unidimensional

از یک توانایی یا یک صفت مکنون نظیر هشیاری یا توانایی کلامی در نزد پاسخ دهنده تعریف می‌شود(هالین، دراسگو و پارسونز[5] [5].

احتمال وقوع هر پاسخ به عنوان تابعی از پارامترها و به طور جداگانه برای هر سوال و مشخصه‌های هر شخص، مدلسازی می‌شود. پارامترهای سوال معرف خواصی از سوال است که آن را به عنوان میزان دشواری و توانایی تشخیص سوال می‌شناسیم و مشخصه‌های شخص معرف میزان توانایی آزمون شونده است. اگر احتمال وقوع پاسخ به عنوان تابعی از مشخصه‌های (توانایی) فرد معرفی شود، تابع مورد نظر در IRT بعنوان تابع پاسخ شناخته می‌شود. در این مقاله گزارشی این از تحلیل و بررسی در آزمون تخصصی زبان انگلیسی سازمان سنجش آموزش کشور که با استفاده از مدل سه پارامتری انجام پذیرفته است عرضه می‌شود.

۲ تابع لجستیک به عنوان مبنای نظریه سوال-پاسخ

اگر بتوان با متغیر تصادفی X مشخصه‌های رفتار یک سیستم را در مقابل یک عمل یا آزمایش بیان کرد، قانون توزیع X، توزیعی است که به نام توزیع لجستیک خوانده می‌شود. مثلا میزان پاسخ‌های گروهی از امتحان دهندگان در مقابل یک پرسش، یا اندازه و میزان نحوه رفتار یک جامعه در برخورد با یک پدیده اجتماعی، دارای توزیع لجستیک است. توابع لجستیک اغلب برای توضیح گونه‌های خاصی از توزیع رشد بکار می‌رود. این مدلها، مدل‌های خوبی برای توضیح دادن توزیع و گسترش اطلاعات در جوامع و جمعیتها می‌باشند. همچنین مدل‌های خوبی برای تشریح و تبیین توزیع رشد بیولوژیک جمعیت درگونه‌هایی است که میزان رشد جمعیت‌شان آن قدر زیاد بوده که اکو سیستم آنها را به نقطه اشباع نزدیک کرده است. بنابراین می‌توان انتظار داشت که احتمال ارایه پاسخ صحیح به سوالاتی که یک توانایی معین را اندازه می‌گیرند به موازات افزایش توانایی در آزمودنی، افزایش یابد. منحنی تابع توزیع این توابع همانند تابع توزیع توابع نمایی، در ابتدا افزایش یا رشد سریع دارند اما به دلیل وجود بعضی محدودیت‌ها که حد و حدودی را بر جمعیت اعمال می‌کند، این افزایش به آرامی کاهش پیدا کرده و سپس قطع می‌شود. بنابراین مدل استاندارد ریاضی برای نظریه سوال-پاسخ و برای منحنی مشخصه‌های سوال، مدل تابع لجستیک[6] می‌باشد. نظریه سوال-پاسخ با فرض وجود رابطه‌ای ریاضی بین توانایی‌ها (یا دیگر صفت‌های مفروض) و احتمال جوابگویی به سوالها، به مطالعه و بررسی نمرات سوال و امتحان می‌پردازد.

5) Hulin, Drasgow & Parsons 6) logistic function

۳ مدل‌های یک، دو و سه پارامتری (بیکر [1])

وقتی در تحلیل هر سوال برآورد دو مشخصه آن یعنی میزان دشواری و تشخیص آن مورد نظر باشد مدل دو پارامتری با تابع توزیع احتمال زیر را انتخاب می‌کنیم.

$$p(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} \qquad (1)$$

در رابطه فوق θ نمایشگر متغیر توانایی و a نمایشگر پارامتر تشخیص و b نمایشگر پارامتر دشواری سوال است. واضح است که پارامترهای a و b برای هر سوال ثابت‌اند و دامنه تغییرات آنها از $-\infty$ تا $+\infty$ است ولی در عمل $-2/80 < a < 2/80$ و $-3 < b < 3$ در نظر گرفته می‌شود. b نقطه ای روی محور توانایی است که احتمال پاسخ گویی متناظر باآن ۰٫۵ باشد. باید توجه داشت که بر اساس مفروضات نظریه سوال-پاسخ، متغیر توانایی از پارامترهای تشخیص و دشواری سوال مستقل است.

اگر در رابطه فوق مقدار a را برابر ۱ قرار دهیم، مدل فوق به مدل لجستیک یک پارامتری تبدیل خواهد شد.

برین باوم[7] [3] با اندکی دستکاری در تابع لجستیک، مدل سه پارامتری را ارایه داد. اما با این وجود مدل وی همواره به عنوان تابع لجستیک سه پارامتری خوانده شده است. وقتی که عامل شانس در پاسخگویی دخیل باشد، مدلی سه پارامتری خواهیم داشت. برای مدل سه پارامتری می‌توان از فرمول زیر استفاده کرد:

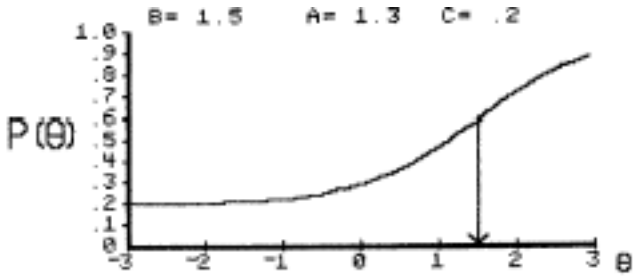
$$p(\theta) = c + (1 - c)\frac{1}{1 + e^{-a(\theta - b)}}$$

یعنی در این مورد تنها ضریب حدس ($0 < c < 1$) وارد می‌شود که معمولا حدس بالای ۰٫۳۵ قابل قبول نیست. با توجه به آنچه گفته شد در واقع می‌توان برای توانایی هر امتحان دهنده یک مقدار عددی یا یک نمره منظور کرد که جای او را روی محور توانایی مشخص می‌کند. این مقدار توانایی را با حرف θ و احتمال اینکه امتحان دهنده با این سطح توانایی به این سوال پاسخ درست بدهد را با $p(\theta)$ نشان می‌دهیم. در شکل ۱ محور x نشانگر سطح توانایی و محور y نشان دهنده احتمال وقوع پاسخ درست به یک سوال از امتحان می‌باشد. منحنی S گونه نشان دهنده احتمال وقوع پاسخ درست از سوی افرادی است که مقدار توانایی آنها متفاوت است.

۴ برآورد پارامترهای یک سوال: بیکر [2]

فرض کنیم در آزمونی، یک نمونه مرکب از n تا امتحان شونده به یک سوال موجود در امتحان پاسخ داده‌اند. سطح توانایی این امتحان دهندگان روی محور توانایی توزیع شده است. این امتحان

7) A. Birnhaum



شکل ۱: منحنی مشخصه‌های سوال برای مدل سه پارامتری

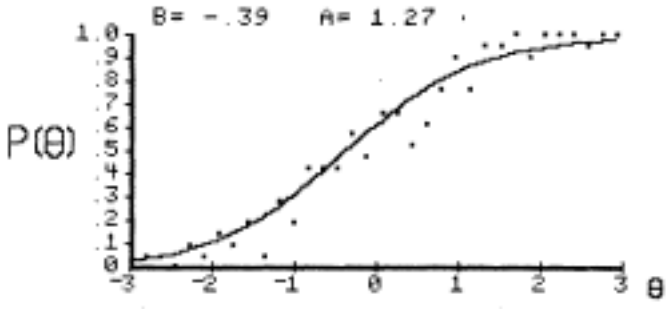
دهندگان را به t گروه توانایی در طول محور توانایی تقسیم می‌کنیم. فرض کنیم n_i امتحان دهنده در گروهi، $i = ۱, ۲, ۳, \ldots, t$، با سطح توانایی یکسان θ_i قرار دارند و فرض کنیم r_i امتحان دهنده از این گروه به سوال، پاسخ صحیح داده‌اند. بنابراین برآورد احتمال پاسخ‌های صحیح در سطح توانایی θ_i، با محاسبه نسبت زیر (یعنی نسبت پاسخ‌های صحیح مشاهده شده به کل پاسخها در همین زیرگروه) امکان پذیر است:

$$p(\theta_i) = \frac{r_i}{n_i}$$

کار اساسی ما این است که منحنی مشخصه‌های سوال را که بهترین برازش[8] برای نسبت پاسخ‌های مشاهده شده است، به دست آوریم. فرآیند برازش منحنی، با استفاده از روش برآورد حداکثر درست‌نمایی[9] انجام می‌گیرد.

