بنام خدا

ششمین کنفرانس بین‌المللی آمار ایران

مجموعه مقالات

۴ تا ٦ شهریور ماه ۱۳۸۱

دانشگاه تربیت مدرس

تهران، ایران

خلاصه مقالات ششمین کنفرانس بین‌المللی آمار ایران

تیراژ : ۱۰۰۰

تنظیم و صفحه‌آرایی : رحیم صفری

چاپ و صحافی : سازمان سنجش آموزش کشور

تاریخ انتشار : تابستان ۱۳۸۱

بنام خدا

پیشگفتار

با یاری خداوند متعال و همکاری انجمن آمار ایران و همت والای همکاران و دانشجویان محترم و حمایت مسئولین محترم دانشگاه و سازمانها و دستگاههای مختلف ششمین کنفرانس بین المللی آمار ایران در روزهای ۴ تا ٦ شهریور ماه۱۳۸۱ در دانشگاه تربیت مدرس برگزار می‌شود. استقبال چشمگیر اساتید و دانشجویان دانشگاه و همچنین محققین و دست اندرکاران فعالیتهای آماری در دستگاهها و سازمانهای مختلف در ارسال مقالات و گزارشات علمی در زمینه‌های مختلف علوم آماری موجب گردید این کنفرانس از سطح علمی مطلوبی برخوردار شود و زمینه‌های ارتباط بین دانشگاه و کاربران آمار در دستگاههای مختلف و واحدهای صنعتی بیش از گذشته فراهم گردد.

متقاضیان شرکت در کنفرانس تعداد ۳۴۰ مقاله به دبیرخانه ارسال نمودند که کمیسیونهای تخصصی شش‌گانه کمیته علمی کنفرانس تعداد ۱۵۷ مورد را بعنوان سخنرانی، تعداد ۷۳ مقاله را برای ارائه بصورت پوستر و ۵ مورد را بصورت کارگاه آموزشی پذیرش نمودند و بدلیل محدودیت مدت زمان برگزاری کنفرانس ٦۹ مقاله را واجد اولویت لازم برای ارائه در کنفرانس ندانستند. بعلاوه از دانشمندان صاحب نام علم آمار داخلی و خارج کشور برای ارائه سخرانی‌های عمومی و تخصصی دعوت بعمل آمد که تعدادی از آنها به دلایلی از شرکت در کنفرانس عذرخواهی نمودند و نهایتاً بالغ بر ۳٦ مقاله توسط مدعوین داخلی و خارجی در کنفرانس ارائه شد.

برگزاری ششمین کنفرانس آمار ایران بدون همکاری صمیمانه تعدادی از اساتید دانشگاههای سراسر کشور بعنوان هیات داوران و تلاشهای خستگی ناپذیر همکاران گرامی دانشگاه علی‌الخصوص همکاران دانشکده علوم پایه و دانشجویان و دانش آموختگان محترم دانشگاه امکان پذیر نبود و جا دارد از تمامی عزیزانی که در برگزاری کنفرانس زحماتی را متحمل شدند قدردانی و تشکر شود.

انجام کلیه امور کامپیوتری و تهیه و تنظیم مجموعه‌های خلاصه و اصل مقالات با همکاری و پشت کار بی‌دریغ و خستگی ناپذیر آقایان رحیم صفری، امیر کاووسی، شهاب جولانی و سرکار خانم ثریا خفری صورت پذیرفت که از این عزیزان نهایت تشکر و امتنان را دارد.

حمایت کنندگان:

این کنفرانس با حمایتها و پشتیبانی انجمن آمار و سازمانها و موسسات زیر برگزار می‌گردد که بدینوسیله نهایت قدردانی خود را از آنانی که تا زمان چاپ این مجموعه کنفرانس را مورد حمایت‌های مختلف قرار دادند ابراز می‌داریم.

۱– وزارت علوم، تحقیقات و فناوری
۲– وزارت نفت
۳– وزارت پست و تلگراف و تلفن
۴– وزرات جهاد کشاورزی
۵– وزارت نیرو و برق منطقه تهران
٦– وزارت صنایع و معادن
۷– سازمان سنجش آموزش کشور
۸– نهاد ریاست جمهوری
۹– موسسه پژوهش و برنامه ریزی آموزش عالی
۱۰– مرکز آمار ایران
۱۱– پژوهشکده آمار
۱۲– پژوهشکده تربیت بدنی
۱۳– پژوهشکده هوا فضا
۱۴– پژوهشکده صنایع رنگ
۱۵– دانشگاه امام حسین
۱٦– انستیتو پاستور
۱۷– بانک رفاه کارگران
۱۸– بانک کشاورزی
۱۹– شرکت مهندسی سیستم یاس
۲۰– سازمان انتقال خون
۲۱– شرکت هواپیمایی جمهوری اسلامی ایران
۲۲– کمیسیون یونسکو
۲۳– دانشگاه پیام نور
۲۴– وزارت کار و امور اجتماعی

دبیر ششمین کنفرانس بین المللی آمار ایران

شهریور ۱۳۸۱

اعضای کمیته برگزاری:

۱– دکتر محسن محمدزاده (مسئول کمیته و دبیر کنفرانس)
۲– دکتر محمدقاسم وحیدی اصل (مسئول کمیته علمی)
۳– دکتر سقراط فقیه زاده (مسئول کمیته برنامه‌ریزی)
۴– دکتر انوشیروان کاظم‌نژاد (مسئول کمیته اجرایی)
۵– دکتر غلامرضا بابایی (مسئول کمیته تبلیغات)
٦– دکتر سید محمد حسینی

اعضای کمیته علمی کنفرانس:

۱– دکتر ابوالقاسم وحیدی اصل (مسئول کمیته)	دانشگاه شهید بهشتی
۲– دکتر محسن محمد زاده (دبیر کنفرانس)	دانشگاه تربیت مدرس
۳– دکتر محمد رضا مشکانی	دانشگاه شهید بهشتی
۴– دکتر احمد پارسیان	دانشگاه صنعتی اصفهان
۵– دکتر سقراط فقیه زاده	دانشگاه تربیت مدرس
٦– دکتر انوشیروان کاظم نژاد	دانشگاه تربیت مدرس
۷– دکتر حمید رضا نواب پور	دانشگاه علامه طباطبایی
۸– دکتر یدا... محرابی	دانشگاه علوم پزشکی شهید بهشتی
۹– دکتر عباس گرامی	پژوهشکده آمار

فهرست مندرجات

عنوان صفحه

اثرات طبقه‌بندی نادرست در تحلیل داده‌های گسسته

علیرضا ابدی [۱]، کاظم محمد [۲]، محمدرضا مشکانی [۳]، انوشیروان کاظم نژاد [۴]

$P۱۱۰۷۰$

[۱] دانشگاه تربیت مدرس
[۲] دانشگاه علوم پزشکی تهران
[۳] دانشگاه شهید بهشتی
[۴] دانشگاه تربیت مدرس

چکیده: طبقه‌بندی نادرست از منابع اریبی است که باعث کاهش کارایی در تحلیل داده‌های گسسته می‌شود. در این مقاله با هدف معرفی خطای طبقه‌بندی نادرست ابتدا تاریخچه این موضوع و مباحث مهم مطرح شده پیرامون آن ارائه و سپس انواع منابع خطا در طبقه‌بندی نادرست بررسی شده است. به منظور نشان دادن چگونگی تاثیر این خطا در تحلیل نتایج مطالعات، شاخصهای مهم قابل محاسبه در یک جدول دو در دو که شامل میزانهای وقوع، خطر نسبی، تعداد موردنیاز برای تیمار و نسبت بخت می‌باشد معرفی و آنگاه در قالب یک مثال فرضی به بررسی تاثیر طبقه‌بندی غلط در نتایج حاصله پرداخته شده است. تاثیر طبقه‌بندی نادرست وقتی که هر یک از متغیرهای وضعیت بیماری یا در معرض خطر بودن یا هر دو متغیرتواماً غلط طبقه‌بندی شده باشند مورد بررسی قرار گرفته و مهمترین روشهای تعدیل تاثیر طبقه‌بندی نادرست براساس روشهای ماتریسی، مدل سازی و روشهای بیزی تشریح شده است.

واژه‌های کلیدی: خطای طبقه‌بندی نادرست، تحلیل داده‌های گسسته.

۱ مقدمه

خطای اندازه‌گیری متغیرهای با مقیاس اسمی یا رتبه‌ای عموماً خطای طبقه‌بندی نادرست [۱] نامیده می‌شود. این خطا در هر نوع مطالعه‌ای ممکن است واقع شود.

در مطالعات گذشته‌نگراگر تعریف دقیق و روشنی از مواجهه با خطر نداشته باشیم انتساب افراد به گروههای موردنظر با خطا همراه خواهد بود.

[۱] Misclassification Error

همچنین در مواردی که در تشخیص بیماری از تستهای آزمایشگاهی استفاده می‌نمائیم و این روشها از حساسیت [۲] و ویژگی [۳] صددرصد برخوردار نباشد انتساب افراد به گروههای بیمار و غیربیمار نیز با خطا همراه خواهد بود که این امر بیشتر در مطالعات آینده‌نگر واقع می‌شود.