در این روش، برآوردهای اولیه برای مقادیر a و b بدین قرار فرض می‌شود $a = ۱$ و $b = ۰$. آنگاه با تکیه بر این برآوردها، مقدار $P(\theta_j)$ با استفاده از معادله تابع لجستیک مدل دو پارامتری در سطح توانایی θ_j محاسبه می‌شود. سپس این مقدار، با مقدار مشاهده شده $p(\theta_i)$ در هر زیرگروه توانایی مقایسه می‌شود. بعد برای پارمترهای برآوردشده سوال سرشکنی تعیین می‌شود. نتیجه این کار آن است که توافق بهتری بین منحنی مشخصه‌های سوال (که با استفاده از مقادیر برآوردشده برای پارامترها رسم می‌شود) و نسبت مشاهده شده پاسخهای درست، حاصل می‌شود. فرآیند سرشکنی برآوردها تا جایی ادامه می‌یابد که مقدار آن به اندازه‌ای کوچک شود که دیگر امکان بهبود این توافق بسیار ناچیز باشد. در این نقطه، تکرار عملیات مراحل برآورد متوقف می‌شود و آخرین مقادیر بدست آمده a و b برآورد پارامترهای سوال می‌باشند. با قرار دادن این برآوردها در معادله منحنی مشخصه‌های سوال می‌توان احتمال پاسخ درست $P(\theta_i)$ را برای هر سطح توانایی محاسبه کرد. منحنی حاصل، منحنی مشخصه‌های سوال است که بهترین برازش برای

8) fitting 9) maximum likelihood estimation



شکل ۲: منحنی مشخصه‌های سوال برازش شده با نسبت مشاهده شده پاسخ‌های درست

پاسخ‌های داده شده به آن سوال است. شکل ۲ یک منحنی مشخصه سوال را نشان می‌دهد که با نسبت پاسخ‌های صحیح مشاهده شده برازش شده است. نتیجه حاصل آن است که در مدل IRT پارامترهای a و b محاسبه شده در هر زیرگروه از توانایی همیشه یکسان خواهد بود. این در حالی است که اندیس دشواری سوال در مدل کلاسیک از یک زیرگروه به زیرگروه دیگر متغیر می‌باشد. به همین دلیل، تعبیر و تفسیر دشواری سوال آن طورکه در نظریه IRT تعریف شده، ساده‌تر می‌باشد. در اینجا برآورد پارامترهای سوال را برای مدل لجستیک سه پارامتری توضیح می‌دهیم. فرض کنیم x_j، تابعی دو مقداری باشد که مقدار ۱ در آن برای پاسخ درست و صفر برای پاسخ نادرست به سوال jام در نظرگرفته شده است. فرض کنیم X^m ماتریس $k \times ۱$ تایی است که مؤلفه‌های آن صفر یا یک می‌باشند یعنی پاسخهای دو حالتی شخص mام به مجموعه‌ای از k سوال باشد. پس فرض کنیم که ماتریس X ماتریس پاسخ گویی n تا امتحان دهنده به k تا سوال باشد. احتمال مشاهده X پیش از مشاهده واقعی پاسخهای نمونه به قرار زیر است:

$$p(X|\theta,\beta) = \prod_i p(x_i|\theta_i,\beta) = \prod_i \prod_j p_j(\theta_i)^{x_i^j} q_j(\theta_i)^{۱-x_i^j} \qquad (۲)$$

که $q_j(\theta) = ۱ - p_j(\theta)$ و θ و بردارهای پارامتری ثابت نامعلوم در مدل هستند. برای مدل لجستیک سه پارامتری، مؤلفه‌های عبارتند از پارامترهای دشواری، تشخیص و حدس برای k تا سوال. معادله (۲) بر این نکته متکی است که احتمال پاسخگویی صحیح از سوالی به سوالی (هر جفت سوال) در سطح توانایی خاص θ، مستقل از یکدیگر هستند. این قضیه فرض بنیادی تمام مدلهای IRT است. بعد از اینکه مشاهده عملی نمونه‌ها انجام شد، دیگر معادله (۲)، یک احتمال نخواهد بود و به جای آن به ازای مقدار x_i معین، معادله را می‌توان به عنوان تابع درست‌نمایی x_i از θ و به شرط x_i تعریف کرد. مقدار تابع درست‌نمایی، به دسته‌ای خاص از مقادیر و پارامترهای واقعی درست‌نمایی بستگی دارد که در نقطه x_i مشاهده بشود. در

مدل لجستیک سه پارامتری، رایج‌ترین روش برآورد پارامترهای سوال، برآورد حداکثر درست‌نمایی کناری[10] است. در این روش برآورد پارامترها از طریق انتگرال‌گیری از تابع درست‌نمایی روی توزیع توانایی به دست می‌آید. یعنی

$$L(\beta|X) = \prod_i \int p(x_i|\theta,\beta) f(\theta) d\theta \qquad (۳)$$

که $f(\theta)$ تابع چگالی توانایی می‌باشد. برآوردهای حداکثر درست‌نمایی کناری، مقادیری از هستند که معادله (۳) را ماکزیمم می‌کند.

اگر پارامترهای سوال در مدل IRT معلوم باشند آنگاه برآورد توانایی برای نمونه‌ای از امتحان دهندگان با استفاده از روش برآورد حداکثر درست‌نمایی[11] آسانتر و صریحتر خواهد بود. در همین نوشته خواهیم دید که چگونه می‌توان میزان توانایی یا θ را با استفاده از برآورد پارامترهای سوال که در جریان کالیبره کردن سوالات حاصل می‌شود، برآورد کرد. برآورد همزمان توانایی و پارامترهای سوال در مدل‌های مختلف (اعم از یک، دو یاسه پارامتری) کاری دشوار و پر دردسر است. البته تهیه نرم‌افزارهای کامپیوتری یا نرم‌افزارهای موجود این کار را نسبتا آسان ساخته است.

۵ برآورد توانایی امتحان دهنده

در نظریه سوال-پاسخ، هدف اولیه ارایه یک آزمون به امتحان دهنده این است که جای او روی مقیاس توانایی مشخص شود. برای اندازه‌گیری توانایی مبتنی بر تعدادی سوال (N) خواهد بود که هرکدام جنبه‌ای از آن توانایی را اندازه می‌گیرد. در بحث پارامترهای سوال و برآورد آنها فرض می‌شود که پارامتر توانایی امتحان دهندگان معلوم است. بر عکس برای برآورد (پارامتر) توانایی نامعلوم یک امتحان دهنده، فرض بر این خواهد بود که مقادیر عددی پارامترهای سوالات آزمون معلوم‌اند. وقتی امتحانی اجرا می‌شود، هر امتحان دهنده به هر یک از سوالات آن پاسخی می‌دهد و به پاسخ‌ها به صورت دو وجهی[12] نمره داده می‌شود. پس برای هر یک از سوالات آزمون، نتیجه نمره‌ای برابر با یک یا صفر خواهد بود. در نظریه سوال-پاسخ برای برآورد توانایی امتحان دهنده، از روش حداکثر درست‌نمایی استفاده می‌شود. این روش، همان طور که برای برآورد پارامترهای سوال نیز در این حالت وجود داشت، فرآیندی است که به دفعات تکرار می‌شود. انجام آن به این شکل است که با ملحوظ داشتن مقادیر معلوم پارامترهای سوالات، ابتدا عددی به عنوان مقدار پیشین[13] برای توانایی امتحان دهنده تعیین می‌شود. از این مقادیر برای محاسبه میزان احتمال وقوع پاسخ صحیح به سوال از سوی امتحان دهنده استفاده می‌شود. آنگاه سرشکنی[14] برای

10) Marginal Maximum Likelihood estimation 11) Maximum Likelihood estimation
12) dichotomously 13) priori value 14) adjustment

برآورد توانایی به دست می‌آید که توافق احتمالات محاسبه شده را با بردار پاسخ‌های امتحان دهنده به سوالات اصلاح کرده و بهبود می‌بخشد. این فرآیند تکرار و تکرار می‌شود تا اندازه میزان حاصل به به قدری کوچک شود که تغییرات در توانایی برآوردشده قابل اغماض باشد. نتیجه عبارت است از برآورد پارامتر توانایی امتحان دهنده. بعد این فرآیند برای هر یک از امتحان دهندگانی که در آزمون شرکت داشته‌اند به طور جداگانه تکرار می‌شود. به هر حال این روش مبتنی بر رویکردی است که با هر امتحان دهنده، جداگانه برخورد می‌کند. از این رو موضوع اصلی آن این است که چطور می‌شود توانایی یک فرد واحد را برآورد کرد. معادله برآورد در فرآیند بر آورد حداکثر درست‌نمایی به صورت زیر است:

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \frac{\sum_{i=1}^{N} a_i[u_i - P_i(\hat{\theta}_s)]}{\sum_{i=1}^{N} a_i^2 - P_i(\hat{\theta}_s)Q_i(\hat{\theta}_s)}$$

که در آن $\hat{\theta}_s$ توانایی برآورد شده برای امتحان دهنده با تکرار s است، a_i پارامتر تشخیص سوال i برای $i = 1, 2, 3, \ldots, k$ است، $u_i = 1$ برای پاسخ درست به سوال i و $u_i = 0$ برای پاسخ ناردست به سوال i است، $p_i(\hat{\theta}_s)$ احتمال پاسخ صحیح به سوال i در مدل منحنی مشخصه‌های سوال داده شده در سطح توانایی $\vec{\theta}$ با تکرار s است، $Q_i(\hat{\theta}_s) = 1 - p_i(\hat{\theta}_s)$ احتمال پاسخ نادرست به سوال i در مدل منحنی مشخصه‌های سوال داده شده در سطح توانایی $\hat{\theta}$ با تکرار s است. اینک می توانیم خطای استاندارد برآورد را طبق فرمول زیر محاسبه کنیم:

$$SE(\hat{\theta}) = \frac{1}{\sqrt{\sum_{i=1}^{N} a_i^2 P(\hat{\theta})Q(\hat{\theta})}}$$

قابل توجه است که زیر رادیکال این معادله، همان مخرج معادله برآورد توانایی است. مراحل برآورد حداکثر درست‌نمایی در حالتی که امتحان دهنده به همه سوالات پاسخ غلط و یا در حالتی که امتحان دهنده به همه سوالات پاسخ درست داده باشد قابل محاسبه نیست. یعنی برآورد توانایی در حالت اول $-\infty$ و در حالت دوم $+\infty$ می‌شود.

۶ مربع خی

در تحلیل‌های IRT مسئله مهم این است که آیا یک منحنی مشخصه سوال معین با داده‌های مشاهده شده، خوب برازش[15] شده است یا نه؟ روش‌های مختلفی برای آزمون نکویی برازش پیشنهاد شده است. برای بررسی نکویی برازش از آزمون آماری مجذور خی[16] استفاده شد.

$$\chi^2 = \sum_{i=1}^{t} n_i \frac{[p(\theta_i) - P(\theta_i)]^2}{P(\theta_i)Q(\theta_i)}$$

15) goodness-of-fit 16) chi-square (chi-square goodness-of-fit)

که t عبارت است از تعداد زیر گروههای توانایی، θ_i سطح توانایی گروه iام، n_i تعداد امتحان دهندگانی که سطح توانایی θ_i دارند، $p(\theta_i)$ احتمال مشاهده پاسخ صحیح گروه i، و $P(\theta_i)$ احتمال پاسخ صحیح برای گروه i است که با استفاده از برآورد پارامترها در مدل منحنی مشخصه سوال محاسبه شده است و $Q(\theta_i) = ۱ - P(\theta_i)$.

اگر مقدار بدست آمده از یک مقدار معین بزرگتر باشد، منحنی مشخصه سوال که با برآورد پارامترهای سوال به دست آمده، برازنده داده‌ها نیست. این امر ممکن است به دو دلیل اتفاق بیافتد. اول اینکه مدل نادرستی برای ترسیم و تحلیل منحنی مشخصه‌های سوال انتخاب شده باشد. دوم اینکه مقدار احتمالات مشاهده شده پاسخ های صحیح آن قدر متفرق باشند که نتوان منحنی برازش شده ای را برای مدل مورد انتظار به دست آورد.

تحت مدل مشخص IRT. فراوانی مورد انتظار انتخاب k توسط پاسخ دهندگان با استفاده از فرمول زیر محاسبه می‌شود.

$$E_i(k) = N \int P(v_i = k|\theta = t) f(t) dt$$

$f(t)$ در این فرمول عبارت است از چگالی توانایی که معمولا استاندارد نرمال فرض می‌شود زیرا توابع سوال-پاسخ بر مبنای این توزیع مقیاس سازی می‌شود. فون در والنبرگ[۱۷] نشان داده است که آماره مربع خی برای هر یک از تک سوال‌ها در بسیاری از موارد به فرض تک بعدی بودن آزمون، حساس نیست. برای اجتناب از این مسئله پیشنهاد شده است که مجذور خی برای سوالات در دسته های دو تایی و سه تایی محاسبه شود. در صورتی که این دسته سوالات از نابرازندگی مشابهی[۱۸] برخوردار باشند، مقادیر بزرگی از مربع خی را نشان خواهند داد. فراوانی مورد انتظار برای یک جفت سوال برای خانه‌های (k, k') در جدول دو طرفه برای سوال‌های $(i$ و$i')$ به صورت زیر محاسبه می‌شود:

$$E_{i,i'}(k, k') = N \int P(v_i = k|\theta = t) P(v_i = k'|\theta = t) f(t) dt$$

اگر عدم برازشی بین مدل و داده‌ها رخ دهد مربع خی برای دسته سوالات دوتایی و سه‌تایی در سطح توانیی یکسان افزایش خواهد یافت. دراسگو و همکارانش ([4]) برای آسان شدن مقایسه مربع خی بر حسب حجم نمونه پیشنهاد کرده‌اند که مقادیر مربع خی برای نمونه‌ای سه هزار نفره سرشکن شده و به درجه آزادی تقسیم شود. آنها در پی مطالعات مختلفی در یافته‌اند که نسبت‌های سرشکن شده χ^2 بر درجه آزادی که کمتر از ۳ باشد نشان دهنده نکویی برازش مدل و داده‌هاست. به دست آمدن مقادیر بزرگ برای این نسبت‌ها ممکن است غلط بودن فرض استقلال موضعی یا تک بعدی بودن آزمون را نشان دهد.