در برخی مطالعات نیز بیماری و مواجهه با خطر تواماً با خطای اندازه‌گیری همراه است که بیشتر در مطالعات مقطعی ممکن است واقع شود.

اندازه گیری غیردقیق مواجهه با خطر، تشخیص غلط، گزارشهای فردی غیر صحیح یا هر عامل دیگری ممکن است باعث شود تا فردی بطور غلط در یکی از خانه‌های جدول توافقی قرار گیرد. خطای فوق الذکر عامل انحراف در تعیین شاخصها و تحلیل داده‌ها می گردد.

در بسیاری از تحقیقها متاسفانه بدون توجه به احتمالات طبقه‌بندی نادرست به تحلیل نتایج پرداخته می‌شود که می‌تواند نتایج گمراه کننده‌ای به همراه داشته باشد. در نتیجه توجه به ساختار احتمالات طبقه‌بندی نادرست برای دستیابی به نتایج واقع بینانه ضروری است.

موضوع طبقه‌بندی نادرست ابتدا توسط براس (۱۹۵۴) مورد توجه قرار گرفته است. توسط چن (۱۹۸۹) مطالعات مروری در این خصوص صورت پذیرفته و تنبین (۱۹۷۲) روش برآورد براساس نمونه‌گیری دو مرحله‌ای را با وجود طبقه‌بندی نادرست ارائه نموده و توسط هیتوسکی و راپ (۱۹۹۲) به دوبار نمونه‌گیری شرطی تعمیم داده شده است. اسپلندوهویی (۱۹۸۷) مدلهای عمومی خطی را برای داده‌های ناکامل براساس اطلاعات کمکی بکار برده‌اند. لیو و لیانگ (۱۹۹۱) تصحیح خطای طبقه‌بندی نادرست را براساس مدلهای عمومی خطی ارائه نموده‌اند. هیرونوری و شیز (۲۰۰۰) استنباط در مورد احتمالات طبقه‌بندی نادرست براساس پاسخهای تکرار شده دوتایی را ارائه نموده‌اند.

۲ برخی شاخصها و تحلیلها در جداول دو طرفه

تحلیل داده‌ها در بسیاری از مطالعات با محاسبه برخی شاخصها یا میزانها در جداول توافقی همراه می‌باشد. جدول فرضی زیر را در نظر بگیرید:

---

[۲] Sensitivity
[۳] Specificity

	بیمار	غیر بیمار	جمع
مواجهه با خطر	A	B	$A+B$
عدم مواجهه با خطر	C	D	$C+D$
جمع	$A+C$	$B+D$	

برخی شاخصهای قابل محاسبه در این جدول را داوسون و تراپ به‌شرح زیرارائه نموده‌اند: ۱) میزان وقوع در گروه تجربه

$$EER = \frac{A}{A+B}(Experimental\ Eevent\ rate)$$

۲) میزان وقوع در گروه کنترل

$$CER = \frac{C}{C+D}(Control\ Event\ Rate)$$

۳) قدر مطلق کاهش خطر

$$ARR = |EER - CER|\,(Absolute\ Risk\ Reduction)$$

۴) تعداد مورد نیاز تیمار برای جلوگیری از یک مورد بیماری

$$NNT = \frac{۱}{ARR}(Number\ Needed\ toTreat.)$$

۵) خطر نسبی

$$RR = \frac{EER}{CER}(Relative\ Risk)$$

٦) کاهش خطر نسبی

$$RRR = \frac{ARR}{CER}(Relative\ Risk\ Reduction)$$

۷) نسبت بخت

$$OR = \frac{[A/(A+C)]/[C/(A+C)]}{[B/(B+D)]/[D/(B+D)]} = \frac{A/C}{B/D} = \frac{AD}{BC}$$

براساس علامت گذاریهای فوق‌الذکر میزان وقوع ER (Event Rate) برای محاسبه اندازه خطر استفاده می شود. EER میزان وقوع در گروه تجربه است که براساس تعداد بیمار در گروه مواجهه یافته با عامل خطر محاسبه می‌شود. CER همین میزان در گروه کنترل است که مواجهه با عامل‌خطر نداشته‌اند. ARR نشانگر قدرمطلق کاهش خطر در مقایسه با خطر مبنا می‌باشد. NNT نمایشگر تعداد افرادی است که باید تحت این تیمار یا درمان قرار بگیرند تا انتظار داشته باشیم یکنفر به‌بیماری مبتلا نشود. به‌عبارتی این شاخص فایده درمان را مشخص می‌سازد.

RR نشانگر نسبت بروز در افراد مواجهه یافته با عامل خطر به بروز در افراد مواجهه نیافته می‌باشد. RRR نیز اندازه کاهش خطر نسبی براساس خطر مبنا را مشخص می‌کند. OR نیز شاخص برای بررسی خطر در مطالعات مورد شاهدی است که نسبت حاصلضرب متقاطع نیز گفته شده است. odds یا بخت ، نسبت احتمال وقوع واقعهٔ موردنظر به عدم وقوع آن است که هم می‌تواند برای بروز و هم می‌تواند برای شیوع تعریف شود. بخت و نسبت می‌توانند برای بیان فراوانی بیماری بکار روند. وقتی که نسبت مقدار کوچکی باشد بخت تقریبی از نسبت می‌باشد. برای مثال اگر نسبت سیگاریها در جامعه‌ای ۰/۲ باشد بخت آن عبارت است از:

$$Odds = \frac{۰/۲}{۰/۸} = ۰/۲۵$$

یعنی در جامعه مورد نظر برای هر یکنفر سیگاری ۴ نفر غیر سیگاری وجود دارد. نسبت بخت نیز حاصل تقسیم‌بخت در گروه مورد به‌بخت در گروه شاهد است.

۳ تاثیر طبقه‌بندی نادرست در تحلیل اطلاعات

با افزایش تعداد متغیرهایی که مواجه به طبقه‌بندی نادرست می‌باشند تحلیل اطلاعات پیچیده‌تر خواهد شد. در دائره المعارف آمارزیستی موضوع براساس آنالیز تک متغیر و دو متغیره و چندمتغیره بررسی شده است. در حالت تک متغیره A را متغیر گروه‌بندی شده واقعی و A^* را متغیر گروه بندی شده نادرست در نظر می‌گیریم. با فرض استقلال بین واحدها احتمال طبقه‌بندی نادرست بصورت

$$P_r = (A^* = j|A = k) = \theta_{j,k} \qquad j,k = ۱,۲,\ldots,m$$

است که پارامتر $\theta_{j,k}$ ساختار طبقه‌بندی غط را نشان می‌دهد و می‌تواند به صورت ماتریس $\phi = [\theta_{j,k}]_{m\times m}$ با عضوهای غیر منفی و جمع ستونهای یک باشد. برای متغیرهای پاسخ دوتایی با فرض A=2(وجود بیماری) و A=1 (عدم وجود بیماری) این ماتریس به صورت

$$\begin{pmatrix} \beta & ۱-\alpha \\ ۱-\beta & \alpha \end{pmatrix} \qquad (۱)$$

می‌باشد که α نشان دهنده حساسیت و β نشان دهنده ویژگی می‌باشد. اثر بکارگیری طبقه‌بندی جایگزین A^* می‌تواند بصورت زیر خلاصه شود:

$$\Pi_{A^*} = \phi\pi_A \qquad (۲)$$

در حالیکه

$$\Pi_{A^*} = (\Pi_{A^*}(۱), \ldots, \Pi_{A^*}(m))'$$

و

$$\Pi_A = (\Pi_A(۱), \ldots, \Pi_A(m))'$$

نسبتهای جامعه در گروه متغیر جانشین و متغیر اصلی می‌باشند. بنابراین نسبتهای نمونه A^* برآوردهای اریبی‌از π_A خواهند بود: ماهیت این اریبی در حالت دوتایی براساس رابطه (۲) بصورت زیر خواهد بود:

$$\Pi_{A^*}(۱) = \beta\Pi_A(۱) + (۱-\alpha)\Pi_A(۲) \qquad (۳)$$

$$\Pi_{A^*}(۲) = (۱-\beta)\Pi_A(۱) + \alpha\Pi_A(۲) \qquad (۴)$$

در آنالیز دو متغیره اگر A معرف وضعیت بیماری و B معرف وضعیت مواجهه با خطر باشد ابتدا حالتی را در نظر می‌گیریم که طبقه‌بندی غلط دررابطه با متغیر پاسخ باشد. در این حالت متغیر پاسخ A بوسیله A^* اندازه گیری می شود و $\Pi_{A|B}(j|l)$ نشانگر نسبت واحدهایی از جامعه که متغیر A در وضعیت j برای گروه در معرض عامل خطر با B=l باشد و همچنین $\Pi_{A^*|B}(j|l)$ نشانگر نسبت متناظر برای A^* باشد. اگر توجه به تفاوت نسبت پاسخ در دو گروه در معرض نمائیم آنگاه برآوردگر نااریب

$$\Pi_{A^*|B}(۲|۲) - \Pi_{A^*|B}(۲|۱)$$

در حالت کلی برای تفاضل واقعی

$$\Pi_{A|B}(۲|۲) - \Pi_{A|B}(۲|۱)$$

اریب خواهد بود. در حالتی که هر دو گروه در معرض خطر حساسیت α و ویژگی β یکسان داشته باشند مکانیزم طبقه‌بندی نادرست, غیر افتراقی [۱] نامیده می‌شود. در این حالت براساس رابطه (۴) خواهیم داشت.