17) Van der Wollenberg 18) similar misfits

۷ تحلیل نتایج آزمون تولیمو ــ الف ــ تحلیل سؤالات آزمون تولیمو

بر این اساس آنچه تا اینجا گفته شد، برای تمام سؤالات آزمون تولیمو ۱۲، پارامترهای سؤالات (یعنی دشواری b، تشخیص a و حدس c) برآورد شد. چون سؤالات از نوع چند گزینه‌ای بودند، امکان پاسخگویی به سؤالات از طریق حدس وجود داشت، بنابراین پارامتر حدس نیز برای تعیین احتمال پاسخگویی صحیح به سؤال وقتی که دانش یا توانایی درکار نیست، برآورد شد. منحنی مشخصه‌های سؤالات نیز بر اساس مقادیر برآورده شده‌ی پارامترها رسم شد و سرانجام سؤالات بر اساس مشابهت مقدار پارامترهای آنها و نیز بر اساس مقایسه منحنی‌های مشخصه سؤالات، به ۸ گروه زیر تقسیم شدند(جدول ۱).

حال که مشخصات هر یک از گروه‌های سوالات را بیان کردیم، این مسئله مطرح می‌شود که کدامیک از انواع سؤالات برای سنجش توانایی یا دانش زبان آزمودنی‌ها مناسبتر است. اگر هدف از اجرای آزمون تولیمو سنجش توانایی دانش زبان و تعیین سطوح مختلف توانایی افراد باشد. اولاً) دشواری سؤالات آزمون باید به گونه‌ای باشند که احتمال اندکی از پاسخگویی درست به سوال برای افراد برخوردار از سطح توانایی پایین نیز وجود داشته باشد و در تمام موارد با بالا رفتن سطح توانایی، احتمال پاسخگویی درست نیز افزایش یابد. در سوالات چهارگزینه‌ای مقدار این احتمال برای سطوح پایین توانایی چیزی در حد و حدود احتمال پاسخگویی تصادفی یا پاسخگویی درست از روی شانس است. در واقع سؤالات باید از دشواری متوسط برخوردار باشند. ثانیاً) سؤالات باید از پارامتر تشخیص مناسبی نیز برخوردار باشد تا هم بتواند افراد دارای سطح بالای توانایی را از افراد با سطح پایین توانایی صراحتا متمایزکند و هم سطوح مختلف بین این دو را شناسایی کند.

با مفروض دانستن هدف فوق، سؤالات گروه اول و دوم، سوالات مناسبی نیست، چون مقدار پارامتر دشواری سؤالات این دوگروه خیلی بالا است و احتمال پاسخگویی درست از سوی تواناترین افراد تنها ۰/۵ بوده است و احتمال پاسخگویی درست از سوی افراد با توانایی متوسط و پایین نیز بسیار ناچیز بوده است. در نتیجه این دوگروه برای تعیین سطوح مختلف توانایی دانش زبان مناسب نیستند و با این دوگروه از سوالات می‌توان تا حدی فقط تواناترین افراد را در زمینه دانش زبان مشخص نمود.

با فرض فوق، سؤالات گروه سوم و چهارم نیز مناسب نیست، چون پارامتر دشواری بالایی دارند و احتمال پاسخگویی درست تنها برای افراد برخوردار از سطح توانایی بالاتر از یک ($\theta > ۱$)، به مقدار قابل توجهی بوده و از این سطح به بعد رو به افزایش بوده است. افراد با توانایی متوسط و پایین هم به سختی و با احتمال بسیار اندکی توانسته‌اند به این گروه از سؤالات پاسخ درست بدهند. پس با این گروه سوالات نیز فقط می‌توان تواناترین افراد را مشخص کرد.

سؤالات گروه پنجم نیز برای تعیین سطوح مختلف توانایی دانش زبان مناسب نیستند چون افرادی که در سطوح متوسط و پایین توانایی قرار گرفته‌اند ($\theta < ۱$) با احتمال بسیار ناچیزی توانسته‌اند به این گروه از سؤالات پاسخ درست بدهند. با این گروه از سوالات تنها می‌توان

جدول ۱: طبقه‌بندی سوالات آزمون تولیمو

گروه	
اول	این گروه که مشتمل بر ۵۰ سوال می‌باشد دارای خصوصیات زیر است: پارامتر دشواری ($b = ۳$)، پارامتر تشخیص بالاتر از یک ($a > ۱$)، و پارامتر حدس کمتر از ۰٫۲ ($c < ۰٫۲$). همان طور که پارامتر دشواری و منحنی مشخصه‌های سؤالات نشان می‌دهد، سؤالات موجود در این گروه، سوالاتی بسیار دشوار هستند به گونه‌ای که احتمال ارایه پاسخ درست در بالاترین سطح توانایی یعنی $\theta = ۳$ فقط ۰٫۵ بوده است و این احتمال برای افراد با توانایی متوسط و پایین هم بسیار ناچیز بوده است. با این سؤالات تنها تا حدودی می‌توان تواناترین افراد را از سایرین تشخیص داد و می‌توان گفت که سؤالات اصلا قادر به شناسایی افراد با توانایی متوسط و پایین و تمیز آنه از یکدیگر نیست.
دوم	این گروه که مشتمل بر ۱۳ سوال می‌باشد دارای خصوصیات زیر است: پارامتر دشواری ($b = ۳$)، پارامتر تشخیص بالاتر از یک ($a > ۱$) و پارامتر حدس بالاتر از ۰٫۲ ($c > ۰٫۲$). در این گروه ۱۳ سؤال جای گرفته است که همانند سؤالات گروه قبلی بسیار دشوارند، طوری که احتمال ارایه پاسخ درست به این سوالات از سوی افراد با سطح توانایی بالا $\theta = ۳$ تنها ۰٫۵ بوده است و افراد با توانایی متوسط و پایین چندان قادر به پاسخگویی درست به سؤالات نبودند. البته با این تفاوت که در این دسته از سوالات احتمال پاسخگویی درست با تکیه بر شانس برای آنها بیشتر بوده یا می‌توان گفت در حد معقولی بوده است. با این گروه از سوالات نیز تنها تا حدودی می‌توان تواناترین افراد را مشخص کرد.
سوم	این گروه که مشتمل بر ۱۵ سوال می‌باشد دارای خصوصیات زیر است: پارامتر دشواری ($b = ۳$)، پارامتر تشخیص بین ۰٫۵ تا یک ($۰٫۵ < a < ۱$)، پارامتر حدس کمتر از ۰٫۱۵ ($c < ۰٫۱۵$).). این گروه از سوالات شامل ۱۵ سؤال می‌شود. با توجه به منحنی مشخصه‌های سوالات می‌توان گفت از سطح توانایی یک ($\theta > ۱$) به بالا، به موازات افزایش توانایی احتمال دادن پاسخ درست به این سؤالات هم افزایش می‌یابد. اما برای افرادی که روی مقیاس توانایی در سطح متوسط و پایین قرارگرفته‌اند احتمال ارایه پاسخ درست به سوالات بسیار پایین بوده است و با توجه به دشواری سؤالات می‌توان گفت که این احتمال حتی برای تواناترین افراد ($\theta = ۳$) هم فقط ۰٫۵ بوده است. در نتیجه با این گروه از سوالات هم تنها تا حدودی می‌توان توانایی‌های بالاتر از متوسط و نیز تواناترین افراد را در زمینه دانش زبان شناسایی کرد.
چهارم	۳۴ سؤال در این گروه قرارگرفته که ۱۵ سؤال دارای پارامتر دشواری $b = ۳$، ۱۴ سؤال پارامتر دشواری بین ۲ تا ۳ ($۳ > b > ۲$)، و ۵ سؤال هم پارامتر دشواری کمتر از ۲ ($۲ > b$) دارند. منحنی مشخص‌های سؤالات این گروه شیب بسیار ملایمی دارد زیرا تمام سؤالات پارامتر تشخیص پایینی دارند و در نتیجه سؤالات از قدرت تمیزکافی برخوردار نیستند و نمی‌تواند افرادی را که در سطح توانایی بالا قرارگرفته‌اند از افرادی که در سطح توانایی پایین قرار دارند به صراحت تفکیک کند.

ادامه جدول ۱

گروه	
پنجم	این گروه که مشتمل بر ۱۹ سوال می‌باشد دارای خصوصیات زیر است: دشواری ($۱ < b < ۲/۵$) و تشخیص بالاتر از ۰/۵ ($۰/۵ < a$) همان گونه که منحنی مشخصه‌های سؤالات نشان می‌دهد از سطح توانایی یک به بالا ($۱ < \theta$)، احتمال پاسخگویی درست به سؤالات، روند افزایشی داشته است، یعنی با بالا رفتن توانایی، احتمال پاسخگویی درست نیز افزایش یافته است. بنابراین این گروه از سوالات به خوبی توانسته است افرادی را که در زمینه دانش زبان دارای توانایی بالاتر از متوسط بوده‌اند، شناسایی کند. امادر سطوح توانایی متوسط و پایین تر از آن قادر به تفکیک و تشخیص تفاوتها نبوده است.
ششم	سه سؤال در این گروه جای دارد با پارامتر دشواری بین ۰/۵ تا ۱ ($۰/۵ < b < ۱$)، پارامتر تشخیص بزرگتر از ۰/۵ ($۰/۵ < a$) و پارامتر حدس بزرگتر از ۰/۲. با توجه به منحنی مشخصه‌های سؤالات از سطح توانایی متوسط به بالا، احتمال پاسخگویی درست به سؤالات رو به افزایش بوده است و با افزایش توانایی، احتمال پاسخگویی درست نیز افزایش یافته است. پارامتر تشخیص سؤالات نیز مطلوب است. با این گروه می‌توان افراد با توانایی متوسط و بالا را در زمینه دانش زبان به خوبی مشخص نمود. اما این سوالات قادر به تفکیک افراد در سطوح پایینتر از متوسط نمی‌باشد.
هفتم	پنج سؤال در این گروه جای دارد با پارامتر دشواری کمتر از ۰/۵ ($b < ۰/۵$)، پارامتر تشخیص بین ۰/۵ تا ۱ ($۰/۵ < a < ۱$) است. با توجه به پارامترهای دشواری این سؤالات می‌توان گفت که احتمال پاسخگویی درست به این سوالات برای کسانی که در سطوح پایین توانایی قرار داشته‌اند کمتر بوده و هر چقدر توانایی افزایش یافته این احتمال نیز افزایش می‌یافته است. البته این موضوع تا سطح توانایی ۲ ($\theta = ۲$) صادق بوده و احتمال پاسخگویی درست برای سطوح مختلف توانایی بالاتر از آن یکسان بوده است. به عبارت دیگر پارامتر تشخیص سؤالات نیز تا حد زیادی مناسب است یعنی این گروه از سوالات تا سطح توانایی $\theta = ۲$ به خوبی افراد را از یکدیگر متمایز می‌کند، اما از این سطح توانایی به بالا دیگر خوب عمل نمی‌کند. در واقع این گروه از سؤالات نمی‌تواند تواناترین افراد را از افراد توانا تفکیک کند. اما برای سنجش و تمیز سطوح توانایی ضعیف تا بالاتر از متوسط کارایی دارد.
هشتم	این گروه فقط شامل یک سؤال می‌شود که دارای پارامتر دشواری $b = -۰/۳۲$، پارامتر تشخیص $a = ۰/۲۶$ و حدس $c = ۰/۳$ است. با توجه به پارامتر دشواری سؤال می‌توان گفت احتمال پاسخگویی درست برای افرادی هم که در سطوح پایین توانایی قرار داشته‌اند، زیاد بوده است. با بالا رفتن توانایی احتمال پاسخ درست به کندی افزایش یافته است طوری که از این لحاظ بین سطوح مختلف توانایی، تفاوت‌ها چندان محسوس نیست. با توجه به پایین بودن مقدار پارامتر تشخیص می‌توان گفت که سؤال نمی‌تواند تفاوت‌ها را در بین توانایی‌های بالاتر از متوسط و در بین سطوح پایینی به صراحت شناسایی کند.

اشخاصی را که در سطح بالا توانایی دانش زبان قرار گرفته‌اند، شناسایی کرد.

سؤالات گروه ششم نیز بار نیل به هدف مذکور مناسب نیست زیرا اشخاصی را که در مقیاس توانایی در سطوح پایینتر از متوسط ($\theta < ۰$) قرار گرفته‌اند را نمی‌توانند مشخص نمایند با این گروه تنها می‌توان افرادی را که در مقیاس توانایی در سطوح متوسط و بالا هستند مشخص کرد. با این وجود می‌توان ادعا کرد که سوالات این گروه و گروه بعدی مناسب‌ترین سوالات این آزمون بوده‌اند.

سؤالات گروه هفتم با توجه به داشتن پارامتر تشخیص مناسب و پارامتر دشواری پایین تقریبا تا سطوح توانایی بالا یعنی ($\theta = ۲$) خیلی خوب عمل کرده‌اند و به خوبی افراد را تفکیک کرده‌اند، اما قادر به تمایزگذاری بین تواناترین افراد و افراد توانا نبوده‌اند. همان طورکه در مورد گروه قبلی گفته شد به رغم این مسایل می توان گفت به لحاظ مقایسه‌ای این گروه از سوالات را می‌توان مناسبترین سوالات این آزمون قلمداد کرد.

گروه هشتم هم که تنها شامل یک سؤال می‌شود برای سنجش سطوح مختلف توانایی مناسب نیست. زیرا اگر چه به علت پایین بودن مقدار پارامتر دشواری سوال، افراد از سطوح پایین‌تر توانایی قادر به پاسخگویی درست به این سؤال بوده‌اند و احتمال پاسخگویی درست نیز با بالا رفتن توانایی افزایش یافته است، اما به دلیل پایین بودن مقدار پارامتر تشخیص سؤال می‌توان گفت که این سوال نمی‌تواند افراد را در سطوح مختلف توانایی به صراحت تفکیک نماید.

آنچه که تاکنون ذکر شد در مورد مناسب بودن سؤالها با فرض این موضوع بود که هدف از آزمون برگزار شده، سنجش تمام سطوح مختلف توانایی دانش زبان بوده است.

حال اگر هدف از اجرای آزمون شناسایی و تعیین افراد با توانایی متوسط و بالا در زمینه دانش زبان باشد، سؤالات آزمون باید به گونه‌ای باشند که از سطوح توانایی متوسط به بالا افراد قادر به پاسخگویی درست به سؤالات باشند و با افزایش سطح توانایی، احتمال پاسخگویی درست نیز افزایش یابد. در واقع دشواری سؤالات باید متناسب با سطوح توانایی متوسط و بالا باشد تا بتوان افرادی را که در این سطوح قرار دارند مشخص کرد. علاوه بر این سؤالات باید از پارامتر تشخیص مناسبی نیز برخوردار باشند بر این اساس سؤالات گروه اول و دوم و سوم مناسب نمی‌باشند، زیرا همان طورکه گفته شد، مقدار پارامتر دشواری سؤالات این سه گروه بالا است و احتمال پاسخگویی درست برای تواناترین افراد فقط ۰٫۵ بوده است و مقدار این احتمال برای افراد با توانایی متوسط هم بسیار اندک بوده است بنابراین با این سؤالات تا حدودی می‌توان فقط تواناترین افراد را در زمینه دانش زبان مشخص کرد.