$$\Pi_{A^*|B}(۲|۲) - \Pi_{A^*|B}(۲|۱) = (\alpha + \beta - ۱)[\Pi_{A|B}(۲|۲) - \Pi_{A|B}(۲|۱)] \quad (۵)$$

$\alpha + \beta - ۱$ مقادیری بین صفر و یک را می‌گیرد و تنها وقتی که یک باشد تفاضل اندازه‌گیری شده بوسیله A^* با مقدار واقعی A برابر خواهد بود. در حالت طبقه‌بندی نادرست غیر افتراقی نسبت

$$\frac{\Pi_{A|B}(۲|۲)}{\Pi_{A|B}(۲|۱)}$$

به سمت مقدار تحت فرض صفر کاهش خواهد یافت. در حالت دوم اگر متغیر پاسخ درست گروه‌بندی شده باشد و متغیر در معرض عامل خطر B غلط طبقه‌بندی شده باشد آنگاه B* گروه بندی غلط متغیر B خواهد بود. اگر طبقه‌بندی نادرست B غیر افتراقی با توجه به متغیر A باشد آنگاه

$$\Pi_{A|B^*}(۲|۲) - \Pi_{A|B^*}(۲|۱) = \frac{(\alpha_B - \beta_B - ۱)\Pi_B(۲)\Pi_B(۱)}{\Pi_{B^*}(۲)\Pi_{B^*}(۱)} \times \left[\Pi_{A|B}(۲|۲) - \Pi_{A|B}(۲|۱)\right]$$

که

$$\Pi_B = (\Pi_B(۱), \Pi_B(۲))' = (\Pi_{B^*}(۱), \Pi_{B^*}(۲))'$$

که به ترتیب نسبتهای جامعه B و B^* بوده و α_B و β_B حساسیت و ویژگی طبقه‌بندی B هستند. فاکتور $\left[\Pi_{A|B}(۲|۲) - \Pi_{A|B}(۲|۱)\right]$ در (٦) بین صفر و یک است و اثر آن نیز مانند حالت قبل کاهشی می‌باشد.

[۱] Nondifferential

حالت سوم این است که هر دو متغیر A وB در معرض طبقه‌بندی نادرست باشند. در این حالت زوج جانشین (A^* وB^*) تواماً توسط زوج (A وB) با احتمال طبقه‌بندی نادرست

$$P_r(A^* = j, B^* = k|A = j, B = k)$$

بیان می شود. طبقه‌بندی نادرست A وB مستقل گفته می‌شود اگر:

$$P_r(A^* = j^*, B^* = k^*|A = j, B = k) = P_r(A^* = j^*|A = j, B = k) \times P_r(B^* = k^*|A = j, B = k)$$

و همچنین طبقه‌بندی نادرست A وB غیر افتراقی گفته می‌شود اگر:

$$P_r(A^* = j^*|A = j, B = k) = P_r(A^* = j^*|A = j)$$
$$P_r(B^* = k^*|A = j, B = k) = P_r(B^* = k^*|B = k)$$

تحت شرایط استقلال و طبقه‌بندی نادرست غیر افتراقی در A وB گالن و همکاران، نشان داده‌اند که برآوردگر نااریب $\Pi_{A^*|B^*}(۱|۱) - \Pi_{A^*|B^*}(۲|۲)$ مجدداً تفاوتی کمتر از مقدار واقعی را نشان می‌دهد. از سوی دیگر اگر A وB یا هر دو تحت طبقه‌بندی نادرست افتراقی باشند گلدبرگ، نشان داده که اریبی حاصل لزوماً کاهشی نسبت به مقدار تفاوت واقعی نیست و هر دو جهت را ممکن است بگیرد. در رابطه با آزمون فرض در جداول دو طرفه نیز اگر میان A وB وابستگی وجود نداشته باشد با وجود احتمالات طبقه‌بندی نادرست غیر افتراقی، میان جانشینهای A^* و B^* وابستگی وجود نخواهد داشت. مارشال و همکاران، نشان داده‌اند که آزمون عدم وابستگی میان A وB براساس A^* و B^* دارای سطح معنی داری صحیح اما با توان کمتر خواهد بود.

به منظور روشنتر شدن موضوع با مثالی فرضی اثرات طبقه‌بندی نادرست در تحلیل اطلاعات جداولی دوطرفه وقتی که تشخیص بیماری با احتمالات طبقه‌بندی نادرست مواجه باشد را بررسی می‌کنیم: با در نظر گرفتن ساختار جدول (۱) داده‌های فرضی زیر را به عنوان داده‌های واقعی در نظر می‌گیریم:

اگر روش تشخیصی بکار گرفته شده برای تشخیص بیمار از غیر بیمار حساسیت ۰/۹ و ویژگی ۰/۸ داشته باشد آنگاه براساس احتمالات فوق الذکر امید ریاضی مقادیر مشاهده شده به صورت جدول ۳ خواهد بود:

	بیمار	غیر بیمار	مجموع
مواجهه	٦٠	٩٠	١٥٠
غیر مواجهه	٢٠	١٣٠	١٥٠
مجموع	٨٠	٢٢٠	٣٠٠

جدول ۱: طبقه‌بندی واقعی بیمار و غیر بیمار در دو گروه

	بیمار	غیر بیمار	مجموع
مواجهه	٧٢	٧٨	١٥٠
غیر مواجهه	۴۴	١٠٦	١٥٠
مجموع	١١٦	١٨۴	٣٠٠

جدول ۲: طبقه‌بندی بیمار و غیر بیمار بر اساس روشهای جاری (همراه با خطا)

که فراوانیهای مشاهده شده براساس حساسیت و ویژگی روشهای تشخیص بیماری بر اساس جدول ۴ بدست آمده است.

براساس جداول ۲ و ۳ برخی شاخصهای قابل محاسبه در جدول ۵ ارائه شده است .

براساس نتایج فوق تاثیر طبقه‌بندی نادرست در برآورد شاخصهاملاحظه می‌شود. بدیهی است خطای طبقه‌بندی نادرست متاثر از نسبت بیمار به غیر بیمار، حجم مشاهدات در هر گروه , خطای طبقه‌بندی در تشخیص بیماری و تعیین مواجهه و ... می‌باشد و اندازهٔ خطا بر اساس عوامل فوق‌الذکر متفاوت خواهد بود. به منظور نشان دادن این اصل که "هر مقدار خطای طبقه‌بندی بیشتر شود برآورد نسبت بخت به سمت مقدار تحت فرض صفر میل خواهد نمود" محاسبات فوق برای تعیین نسبت بخت در حالتهای مختلف حساسیت و ویژگی در جدول ٦ آورده شده است:

به روشنی ملاحظه می‌شود که نسبت بخت داده‌های واقعی ۴/۳ است و با کم شدن حساسیت و ویژگی روش تشخیصی مقدار نسبت بخت به سمت عدد ۱ که همان مقدار تحت فرض صفر است میل می‌کند.

۴ راههای تعدیل خطای طبقه‌بندی نادرست

به منظور اصلاح اریبی حاصل از طبقه‌بندی نادرست لازم است اطلاعاتی در مورد ساختار طبقه‌بندی نادرست داشته باشیم. در حالت کلی چون این ساختار نامشخص است باید

	بیمار	غیر بیمار
مواجهه	$TP+FP$	$TN+FN$
غیر مواجهه	$TP+FP$	$TN+FN$

جدول ۳:

شاخص	بر اساس طبقه‌بندی واقعی	بر اساس طبقه‌بندی نادرست
EER	۰/۴	۰/۴۸
CER	۰/۱۳	۰/۲۹
ARR	۰/۲۷	۰/۱۹
NNT	۳/۷	۵/۳
RR	۳/۱	۱/۶۵
RRR	۲/۱	۰/۶۵
OR	۴/۳	۲/۲

جدول ۴:

براساس داده‌های کمکی به برآورد احتمالات طبقه‌بندی نادرست بپردازیم. برای بدست آوردن داده‌های کمکی به برآورد احتمالات طبقه‌بندی نادرست بپردازیم. برای بدست آوردن داده‌های کمکی دو روش پیشنهادی عبارتند از:

۱— نمونه‌های معتبر

براساس ابزاری که به آن استاندارد طلائی [۱] گفته می‌شودامکان تعیین مقدار واقعی یا دقیقتر A بر اساس A^* وجود دارد. البته در برخی مطالعات ممکن است این روش بسیار گران باشد. احتمال خطا در استاندارد طلائی صفر نیست اما این مقدار خطا باید قابل اغماض باشد. مناسب است که نمونه معتبر زیر نمونه‌ای از داده‌های اولیه باشد تا نسبتهای گروه‌بندی A در نمونه معتبر برآوردگر نااریب نسبتهای متناظر در جامعه باشند. روشهای دیگری در ارتباط با انتخاب نمونه‌های معتبر برای شرایط متفاوت وجود دارد.

۲— اندازه‌گیریهای مکرر

در صورت عدم وجود استاندارد طلائی با روش اندازه‌گیریهای مکرر به برآورد پارامترهای طبقه‌بندی نادرست می‌پردازیم. تعداد تکرار بستگی به مدل خواهد داشت و این اندازه‌گیریها باید استقلال شرطی داشته باشند. عموماً سه بار اندازه‌گیری برای اکثر مدلها کفایت دارد. در صورت وجود اطلاعات لازم از ساختار طبقه‌بندی غلط با کمک روشهای زیر می‌توان خطای طبقه‌بندی نادرست را تعدیل نمود.

[۱] Gold Standard

		بیمار	غیربیمار
$Sen = ۰/۹$	مواجهه	۵۴	۹۶
$Spe = ۱$			
$OR = ۴/۱$	غیرمواجهه	۱۸	۱۳۲
$Sen = ۰/۸$	مواجهه	۶۶	۸۴
$Spe = ۰/۸$			
$OR = ۲$	غیرمواجهه	۴۲	۱۰۸
$Sen = ۰/۸$	مواجهه	۷۵	۷۵
$Spe = ۰/۷$			
$OR = ۱/۷$	غیرمواجهه	۵۵	۹۵
$Sen = ۰/۷$	مواجهه	۸۷	۶۳
$Spe = ۰/۵$			
$OR = ۱/۲$	غیرمواجهه	۷۹	۷۱
$Sen = ۰/۵$	مواجهه	۷۵	۷۵
$Spe = ۰/۵$			
$OR = ۱$	غیرمواجهه	۷۵	۷۵

جدول ۵:

الف— روشهای ماتریسی

با استفاده از روش برآورد بردار نسبتهای Π_A را برای متغیر با گروه‌بندی نادرست A بدست می‌آوریم. فرض کنیم که مجموعه داده اولیه به حجم n_p و یک مجموعه داده معتبر به حجم n_v در دسترس باشد. داده‌های معتبر که برآوردها را مشخص می‌سازند بصورت

$$\hat{\Theta}(A^*|A)$$

در نظر گرفته می‌شوند. برای ماتریس احتمالات طبقه‌بندی نادرست داریم

$$\Theta_{jk} = P_r(A^* = j|A = k)$$

و برآورد ماتریس Π_A بصورت

$$\hat{\Pi}_A^M = \left[\hat{\phi}(A^*|A)\right]^{-1} \hat{\Pi}_{A^*}$$

این برآوردگر در مطالعات مختلف کاربرد دارد. با روش دلتا نیز امکان برآورد ماتریس واریانس وجود دارد. این روش برای مسائل ساده تر مناسب است و از آن جمله در مواردی

که تعداد گروه‌بندی متغیرها یا واریانس کم باشد.این روش در مواردی که گروه‌بندی متغیرها زیاد باشد جداول پراکنده و برآوردهای غیر دقیق برای برخی پارامترها را به همراه خواهد داشت.

ب — روش مدل سازی

اگر A را یک مجموعه از متغیرهای با احتمال طبقه‌بندی نادرست و A^* نیز مجموعه متغیرهای جانشین و C مجموعه متغیرهای بدون طبقه‌بندی نادرست باشند و مجموعه متغیر L وابستگی داده‌ها را به گروهی مشخص نماید آنگاه توزیع توام (A,A^*, C, L) می تواند تحت دو مدل زیر مشخص شود:

۱) مدلی برای متغیرهای صحیح (A, C, L) که اثر متقابل بین L و (A, C) نشانگر تفاوت میان توزیع متغیرهای صحیح در میان نمونه‌ها می‌باشد.

۲) مدلی برای احتمالات طبقه‌بندی غلط بوسیله اثرات متقابل درون A^* و بین A^* با (A, C) مشخص می‌شود. مدل با توجه به متغیرهای صحیح (A, C) اشباع شده می‌باشد غالباً هر دو مدل لگ خطی سلسله مراتبی در نظر گرفته می‌شود اما مدل توام تولید شده توسط آنها در حالت کلی ممکن است لگ خطی نباشد.

ج — روشهای بیزی

با فرض اینکه A^* نشانگر وضعیت متغیر با طبقه‌بندی غلط و A نشانگر وضعیت واقعی آن متغیر باشد در حالتی که احتمالات طبقه‌بندی نادرست نامشخص باشند تابع درستنمایی قابل تشخیص نبوده و استنباطهای کلاسیک را غیر ممکن می‌سازد. و از سوی دیگر اگر حساسیت و ویژگی مشخص باشند آنگاه استنباط در مورد نسبت‌ها در جداول توافقی آسان و روشن می‌باشد. اما در حالت واقع گرایانه ما اطلاعات تقریبی از حساسیت و ویژگی روشها داریم و این اطلاعات را به عنوان بهترین حدسها برای سایر برآوردها به کار خواهیم برد. براساس روشهای بیزی عدم اطمینان در مورد حساسیت و ویژگی روشها را بایک توزیع پیشین نشان داده و آنرا با اطلاع پیشین ترکیب کرده و برآوردهای بهتری برای احتمالات طبقه‌بندی نادرست بدست خواهیم آورد.

۵ بحث و نتیجه‌گیری

طبقه‌بندی غلط موجب انحراف در برآورد کمیتهای موردنظر براساس مشاهدات می‌گردد. اریبی حاصل از طبقه‌بندی نادرست تابعی از حساسیت و ویژگی روشهای بکار گرفته شده برای طبقه‌بندی است. طبقه‌بندی غلط در برخی حالات موجب تضعیف وابستگی میان

متغیرها می‌گردد. در برخی موارد نیز ممکن است انحرافهای ناشی از خطای طبقه‌بندی غلط در هر جهتی واقع شود. لازم است با انتخاب داده‌های کمکی همچون داده‌های معتبر یا اندازه‌گیریهای مکرر برآوردی از ساختار احتمالات طبقه‌بندی نادرست بدست آوریم تا بتوانیم برآوردهای تعدیل شده‌ای بدست آوریم. در حالت طبقه‌بندی نادرست تلاش براین است که اریبی حاصله را تعدیل کنیم ودر جهت اصلاح اریبی پیشنهادهایی بدهیم در این موارد در صورت امکان از ملاک MSE به جای واریانس استفاده خواهیم نمود که‌معمولا اطلاع واقع بینانه‌تری در اختیار قرار می‌دهد. در این مقاله با یک مثال فرضی اثرات طبقه‌بندی نادرست در برآورد برخی شاخصهای قابل محاسبه در جداول توافقی ۲ × ۲ نشان داده شده است. براساس نتایج حاصله مشخص گردیده برآورد شاخصهایی چون نسبت بخت به مقدار زیاد تحت تاثیر اندازه خطای طبقه‌بندی نادرست بوده و باافزایش خطای طبقه‌بندی نادرست اندازه برآورد نسبت بخت به مقدار تحت فرض صفر میل می‌کند. نتایج مورد اشاره با نتایج ارائه شده در دیگر تحقیقها هماهنگی دارد.

مراجع

Armitage, Peter and Colton (1991) *Encyclopedia of Biostatistics*, John Wiley, 2615 - 2621.

Bross, I. (1954) *Misclassification in* ۲ × ۲ *tables*, Biometrics 10, 488 - 495.

Chen. T.T (1989) *A review of methods for misclassified categorical data in epidemiology, Statistics in medicine 8, 1095 - 1106.*

Dawson, Beth and Trapp, Robert (2001) *Basic and Clinical Biostatistics, MC Graw - Hill, 51 - 54*

Espeland, M. A. & Hui, S.L. (1987) *A general approach to analyzing epidemiologic data that contain misclassification errors, Biometrics 43,1001-1012*

Haitovsky, Y & Rapp, J. (1992) *Conditional resampling for misclassified multinomial data with applications to sampling inspection. Technometrics*

34, 473 - 483.

Hironori, F. and Shizu, I. (2000) *Inference about miscalssification probabilities from repeated binary response, Biometrics 56, 706 - 711.*

Liv, X. and liang, K. Y. (1991) *Adjustment for nondifferential Misclassificantion error in the generalized linear model, statistics in Medicine 10, 1197 - 1211*

Marshal, J.R.,Priore, R.,Graham, S. (1981) *On the distrotion of risk estimates in multiple exposure level cose - control studies, American Journal of Epidemiology 113, 464 - 473*

Tenenbein, A. (1972) *Adouble sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection, Technometrics 14, 187 - 202*

رکوردها و حجم نمونه

جعفر احمدی، مهدی دوست پرست

P۱۳۰۵۴

دانشکده علوم ریاضی دانشگاه فردوسی مشهد

چکیده: یکی از مسائل مهم در استنباط و تحلیل آماری برآورد پارامترهای مجهول جامعه می‌باشد. موارد متعددی وجود دارد که بدلایل نامعلومی تعدادی از مشاهدات ناپدید شده‌اند و پژوهشگر سعی در برآورد تعداد آنها دارد، بنابراین یکی از پارامترهای مجهول در این حالت حجم نمونه (n) می‌باشد. روشهای زیادی وجود دارد که می‌توانند جهت تخمین n بکار گرفته شوند و بکاربردن هر روش نیازمند وجود اطلاعاتی دربارهٔ شاخصهایی از جامعه است.