ب-- برآورد توانایی آزمودنی‌ها در آزمون تولیمو

متاسفانه هیچ راهی برای از پیش دانستن پارامتر واقعی توانایی وجود ندارد و بهترین راه تنها برآورد آن است. به لحاظ منطقی می‌توان گفت مقدار متوسط برآوردهایی که توسط کامپیوتر محاسبه می‌شود به مقدار پارامتر توانایی آزمودنی‌ها نزدیک است. وقتی پارامتر دشواری سؤالات برابر یا نزدیک به پارامتر توانایی آزمودنی باشد، میانگین توانایی برآورد شده نزدیک به مقدار توانایی وی است. نکته حائز اهمیت در برآورد توانایی آزمودنی‌ها در نظریه سؤال-پاسخ این است که توانایی آزمودنی نسبت به سؤالات ثابت است یعنی نسبت به سؤال‌هایی که برای سنجش توانایی

مورد نظر به کار می‌روند نامتغیر است. این بدین معنا است که از یک آزمون با هر جایگاهی که در سرتاسر مقیاس توانایی داشته باشد (یعنی با هر میزانی از دشواری) می‌توان برای برآورد توانایی آزمودنی استفاده کرد. بر این اساس توانایی هر یک ازامتحان دهندگان آزمون تولیمو با استفاده از نرم‌افزارهای تهیه شده برآورد شد (جدول نتایج پیوست). برای هر آزمودنی مقدار توانایی، نمره براساس توانایی، نمره خام (روش کلاسیک)، رتبه‌بندی براساس نمرات توانایی و بر اساس نمرات خام و اختلاف دو روش رتبه‌بندی ارائه شده است.

در آزمون تولیمو ۷۸۴ آزمودنی شرکت کرده‌اند که بر اساس جدول نتایج پیوست دامنه توانایی آزمودنی‌ها بین ۳٫۳۵۸ـ تا ۴٫۶۶۵+ است. توزیع فراوانی توانایی آزمودنیها به قرار زیر است: ۹۰ آزمودنی توانایی بالاتر از ۳+، ۴۰۷ آزمودنی توانایی بین ۲+ تا ۳+، ۲۴۶ آزمودنی توانایی بین ۱+ تا ۲+، ۳۱ آزمودنی توانایی بین صفر تا ۱+ و ۱۰ آزمودنی توانایی زیر صفر دارند.

محاسبه نمره در هر سطح توانایی و مقایسه رتبه‌هاـ بعد از برآورد توانایی هر یک از آزمودنی‌ها مرحله بعدی محاسبه نمره هر یک از آنها است. برای محاسبه نمره آزمودنی‌ها در هر سطح از توانایی، احتمال پاسخگویی درست به تک تک سؤالاتی را که آزمودنی به آن پاسخ درست داده است را با هم جمع می‌کنیم. در آزمون تولیمو دامنه نمرات محاسبه شده بر اساس نظریه سؤال-پاسخ بین ۱٫۵۴ تا ۸۵٫۸۸ است.

پس از محاسبه نمرات هر یک از آزمودنی‌ها، رتبه‌بندی آنها بر اساس نمرات نظریه سؤال-پاسخ و نمرات خام (روش کلاسیک) انجام پذیرفت. آنگاه نمرهٔ آزمودنی‌ها بر پایه هر دو روش نمره‌گذاری، رتبه‌بندی شد و نتایج نیز با یکدیگر مقایسه شد. همانطور که در جدول نتایج پیوست مشاهده می‌شود در بعضی موارد نمرات خام آزمودنی‌ها یکسان است، اما نمرات محاسبه شده آنها بر اساس نظریه سؤال-پاسخ یکسان نیست، چون احتمال پاسخگویی درست به سؤالات یکسان و یا مختلف، در سطوح مختلف توانایی متفاوت است. در واقع بر اساس نظریه سؤال-پاسخ هیچ دو آزمودنی نمره مشابه کسب نخواهند کرد و تفاوتی هر چند اندک بین نمرات مشاهده می‌شود.

مقایسه رتبه‌بندی آزمودنی‌ها نشان می‌دهد که فقط در ۳۱ مورد بین دو روش رتبه‌بندی تفاوتی وجود ندارد و این آزمودنی‌ها بر اساس هر دو روش رتبه یکسانی را کسب کرده‌اند. به عنوان مثال از رتبه ۱ تا ۶ نتایج مشابه هستند و هم چنین رتبه‌های ۷۸۱ تا ۷۸۴ نیز یکسان هستند به این معنا که یک آزمودنی بر اساس هر دو روش رتبه ۷۸۴ را کسب کرده است. همچنین آزمودنی دیگری نیز بر اساس هر دو روش نمره‌گذاری رتبه ۳۱۰ را کسب کرده است.

اما در بیشتر موارد بین رتبه‌ها همسانی وجود ندارد. در بعضی از موارد اختلاف رتبه‌ها ناچیز است و در بعضی از موارد این اختلاف خیلی چشمگیر است. به عنوان مثال آزمودنی خاصی، بر اساس نظریه سؤال-پاسخ، رتبه ۴۶۸ را کسب کرده است و بر پایه روش کلاسیک رتبه‌اش ۵۷۳ شده است و اختلاف بین دو روش رتبه‌گذاری برابر ۱۰۵ است. شاید یکی از دلایل این تفاوت‌های چشمگیر بین دو روش رتبه‌گذاری این باشد که این قبیل آزمودنی‌ها به سؤالاتی پاسخ درست داده‌اند که بر طبق تحلیل صورت گرفته مبتنی بر نظریه سؤال-پاسخ، سؤالات مناسبی برای اهداف سنجش بوده‌اند.

اختلاف بین رتبه‌ها، دو الگو دارد. در الگوی اول رتبه محاسبه شده بر اساس نظریه سؤال-پاسخ کمتر از رتبه محاسبه شده بر اساس روش کلاسیک است و در واقع اختلاف بین رتبه‌ها مثبت شده است. به عنوان مثال یک آزمودنی بر اساس نظریه سؤال-پاسخ رتبه ۹۰ کسب کرده است، اما بر اساس نمرات خام (روش کلاسیک) رتبه ۱۰۹ کسب کرده که در اینجا اختلاف رتبه برابر ۱۹ است. یا نمونه دیگر آزمودنی دیگری بوده است که بر اساس نظریه سؤال-پاسخ رتبه ۳۵۴ کسب کرده است، اما بر اساس روش کلاسیک رتبه ۴۱۴ را کسب کرده، که اختلاف بین دو رتبه ۵۷ است.

در الگوی دوم رتبه کسب شده بر اساس نمرات خام (روش کلاسیک) بالاتر از رتبه های مبتنی بر نظریه سؤال-پاسخ است. برای مثال یکی از آزمودنی‌ها بر پایه نمرات خام رتبه‌اش ۱۱۸ شده است، اما بر پایه رویه سؤال-پاسخ رتبه‌اش ۱۳۸ شده است که اختلافی برابر ۲۰ بین دو روش رتبه‌بندی وجود دارد. یا نمونه دیگر آزمودنی است که بر اساس رویه کلاسیک رتبه‌اش ۴۷۶ شده است، اما طبق رویه سؤال-پاسخ رتبه ۵۲۵ را کسب کرده است که اختلافی برابر ۴۹ بین دو روش مذکور مشاهده می‌شود.