فرض کنید $X_1, X_2, \ldots$ دنباله‌ای از مشاهدات باشد، X_j را یک رکورد پائین (بالا) می‌گوئیم هرگاه از تمام مشاهدات ما قبل خود کوچکتر (بزرگتر) باشد. در این مقاله، با درنظر گرفتن $N^{(n)}$ بعنوان تعداد رکوردها در نمونه‌ای به حجم n، ابتدا به بررسی خواص توزیعی $N^{(n)}$ می‌پردازیم، سپس براساس $N^{(n)}$، برآوردگرهای نا اریب، درستنمایی ماکزیمم و گشتاوری برای n بدست می‌آوریم. درپایان، این برآوردگرها را مورد مقایسه قرار می‌دهیم.

واژه‌های کلیدی: رکوردهای پائین (بالا)، تقعّر (تحدّب)، برآوردگر نااریب، برآوردگر درستنمایی ماکزیمم، برآوردگر گشتاوری، اعداد استرلینگ نوع اول.

۱ مقدمه

نظریه رکوردها شاخه ای نسبتاً جدید می‌باشد و رشد اصلی خود را در چند دههٔ اخیر نموده است. این نظریه علاوه بر ویژگیهای مهم تئوری دارای کاربردهای خاص عملی است. تغییرات جوی، بعضی از مسائل ترافیک، پیشامدهای طبیعی مانند باد و مسائل مربوط به تعیین مقاومت مصالح و محاسبه احتمال کارافت و ... از جمله کاربردهای مهم آن است.

وقتی ما با مشاهدات متوالی سرکار داریم، مشاهدات تصادفی مربوط به آنها مورد توجه قرار می‌گیرند. مثلاً ممکن است طی سالهای متمادی میزان بارندگی روزانه یا هفتگی را در منطقه‌ای ثبت کرده و بخواهیم در مورد تعداد روزها یا هفته هایی که ممکن است میزان بارندگی بیش از گذشته باشد را مطالعه کنیم. یا طراح یک سازه بخواهد بداند که

در طول عمر مفید سازه مورد طرح وی احتمال وقوع زلزله‌ای بزرگتر یا مخربتر از آنچه که در گذشته اتفاق افتاده، چقدر است؟ بعبارت دیگر اگر سازه‌ای براساس بزرگترین زلزله درتاریخ منطقه‌ای مورد نظر محاسبه و طراحی شود، احتمال تنزل کارائی آن در اثر بار زلزله در ۲۰ یا ۵۰ سال آینده چقدر است؟ یا فرض کنید دنباله از تولیداتی را که ممکن است در اثر فشار متلاشی شوند در نظر بگیریم و مایل به تعیین مینیمم فشار متلاشی شدن برای این دنباله از تولیدات باشیم، ما اولین محصول را آزمایش می‌کنیم تا اینکه با حداقل فشار متلاشی شود، X_1 را بعنوان فشار گسیختگی اولین محصول ثبت می‌کنیم، سپس محصولات بعدی را درنظر می‌گیریم و X_m را بعنوان فشار گسیختگی محصول m ام نامیده و آنرا در صورتی ثبت می‌کنیم که

$$X_m < min(X_1, ..., X_{m-1})$$

بنابراین در صورتی فشار گسیختگی را ثبت خواهیم کرد که کمترین مقدار در بین مقادیر ثبت شده قبلی را داشته باشد.

بعنوان مثال دیگر X_{ij} را بالاترین سطح آب رودخانه‌ای در روز j ام در مکان i ام باشد. فرض کنید علاقمند به مطالعه ماکزیمم موضعی X_{ij} در هر مکان باشیم. بنابراین مشاهده می‌کنیم گاهی اوقات ماکزیمم موضعی و گاهی اوقات مینیمم آن برای ما اهمیت دارد.

مزیت این نظریه را می‌توان در عدم نیاز به مشاهدات غیر رکوردی، سادگی و غیر تقریبی بودن برخی نتایج آن بویژه آن دسته که مستقل از توزیع مشاهدات هستند، دانست.

تحقیق براساس رکوردها تاکنون مورد توجه متخصصین بسیاری واقع شده است. شاید بتوان گفت چاندلر بطور برجسته‌ای مطالعه مقادیر رکورد را شروع کرد و رزنیک (۱۹۷۳) و شورک (۱۹۷۳)، نظریه مجانبی رکوردها را تکمیل کردند. بعدها، نقطه نظرات جالبی برای نسبت دادن مقادیر رکورد با فرایندهای فرین معین معرفی شدند. ند گلیک (۱۹۷۸) با استفاده از عنوان فریبنده " شکستن رکوردها و شکستن مرزها " تحقیقات بیست و پنج ساله اول را جمع آوری کرد. استنباط آماری براساس رکوردها در سالهای اخیر نیز مورد توجه قرار گرفته است می‌توان به کارلین و گیلفاند (۱۹۹۳)، فیورویرگر و هال (۱۹۹۸)، گالاتی و پادجت (۱۹۹۴)، احمدی و ارقامی (۱۳۷۵ش) (۲۰۰۱) (۲۰۰۲الف) (۲۰۰۲ب) و احمدی (۱۳۷۳ش) (۲۰۰۰) مراجعه کرد. احمدی و ارقامی (۲۰۰۱) با ارائه نظریه اطلاع فیشر براساس رکوردها نشان دادند که در بعضی موارد، رکوردها بسیار مناسب تر از مشاهدات و یا آماره های ترتیبی عمل می‌کنند. مطالب جامع دربارهٔ رکوردها را می‌توان در کتابی با عنوان "رکوردها" نوشتهٔ آرنولد، بالاکریشنان و ناگاراجا (۱۹۹۸) یافت.

در این مقاله، بخش ۲ شامل تعاریف پایه‌ای ازداده‌های رکوردی می‌باشد. دربخش ۳

دربارهٔ چگالی و خصوصیات $N^{(n)}$ بحث خواهیم کرد و باتوجه به قضایای ارائه شده در بارهٔ آمارهٔ تعداد رکوردها، در بخش ۴ برآوردگرهای نقطه‌ای نااریب، درستنمایی ماکزیمم و گشتاوری برای حجم نمونه معرفی می‌کنیم. بخش (۵) شامل مقایسه این برآوردگرها می‌باشد.

۲ آماره‌های رکوردی

در این بخش چند تعریف پایه‌ای که در این مقاله از آنها استفاده می‌کنیم، بیان می‌کنیم.

۱.۲ رکورد

فرض کنید $X_۱, X_۲, \ldots$ دنباله‌ای از متغیرهای تصادفی هم توزیع، مستقل و از توزیع مشترک F باشند. مشاهده X_j را یک رکورد پائین (بالا) می‌گوئیم هرگاه از همهٔ مشاهدات ما قبل خود کوچکتر (بزرگتر) باشد.

درادامه، خود را به رکوردهای پائین محدود می‌کنیم. نتایج حاصل را می‌توان برای رکوردهای بالا با اندکی تغییر مورد استفاده قرار داد.

۲.۲ زمان رکورد

طبق تعریف ۱.۲، زمانی که در آن رکورد رخ می‌دهد یک متغیر تصادفی می‌باشد که بصورت زیر تعریف می‌شود

$$L(۱) = ۱, L(k) = min[j : j > L(k-۱), X_j < X_{L(k-۱)}]$$

که $L(k)$ زمان رخ دادن k امین رکورد پائین می‌باشد .

۳.۲ تعداد رکوردها

قرار دهید $N^{(n)} = \sum_{i=۱}^{n} I_i$ آنگاه $N^{(n)}$ متغیر تصادفی تعداد رکوردهای پائین در بین $X_n, .., X_۱$ می‌باشد که متغیرهای نشانگر I_i را بصورت زیرتعریف می‌کنیم

$$I_i = \begin{cases} ۱ & X_i = min(X_۱, ..., X_i) \\ \\ ۰ & \text{درغیر اینصورت} \end{cases}$$

۴.۲ مثال

فرض کنید دنباله‌ای از مشاهدات با $n = ۱۰$ با مقادیر زیر داریم

$$۲۰, ۲۲, ۱۸, ۲۵, ۳۰, ۱۲, ۴۰, ۳۲, ۱۰, ۱۱$$

دراین صورت طبق تعریف ۱.۲، رکوردهای پائین عبارتند از:

$$۲۰, ۱۸, ۱۲, ۱۰$$

وطبق تعریف ۲.۲، زمان رخدادن رکوردهای پائین عبارتند از:

$$L(۱) = ۱, L(۲) = ۳, L(۳) = ٦, L(۴) = ۴$$

وطبق تعریف ۳.۲، تعداد رکوردهای پائین

$$N^{(۱۰)} = ۴$$

می‌باشد.
لازم به ذکر است که دنباله رکوردهای بالای نمونه فوق عبارتند از

$$۲۰, ۲۲, ۲۵, ۳۰, ۴۰$$

۳ خواص توزیعی تعداد رکوردها

دراین بخش با ارائه قضایا و لمهایی، رفتار $N^{(n)}$ را بررسی می‌کنیم وبا ایده‌هایی که از این بخش می‌گیریم، زمینه لازم برای رسیدن به هدف خود یعنی برآورد $N^{(n)}$ فراهم می‌کنیم.
لم ۱ : متغیرهای تصادفی $I_i (i = ۱, .., n)$ دارای توزیع $B(۱, \frac{۱}{i})$ هستند و از هم مستقل‌اند.
برهان : برای هر $i < j$

$$P(I_i = ۱, I_j = ۱) = P(Y_i = min(Y_۱, ..., Y_i), Y_j = min(Y_۱, ..., Y_j))$$
$$= P(Y_i = min(Y_۱, .., Y_i) > Y_j = min(Y_{i+۱}, .., Y_j))$$
$$= P(Y_i = min(Y_۱, ..., Y_i)) \times P(min(Y_۱, ..., Y_i) > min(Y_{i+۱}, ..., Y_j)) \times P(Y_j = min(Y_{i+۱}, ..., Y_j)) = \frac{۱}{i}\frac{j-i}{j}\frac{۱}{j-i} = \frac{۱}{ij} = P(I_i = ۱)P(I_j = ۱)$$