همان طور که قبلا ذکر شد در ۳۱ مورد بین دو روش رتبه‌گذاری تفاوتی وجود ندارد اما در سایر موارد رتبه‌ها یکسان نیستند. در بعضی موارد اختلاف رتبه‌ها ناچیز و در بعضی موارد بسیار فاحش است. به همین منظور تفاوت بین رتبه‌های دو روش نمره‌گذاری را در ۴ طبقه با فاصله طبقاتی ۲۵ گروه‌بندی و فراوانی هر طبقه مشخص شده است (جدول ۲). همان طور که در جدول مشخص است اختلاف رتبه زیر ۲۵ بیشترین فراوانی را دارد و با بیشتر شدن اختلاف رتبه‌ها فراوانی کمتر می‌شود. چنانچه فراوانی تفاوت رتبه بالاتر از ۷۶ کمترین مقدار را دارد یعنی همان طورکه از جدول ۳ پیداست بیشترین جابجایی در رتبه‌ها را می‌توان درگروه افزایش رتبه‌های از ۱ تا ۲۵ دید که رتبه ۲۸۰ نفر تغییر کرده است و این میزان تغییر در یک جمعیت حدود ۸۰۰ نفره تغییر ناچیزی به حساب نمی‌آید. به ویژه که باید توجه داشت که به همین نسبت جابجایی قرینه‌وار یعنی کاهش رتبه را شاهد بوده‌ایم. پس از آن تغییرات را در افزایش رتبه‌ها از ۲۶ تا ۵۰ رتبه شاهد هستیم و کمترین تغییرات را درگروه پایینی تغییرات رتبه‌ها می‌توان سراغ گرفت.

برای مقایسه بیشتر بین نمرات و رتبه‌های دو شیوه سنتی و IRT همبستگی بین سطوح مختلف آن محاسبه شد. جالب توجه است که همبستگی بین نمرات و رتبه‌ها در دو سطح ده درصدی بالا و پایین نمرات بیش از ۰٫۹۸ بود. همبستگی بین سطوح میانی نمرات نیز بیش از ۰٫۹۶ بوده است. محدود بودن تغییرات فاحش بین دو روش حاکی از آن است که اندازه‌گیری و نمره‌دهی بر حسب نظریه سؤال-پاسخ به تغییرات فاحش در جابجایی رتبه‌ها نمی‌انجامد، بلکه با توجه به دقت در تعیین پارامترهای سوالات و تکرار محاسبات در برآورد توانایی در چندین دور محاسباتی می‌توان استنتاج کرد که این شیوه اندازه‌گیری تصویری دقیقتر از نتایج به دست می‌دهد، علاوه بر این می‌توان ادعا کرد که حتی اگر شاهد همین تغییرات هم نبودیم حداقل نتایج حاصل که مزیت IRT نسبت به رویه سنتی است عبارت است از فراهم ساختن امکان برآورد تواناییها، تعیین احتمال پاسخگویی به هر سؤال روی سطوح مختلف توانایی، ترسیم مقیاس واحدی برای

جدول ۲: توزیع فراوانی جابجایی رتبه‌ها درکل نمرات توانایی صرف نظر ازکاهش یا افزایش رتبه

اختلاف رتبه	فراوانی
۰	۳۱
۱-۲۵	۵۷۱
۲۶-۵۰	۱۳۷
۵۱-۷۵	۳۶
۷۶ به بالا	۹
جمع کل	۷۸۴

جدول ۳: توزیع فراوانی اختلاف رتبه درکل نمرات توانایی بر حسب افزایش وکاهش رتبه

اختلاف رتبه	فراوانی
۱-۲۵	۲۸۰
۲۶-۵۰	۶۸
۵۱-۷۵	۱۸
۷۶ به بالا	۶
۰	۳۱
-۱-(-۲۵)	۲۹۱
-۲۶-(-۵۰)	۶۹
-۵۱-(-۷۵)	۱۸
۷۶ به پایین	۳
جمع کل	۷۸۴

نشان دادن مشخصات امتحان دهندگان و سؤالات، و ترسیم نمودار مشخصه‌های سؤالات. اینها امکانات فوق العاده‌ای است که IRT علاوه بر در اختیارگذاشتن نتایج دقیقتر، فراروی ما می‌گذارد. پیش‌بینی می‌شود با حذف سؤالاتی که با هدف سنجش همخوانی ندارند، و قدرت تمیزکافی بین توانایی‌های مختلف ندارند، در برآورد توانایی آزمودنی‌ها تفاوتهایی مشاهده شود.

۸ برازش مدل و داده‌ها

همان طورکه قبلا گفته شد مدل انتخابی برای تحلیل آزمون و برآورد توانایی‌ها، مدل سه پارامتری بوده است. مقادیر حاصل از محاسبه مربع خی در اکثر موارد از عدد معیار بزرگتر نبود. با این حال این دو موضوع بایستی مورد بررسی قرار می‌گرفت: یکی موضوع تک بعدی یا چند بعدی بودن مدل و دیگری مسئله نامتغیر بودن پارامترها.

جدول ۴: توزیع فراوانی اختلاف رتبه در ۲۵ درصد بالای نمرات توانایی

اختلاف رتبه	فراوانی	درصد	درصد تراکمی
۱-۲۵	۹۳	۴۷/۴	۴۷/۴
۰	۱۶	۸/۲	۵۵/۶
-۱-(-۲۵)	۸۷	۴۴/۴	۱۰۰
جمع کل	۱۹۶	۱۰۰	

جدول ۵: توزیع فراوانی اختلاف رتبه در ۲۵ درصد پایین نمرات توانایی

اختلاف رتبه	فراوانی	درصد	درصد تراکمی
۱-۲۵	۶۲	۳۱/۶	۳۱/۶
۲۶-۵۰	۱۳	۶/۶	۳۸/۲
۵۱-۷۵	۲	۱/۰	۳۹/۲
۷۶ به بالا	۱	۰/۵	۳۹/۷
۰	۱۲	۶/۱	۴۵/۸
-۱-(-۲۵)	۸۱	۴۱/۳	۸۷/۲
-۲۶-(-۵۰)	۲۱	۱۰/۷	۹۸
-۵۱-(-۷۵)	۳	۱/۵	۹۹/۵
۷۶ به بالا	۱	۰/۵	۱۰۰
جمع کل	۱۹۶	۱۰۰	

جدول ۶: توزیع فراوانی اختلاف رتبه در میانه توزیع نمرات

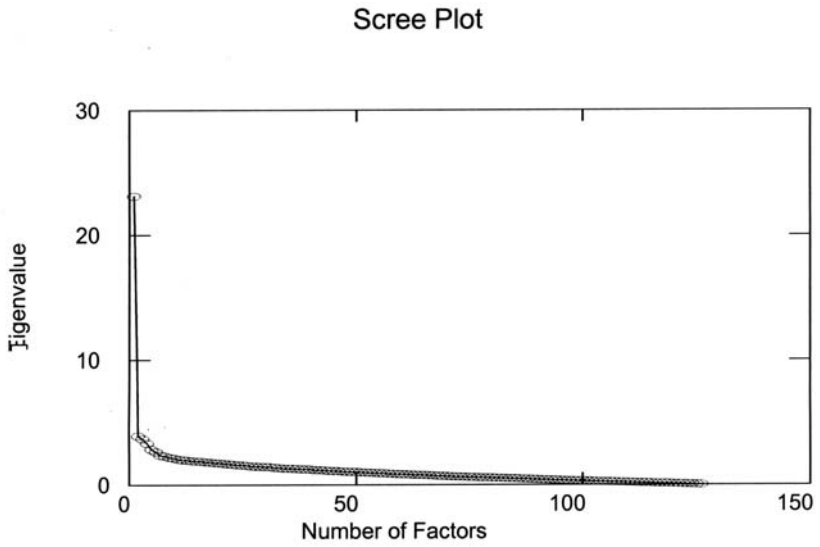
اختلاف رتبه	فراوانی	درصد	درصد تراکمی
۱-۲۵	۱۲۵	۳۱/۸	۳۱/۸
۲۶-۵۰	۵۵	۱۴	۴۵/۸
۵۱-۷۵	۱۶	۴/۱	۴۹/۹
۷۶ به بالا	۵	۱/۳	۵۱/۲
۰	۳	۰/۸	۵۲
-۱-(-۲۵)	۱۲۳	۳۱/۴	۸۳/۲
-۲۶-(-۵۰)	۴۸	۱۲/۲	۹۵/۷
-۵۱-(-۷۵)	۱۰	۳/۸	۹۹/۵
۷۶ به بالا	۲	۰/۵	۱۰۰
جمع کل	۳۹۲	۱۰۰	