وبه همین ترتیب سایر موارد بطور مشابه ثابت می‌شود.
لم ۲ : بدیهی است که

$E(N^{(n)}) = E(\sum_{i=۱}^{n} I_i) = \sum_{i=۱}^{n} \frac{۱}{n}$
$Var(N^{(n)}) = \sum_{i=۱}^{n} \frac{۱}{i}(۱ - \frac{۱}{i}) = \sum_{i=۱}^{n} \frac{۱}{i} - \sum_{i=۱}^{n} \frac{۱}{i^۲}$

نتیجه۱: از روابط فوق می‌توان یک کران بالا برای احتمال وقوع حداقل تعدادی از رکوردها راباتوجه به نامساوی یک طرفه چبیشف ارائه داد :

$$P(N^{(n)} \geq r) \leq \frac{Var(N^{(n)})}{Var(N^{(n)}) + [r - E(N^{(n)})]^۲}$$

بعنوان مثال

$$P(N^{(۱۰۰۰)} \geq ۱۸) \leq ۰.۰۵$$

قضیه ۱ : تابع جرم احتمال $N^{(n)}$ عبارت است از :

$$g(n;k) = P(N^{(n)} = k) =$$
$$\sum_{i_۲=۲}^{n-k+۲} \sum_{i_۳=i_۲+۱}^{n-k+۳} \cdots \sum_{i_k=i_{k-۱}+۱}^{n} \frac{۱}{(i_۲ - ۱)(i_۳ - ۲)...(i_k - ۱)n}$$

برهان : با توجه به تعریف ، اولین مشاهده را رکورد قرار داده‌ایم پس

$P(L(۲) = n) = P($(رکورد پائین دوم درزمان nام رخ دهد$)$
$= P(X_۱ = min(X_۱, ..., X_{n-۱}), X_n < X_۱)$
$= \frac{(n-۲)!*۱*۱}{n!} = \frac{۱}{n(n-۱)}$

امّابرای اعداد صحیح $۱ < n_۱ < n_۲ < ... < n_k$ داریم

$$P(L(۱) = ۱, L(۲) = n_۲, L(۳) = n_۳, ..., L(k) = n_k)$$

$$P(I_۲ = ۰, ..., I_{n_۲-۱} = ۰, I_۲ = ۱, I_{n_۲+۱} = ۰, ..., I_{n_k} = ۱)$$

(از لم ۱)

$$= \frac{۱}{۲}\frac{۲}{۳}\frac{۳}{۴}...\frac{n_۲ - ۲}{n_۲ - ۱}\frac{۱}{n_۲}\frac{n_۲}{n_۲ + ۱}...\frac{۱}{n_k} = [(n_۲ - ۱)(n_۳ - ۱)...(n_k - ۱)n_k]^{-۱}$$

و نتیجه حاصل می‌شود.

در بعضی از مراجع، تابع جرم احتمال $N^{(n)}$ که با استفاده از اعداد استرلینگ نوع اول بیان می‌شود بصورت زیر بیان می‌کنند:

$$g(n;k) = P(N^{(n)}) = \frac{|S_n^{k-1}|}{(k-1)!} \approx \frac{[ln(n)]^{k-1}}{n(k-1)!}$$

تقریب برای n به اندازه کافی بزرگ برقرار است.برای اثبات می‌توان به ند گلیک (۱۹۷۸) رجوع کرد.
مورنو و دیگران (۱۹۹۶)، تابع مولد احتمال $N^{(n)}$، بصورت زیر ارائه شده است :

$$E(S^{N^{(n)}}) = P_n(s) = \frac{\Gamma(n+s)}{\Gamma(n+1)\Gamma(s)} \quad \forall s \in R$$

لم ۳ : $g(1,1) = 1$ و برای $k = 1,..,n+1$

$$g(n+1;k) = \frac{n}{n+1} g(n;k) + \frac{1}{n+1} g(n;k)$$

برهان :

$$g(n+1;k) = P(N^{(n+1)} = k)$$
$$= \sum_{i=1}^{k} P(N^{(n+1)} = k | N^{(n)} = i) P(N^{(n)} = i)$$
$$= P(N^{(n+1)} = k | N^{(n)} = k) \times P(N^{(n)}) = k) +$$
$$P(N^{(n+1)} = k | N^{(n)} = k-1) \times P(N^{(n)} = k-1) + 0 + 0 + ...$$
$$= P(I_{n+1} = 0 | N^{(n)} = k) \times g(n;k) + P(I_{n+1} = 1 | N^{(n)} = k-1) \times g(n;k-1)$$
$$= \frac{n}{n+1} g(n;k) + \frac{1}{n+1} g(n;k-1)$$

لم ۴ : تابع جرم احتمال $N^{(n)}$، بازای n ثابت، تابعی مقعراست .
برهان : $\triangle(n;k)$ و $\triangle^2(n;k)$ را به صورت زیر تعریف می‌کنیم:

$$\triangle(n;k) = g(n;k+1) - g(n;k)$$
$$\triangle^2(n;k) = \triangle(n;k+1) - \triangle(n;k)$$

کافی است ثابت کنیم :

$$\triangle^2(n;k) \leq 0 \qquad \forall k$$

برای $n = 1,2,3$ بارسم شکل دیده می‌شود که $g(n;k)$ تابعی مقعر است . فرض کنیم برای n برقرار باشد . از لم ۳ داریم :

$$\triangle^2(n+1;k) = \frac{1}{n+1}[n\triangle^2(n;k) + \triangle^2(n;k-1)]$$

باتوجه به فرض استقراً $\triangle^2(n;k) \leq 0$ و $\triangle^2(n,k-1) \leq 0$ می‌باشد،لذا $\triangle^2(n+1;k) \leq 0$.

لم ۵ : اگر $Mo(N^{(n)})$، مد $N^{(n)}$ براساس $g(n;k)$ باشد.

الف)$Mo(N^{(n)}) \leq Mo(N^{(n+1)} \leq Mo(N^{(n)} + 1$

ب) برای $n > 2$، $Mo(N^{(n+1)}) = Mo(N^{(n)}) + 1$ اگر و فقط اگر

$$n\triangle(n; Mo(N^{(n)})) + \triangle(n; Mo(N^{(n)}) - 1) \geq 0$$

برهان :

الف) ازلم ۳ نتیجه می‌شود که

$$\triangle(n+1;k) = \frac{1}{n+1}[n\triangle(n;k) + \triangle(n;k-1)]$$

و باتوجه به تقعّر $g(n;k)$

اگر$k \leq Mo(N^{(n)}) - 1$ آنگاه $\triangle(n;k) > 0$ و $\triangle(n;k-1) > 0$ لذا

$$\triangle(n+1;k) > 0 \qquad (1)$$

اگر$k \geq Mo(N^{(n)}) + 1$ آنگاه $\triangle(n;k) < 0$ و $\triangle(n;k-1) < 0$ لذا

$$\triangle(n+1;k) < 0 \qquad (2)$$

فرض کنیم در (۱)، $k = Mo(N^{(n)}) - 1$ لذا

$$g(n+1; Mo(N^{(n)})) > g(n+1; Mo(N^{(n)}) - 1) > ... > g(n+1; 1) \qquad (3)$$

وهمچنین در (۲)، $k = Mo(N^{(n)}) + 1$ لذا

$$g(n+1; Mo(N^{(n)}) + 1) > g(n+1; Mo(N^{(n)}) + 2) > ... > g(n+1; n+1) \quad (4)$$

پس $Mo(N^{(n)} \leq Mo(N^{(n+1)}) \leq Mo(N^{(n)}) + 1$

ب) اگر $\triangle(n+1; Mo(N^{(n)}) > 0$ آنگاه

$$g(n+1; Mo(^{(n)}) + 1) > g(n+1; Mo(N^{(n)}))$$

وباتوجه به رابطه‌های (۳) و (۴)

$$g(n;k) < g(n+1; Mo(N^{(n)}) + 1) \qquad \forall k \neq Mo(N^{(n)}) + 1$$

پس $Mo(N^{(n+1)}) = Mo(N^{(n)}) + 1$. برعکس اگر

$$Mo(N^{(n+1)}) = Mo(N^{(n)}) + 1$$

آنگاه با فرض $k = Mo(N^{(n)})$

$$\triangle(n+1; Mo(N^{(n)}) > 0$$

لم ٦ : بازای k ثابت، $g(n;k)$ برای $n \geq k$، دارای خصوصیات زیر است :
الف)برای $k \geq ۳$، $g(k;k) < g(k+۱;k)$
ب) $\lim_{n\to\infty} g(n;k) = ۰$
ج) $g(n;k)$ تابع تک مدی است یعنی
اگر $g(n+۱;k) > g(n;k)$ آنگاه $g(n;k) > g(n-۱;k)$
اگر $g(n;k) < g(n-۱;k)$ آنگاه $g(n+۱;k) < g(n;k)$
برهان : الف) ازتابع جرم احتمال $g(n;k)$ به آسانی بدست می آید.
ب)برای n های بزرگ درپیفر(۱۹۹۱) ثابت کرد که

$$g(n;k) = e^{-\lambda_n} \frac{\lambda_n^{k-۱}}{(k-۱)!} + O(\frac{۱}{Ln(n)})$$

که $\lambda_n = Ln(n) + \gamma - ۱$ و γ ثابت اویلر است و (ب) بدست می آید.
ج) از لم ۳ داریم:

$$g(n+۱;k) - g(n;k) = \frac{g(n;k-۱)-g(n;k)}{n+۱} = \frac{-\triangle g(n;k-۱)}{n+۱}$$

اگر $g(n+۱;k) - g(n;k) > ۰$ آنگاه $\triangle(n;k-۱) < ۰$ وازآنجا که برای n ثابت $g(n;k)$ مقعر است لذا$Mo(N^{(n)}) \leq k-۱$ وازلم ۵—الف $Mo(N^{(n-۱)}) \leq k-۱$ و مجدداً ازتقعر $g(n-۱;k)$ داریم $\triangle(n-۱;k-۱) < ۰$ لذا

$$g(n;k) - g(n-۱;k) = \frac{-\triangle(n-۱;k)}{n} > ۰ \qquad (۵)$$

حالت دوم بطورمشابه ثابت می‌شود.

۴ برآورد

در این بخش، برآوردگرهای نا اریب، درستنمایی ماکزیمم و گشتاوری با توجه به قضایا و لمهای بخش ۳ بیان می‌کنیم.
قضیه ۲ : $T_۱ = ۲^{N^{(n)}} - ۱$، یک برآوردگر نااریب n با واریانس

$$Var(T_۱) = \frac{(n+۳)(n+۲)(n+۱)}{٦} - (n+۱)^۲$$

برهان : با توجه به $P_n(s)$ تابع مولد احتمال $N^{(n)}$:

$$E(T_۱) = E(۲^{N^{(n)}}) - ۱ = P_n(۲) - ۱ = n$$
$$Var(T_۱) = Var(۲^{N^{N^{(n)}}}) = E(۲^{۲N^{(n)}}) - [E(۲^{N^{(n)}})]^۲ =$$
$$P_n(۴) - (P_n(۲))^۲ = \frac{\Gamma(n+۴)}{\Gamma(n+۱)\Gamma(۳)} - (n+۱)^۲$$

اگر $k_۰$ تعداد رکوردهای پائین مشاهده باشد آنگاه $\hat{n} = T_۲$ (برآوردگر درستنمائی ماکزیمم n) آن n ای است که $g(n;k_۰)$ را ماکزیمم کند یعنی

$$g(\hat{n};k_۰) = \max_n g(n;k_۰)$$

لذا می‌توان برآوردگر درستنمایی n را بصورت زیر بیان کرد.

قضیه ۳ : فرض کنید $k_۰$ تعداد رکوردهای پائین مشاهده شده در نمونه‌ای به حجم n (مجهول) باشد. در این صورت برآوردگر درستنمایی n عبارت است از:

$$\hat{n} = min\{n|Mo(N_{(n)}) = k_۰\}$$

برهان : اگر $\hat{n}$، $g(n;k_۰)$ را ماکزیمم کند، دراین صورت $g(\hat{n}-۱;k_۰) \leq g(\hat{n};k_۰)$ و $g(\hat{n}+۱,k_۰) \leq g(\hat{n};k_۰)$ از (۵)، $\triangle(\hat{n}-۱;k_۰-۱) < ۰$ و $\triangle(\hat{n};k_۰-۱) > ۰$ و از تقعرتابع جرم احتمال $N^{(n)}$:

$$Mo(N^{(\hat{n}-۱)}) \leq k_۰ - ۱ \quad , \quad Mo(N^{(\hat{n})}) \geq k_۰ \qquad (۶)$$

پس

$$Mo(N^{(\hat{n}-۱)}) \leq k_۰ - ۱ < k_۰ \leq Mo(N^{(\hat{n})}) \quad \Rightarrow \quad Mo(N^{(\hat{n}-۱)}) < Mo(N^{(\hat{n})})$$

امّا از لم ۵

$$Mo(N^{(\hat{n}-۱)}) \leq Mo(N^{(\hat{n})}) \leq Mo(N^{(\hat{n}-۱)}) + ۱ \Rightarrow$$
$$Mo(N^{(\hat{n})}) = Mo(N^{(\hat{n}-۱)}) + ۱$$

از (۶) داریم :

$$Mo(N^{(\hat{n})}) - ۱ \leq k_۰ - ۱ \quad , \quad Mo(N^{(\hat{n})}) \geq k_۰ \quad \Rightarrow$$

$$Mo(N^{(\hat{n})}) = k_۰ \quad , \quad Mo(N^{(\hat{n}-۱)}) = k_۰ - ۱$$

یعنی $\hat{n}$ جایی قراردارد که با یک واحد کاستن ازآن، یک واحد از مد $g(n;k_۰)$ کاسته می‌شود.یعنی

$$\hat{n} = min\{n|Mo(N^{(n)}) = k_۰\}$$

و اثبات کامل می‌شود.
می‌دانیم که $N^{(n)} = \sum_{k=1}^{n} I_k$. پس $E(N^{(n)}) = \sum_{k=1}^{n} \frac{1}{k} = ln(n) + \gamma$ براحتی می‌توان یک برآورد گشتاوری برای n بصورت زیر ارائه داد:

$$T_3 = exp(N^{(n)} - \gamma)$$

باتوجه به تابع مولّد احتمال $N^{(n)}$:

$$E(T_3^k) = E(exp(N^{(n)} - \gamma))^k = P_n(exp(k))exp(-k\gamma)$$

چون برآوردگر n باید مقادیر صحیح بگیرد لذا T_3^* را نزدیکترین عدد صحیح به $exp(N^{(n)} - \gamma)$ تعریف می‌کنیم.
ارهارد کرامر ثابت کرد هنگامی که $N^{(n)} \to \infty$، بااحتمال یک

$$\frac{T_2(N^{(\hat{n})})}{exp(N^{(n)} - \gamma)} \to 1 \qquad (7)$$

ممکن است پیشنهاد شود برای n،فاصلهٔ $[T_2(k), T_4(k)]$ که

$$T_4 = max\{m \geq k | Mo(N^{(m)}) = k\}$$

مناسب بنظر برسد. ولی طول این فاصله بطورنمایی افزایش می‌یابد زیرا

$$T_4 = T_2(k+1) - 1$$

و طول فاصله برابر

$$\begin{aligned} T_4(k) - T_2(k+1) &= T_2(k+1) - T_2(k) - 1 \\ &= exp(k)exp(-\gamma)[e-1] \end{aligned}$$

است.

۵ مقایسه

مورنو و دیگران (۱۹۹۶) پیش بینی کردند که برای $N^{(n)} \geq 3$

$$T_1 < T_2 < T_3{}^*$$

اما مدتی بعد، ارهارد کرامر همین رابطه را با استفاده از اعداد استرلینگ نوع اول ثابت کرد. مناسب بنظر می‌رسد که برای چند مقدار از $N^{(n)}$ برآوردهای مختلف ارائه شود.

$N^{(n)}$	$T_۱$	$T_۲$	$T_۳{}^*$
۱	۱	۱	۲
۲	۳	۲	۴
۳	۷	۸	۱۱
۴	۱۵	۲۵	۳۱
۵	۳۱	۷۳	۸۳
۶	۶۳	۲۰۴	۲۲۷
۷	۱۲۷	۵۶۵	۶۱۶
۸	۲۵۵	۱۵۵۶	۱۶۷۴
۹	۵۱۱	۴۲۷۲	۴۵۵۰
۱۰	۱۰۲۳	۱۱۶۹۸	۱۲۳۶۷

جدول ۱: مقایسه برآوردگرها

مراجع

۱. احمدی، ج (۱۳۷۳) «رکوردها در توزیعهای پیوسته» رساله کارشناسی ارشد (آمار ریاضی)، دانشکده علوم ۲، دانشگاه فردوسی مشهد.

۲. احمدی و ارقامی (۱۳۷۵) «رکوردها در توزیع بتای نوع اول» اندیشه آماری، سال اول، شماره ۲، ۹—۱۳.

[3] Ahmadi, J. and Arghami, N. A.(2002a)*Nonparametric confidence intervals based record values datas*, accepted for publication in *Statistical papers*.