۹ ابعاد آزمون

اغلب مدل‌های اندازه‌گیری این فرض را دستمایه کار خود قرار می‌دهند که مفهومی که داریم اندازه می‌گیریم یک بعدی است، یعنی تنها یک عامل برجسته است که مبنای توضیح آن رفتار خاص است. رویکردهای مختلفی برای وارسی و ارزیابی تک بعدی بودن مدل پیشنهاد شده است.(برای اطلاع از آنها می‌توان به کتاب هاتی ([5]) مراجعه کرد). روش‌های رایج عبارت است از تعیین تعداد مقادیر ویژه‌ای[19] که بزرگتر است از ۱، بررسی نمودار اسکری[20]، و مقایسه نسبت اولین مقدار ویژه به دومین مقدار ویژه. در این تحلیل ما از نمودار اسکری و مقایسه نسبت اولین مقدار ویژه با دومین مقدار ویژه استفاده کرده‌ایم که بر مبنای تکنیک تحلیل عاملی به بررسی ابعاد آزمون می‌پردازد. البته در اینجا برای تعیین تعداد عوامل دخیل در مقیاس، به جای روش تحلیل عناصر اصلی[21] (PCA)، از روش عامل یابی محور اصلی[22] (PAF) استفاده کرده‌ایم. زیرا در روش تحلیل عامل یابی محور اصلی فقط واریانس مشترک در محاسبات در نظر گرفته می‌شود در حالی که در روش تحلیل عناصر اصلی هم واریانس اشتراکی و هم واریانس اختصاصی در محاسبات دخیل است.

همان طور که قبلا گفته شد سوالات آزمون تولیمو دو انتخابی هستند یعنی پاسخ آنها از دو حالت صحیح یا غلط بیرون نیست. درباره این گونه داده‌های دو انتخابی (صحیح یا غلط) لازم است بدانیم که تحلیل عاملی بر روی همبستگی تتراکوریک پاسخ‌ها انجام می‌شود. از بین نرم‌افزارهای معروف برای انجام این محاسبه از نرم افزار SYSTAT 10.2استفاده شد که در محاسبه همزمان همبستگی‌های تتراکوریک و عامل یابی محور اصلی کارایی چشمگیری دارد. در زیر ابتدا مقادیر اولین مقدار ویژه (۲۳/۱۰۸) تا بیستمین مقدار ارایه شده است. با مشاهده نخستین مقدار ویژه می‌توان دریافت که اختلاف آن با مقادیر بعدی بسیار فاحش است. در قسمت بعدی نیز سهم هر یک از عوامل در واریانس همبستگی‌ها دیده می‌شود که سهم عامل نخست نسبت به سایر عوامل بسیار بیشتر است. همچنین نمودار اسکری نیز نشان می‌دهد که عامل‌های دوم به بعد در جایی از نمودار قرار گرفته‌اند که نمودار به حالت تخت نزدیک شده است در حالی که عامل اول با فاصله بسیاری در بالای نمودار قرار گرفته است. بنابر این می‌توان از تک بعدی بودن آزمون و مقیاس مطمئن بود.

19) Eigenvalues 20) scree plot 21) Principal Components Analysis 22) Principle Axis Factoring



مقادیر ویژه

۱	۲	۳	۴	۵
۲۳٫۱۰۸	۳٫۹۲۵	۳٫۶۹۳	۳٫۳۳۶	۲٫۸۴۰
۶	۷	۸	۹	۱۰
۲٫۶۵۰	۲٫۳۵۴	۲٫۳۰۱	۲٫۱۹۰	۲٫۱۳۱
۱۱	۱۲	۱۳	۱۴	۱۵
۲٫۰۳۵	۱٫۹۸۹	۱٫۹۶۵	۱٫۹۰۹	۱٫۸۸۷
۱۶	۱۷	۱۸	۱۹	۲۰
۱٫۸۷۹	۱٫۸۲۱	۱٫۷۹۹	۱٫۷۵۲	۱٫۷۴۱

از آنجایی که بررسی ابعاد مدل نشان داد مقیاس ما یک بعدی است ناچار باید بررسی شود که آیا پارامترهای سوالات و توانایی آزمودنی‌ها نسبت به هم نامتغیر هستند یا نه.

برای این کار گروه چهارم از سوالات که ۳۴ سوال بودند را که با توجه به هدف آزمون سوالات مناسبی نبودند حذف گردید. یعنی این سوالات از قدرت تمیزکافی برخوردار نیستند و نمی‌توانستد در سطوح مختلف توانایی آزمودنی‌ها را به صراحت تفکیک نمایند. پس از حذف این سوالات مجددا برای هر یک از آزمودنی‌ها بر آورد توانایی بر اساس روش حداکثر درست‌نمایی، محاسبه نمره خام و نمره بر اساس نظریه IRT و رتبه‌بندی آزمودنی‌ها با توجه به ۱۰۶ سوال باقیمانده صورت گرفت. پس از آن نتایج این مرحله و مرحله قبل مورد مقایسه قرار گرفتند تا مشخص شود حذف سوالات چه تاثیری در برآورد توانایی آزمودنی‌ها دارد. بر این اساس میانگین تفاوت توانایی‌های برآورد شده با احتمال ۹۹ درصد در فاصله ۰٫۱۰۵۷- تا ۰٫۲۷۵۶ قرار گرفته‌اند.

بررسی نتایج نشان می‌دهد که در اکثریت موارد تفاوت بین توانایی‌های بر آورد شده بسیار ناچیز است و در فاصله برآورد شده قرار گرفته‌اند و میزان جا به جایی افراد در مقیاس توانایی با حذف سوالات قابل اغماض است. این بیانگر نامتغیر بودن توانایی آزمودنی‌ها نسبت به سوالهایی است که برای برآورد توانایی به کار می‌روند.

برای بررسی استقلال پارامترهای سوالات از نمونه، ۱۰۰ آزمودنی به طور تصادفی از میان ۷۸۴ آزمودنی انتخاب شد. محاسبات پارامترهای سوالات تفاوت معنی‌داری را نشان نمی‌داد.

مراجع

[1] Baker, F. B. (2001), The basics of item response theory.

[2] Baker, F. B. (1992), Item response theory: Parameter estimation techniques NewYork Marcel Dekker.

[3] Birnhaum, A. (1968), Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley Publishing.

[4] Drasgow, F., Levine M.V., Tsien, S., Williams B.A., & Mead, A.D. (1995), Fitting polytomous item response theory models to multiple-choice tests. Applied Psychological Measurement, 19, 143-165

[5] Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983), Item response theory Applications to psychological measurement. Homewood, IL: Dow Jones.