[4] Ahmadi, J. and Arghami, N. A.(2002b)*Comparing the Fisher information In record values and IID observations*, to apaer in *Statistics*.

[5] Ahmadi, J. and Arghami, N. A.(2001)*On the Fisher information in record values*, *Metrika* 53, 3, 195-206.

[6] Ahmadi, J (2000) Record values - Theory and Aplications, Ph.D thsies, Ferdowsi University of Mashhad, Iran.

[7] Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1992) *A first course in order statistics*, John Wiley, New York.

[8] Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1998) *Records*, John Wiley, New York.

[9] Carlin, P. B. and Gelfand, A. E. (1993)*Parametric likelihood inference for record breaking problems*, *Biometrika*, 80, 3, 507-515. ¿

[10] Erhard Gramer.*Asymptotic properties of estimators of the sample size in a record model.*

[11] Feuerverger, A. and Hall, P. (1998) On statistical inference based on record values, *Extermes*, 12, 169-190.

[12] Glick,N.(1978).*Breaking records and breaking boards*,Am.math. Monthely 85,2-26.

[13] Gulati, S. and Padgett, W. J. (1994) Smooth nonparametric estimation of the distribution and density functions from record-breaking data, *Comm. Stat.-Theory Methods*, 23(5), 1259-1274.

[14] Moreno-Rebollo, J.L, Barranco-Chamorro,I.,L. pez-Bl. squez, F.and G..mez-G. mez,T. (1996). *On the estimation of the unknown sample size from the rocords.* statist. Probab. letters 31, 7-12.

[15] Resnick,S.I. (1973)*Records values and maxima*, Ann. Probab.1,650-662

[16] Shorrock, R. W. (1973) *Record values and inter-record times*, J. Appl. Probab. 10, 543-555.

اطلاع در توزیع بر نوع ۱۲ (BURR XII)

جعفر احمدی، مصطفی رزمخواه

P۱۴۲۸۳

گروه آمار، دانشگاه فردوسی مشهد

چکیده: خانواده توزیعهای بر در سال ۱۹۴۲ توسط بر معرفی شده است. یکی از اعضای این خانواده توزیع بر نوع ۱۲ می‌باشد که دارای کاربردهای وسیعی بویژه در کنترل کیفیت، تحلیل داده‌های نقص فنی، کارآیی مسکن‌ها در آزمایشات پزشکی و مدلهای قابلیت اطمینان در سیستمهای مهندسی می‌باشد.

این توزیع در مقالات بصورت دوپارامتری معرفی شده است. برای بدست آوردن ماتریس اطلاع فیشر به محاسبه انتگرالهایی مواجه می‌شویم که به سادگی قابل حل نیستند. در این مقاله ابتدا امید ریاضی تعدادی توابع پایه‌ای از این توزیع را بدست می‌آوریم و آنگاه نشان خواهیم داد که تمام امید ریاضی‌های مربوط به ماتریس اطلاع فیشر را می‌توان از نتایج بدست آمده محاسبه نمود. بعلاوه با روش پارامترسازی، یک پارامتر مقیاس برای این توزیع معرفی می‌کنیم و ماتریس اطلاع فیشر را برای حالت سه‌پارامتری بدست می آوریم.

واژه‌های کلیدی: تابع درستنمایی، برآوردگر درستنمایی ماکزیمم، ماتریس اطلاع فیشر.

۱ مقدمه

منشأ کار روی این خانواده از توزیعها مقاله‌ای بود که در سال ۱۹۴۲ توسط بر(BURR) ارائه شد که در آن وی قصد دارد تا توزیع‌هایی را که می‌توانند فرم‌های بسیار متنوعی را بگیرند و انعطاف پذیری زیادی دارند ایجاد کند.

خانواده بر توسط جواب‌هایی برای یک معادله تفاضلی در تابع توزیع تجمعی معرفی شده است. یک جواب ویژه که باعث بوجود آمدن توزیع B۱۲ شده، در عمل بسیار پر کاربرد واقع گشته است؛ می‌توان به برخی از این کاربردها در بخش‌هایی از کنترل کیفیت، طول عمر یا مدل بندی زمان نقص، مدل بندی توزیع درآمد،bio-assay و آزمون فرض اشاره نمود.

بر(۱۹۴۲) تعدادی از فرمهای توابع توزیع تجمعی را توصیف کرد که در عمل معلوم

شده است که برای برازش داده‌ها مفید می‌باشند. هدف از استفاده از این توابع توزیع تجمعی، آسان کردن تحلیل‌های ریاضی تا رسیدن به یک برازش منطقی به داده‌ها با استفاده از روش گشتاورها می‌باشد. این سیستم توزیع‌ها با در نظر گرفتن توابع توزیعی که برای معادله تفاضلی زیر مناسب هستند معرفی می‌شوند:

$$dF = F(۱ - F)g(x)dx$$

که $۰ \leq F \leq ۱$ و $g(x)$ یک تابع مناسب و نامنفی برای هر محدوده x است. جواب این معادله تفاضلی به ازای $g(x)$ داده شده عبارتست از:

$$F(x) = \frac{۱}{۱ + e^{-G(x)}}$$

که در آن

$$G(x) = \int_{-\infty}^{x} g(u)du$$

بر، دوازده جواب برای این معادله تفاضلی (با در نظر گرفتن انتخاب $g(x)$) ارائه کرده است که دوازدهمین آنها توزیع بر نوع ۱۲ نام گرفته است.

اگر X بصورت توزیع بر نوع دوازده توزیع شده باشد، از نماد B۱۲ استفاده می‌کنیم. برای این توزیع، تابع توزیع احتمال بصورت زیر معرفی می‌گردد

$$F(x) = ۱ - (۱ + x^c)^{-k}, \qquad x > ۰$$

و تابع چگالی احتمال عبارتست از

$$f(x) = kcx^{c-۱}(۱ + x^c)^{-(k+۱)}, \qquad x > ۰ \qquad (۱)$$

این تابع چگالی تک مدی با مد $x = (\frac{c-۱}{kc+۱})^{\frac{۱}{c}}$ برای $c > ۱$ می‌باشد. و اگر $c \leq ۱$ این چگالی فرم L به خود می‌گیرد.

در بسیاری از پژوهش‌های آماری، متغیرهایی که مدل می‌شوند توابعی از تعدادی متغیرهای دیگر هستند. بنابراین بیان تعدادی از روابط بین توزیع B۱۲ و تعدادی دیگر از توزیع‌های معروف برای تسهیل چنین مدل‌بندی‌هایی مفید است. اگر $X \sim B۱۲$، آنگاه:

۱— $U = (۱ + X^c)^{-۱}$ دارای توزیع بتای نوع اول با پارامترهای k و ۱ می‌باشد.

۲— $V = X^c$ دارای توزیع بتای نوع دوم (پیرسن نوع ٦) با پارامترهای k و ۱ می‌باشد.

۳– $W = kX^c$ دارای توزیع فیشر با درجات آزادی (۲,۲k) می‌باشد. شایان توجه است که Timr.L.Fry.(1993)خواص خانواده توزیع بر را در حالتهای یک متغیری و دو متغیری بررسی نموده است، همچنین احمدی (۱۳۷۳) بر اساس رکوردها این توزیع را توصیف کرد.

هدف اصلی این مقاله، محاسبه ماتریس اطلاع فیشر می‌باشد. می‌دانیم که اطلاع فیشر در نامساوی کرامر ـ رائو برای بدست آوردن کران پایین واریانس برآوردگرهای نااریب و همچنین در واریانس توزیع حدی برآورردگرهای درستنمایی ماکزیمم ظاهر می‌شود.

اخیراً بررسی اطلاع فیشر مورد توجه متخصصین زیادی قرار گرفته است، از جمله می‌توان به Ahmadi j and Arghami NR (2001)، Park S (1996) و Zheng G and Gastwirth JL (2000) اشاره نمود.

در بخش دوم این مقاله ابتدا به تعریف اطلاع فیشر می‌پردازیم و سپس لم‌هایی را اثبات می‌کنیم که در محاسبه ماتریس اطلاع فیشر مورد استفاده قرار می‌گیرند. در بخش سوم لگاریتم درستنمایی و مشتقات آن را محاسبه می‌کنیم. در بخش چهارم امید ریاضی توابعی پایه‌ای از این توزیع را که در تعیین ماتریس اطلاع فیشر مفید واقع می‌گردند محاسبه می‌کنیم. در بخش پنجم ماتریس اطلاع فیشر را معرفی می‌کنیم. و بالاخره در بخش ششم با پارامترسازی یک پارامتر مقیاس برای این توزیع تعریف کرده و ماتریس اطلاع فیشر متناظر را برای حالت سه‌پارامتری بدست می‌آوریم.

۲ نتایج مقدماتی

تعریف اطلاع فیشر: مقدار اطلاع فیشر درباره پارامتر θ در متغیر تصادفی X، که دارای چگالی $f(x;\theta), \theta \in A$ می‌باشد، برابر است با:

$$I_X(\theta) = E([\frac{\partial}{\partial\theta}\log f(X;\theta)]^2) \geq 0$$

امید ریاضی فوق را نسبت به چگالی $f(x;\theta)$ حساب می‌کنیم و فرض می‌کنیم که وجود داشته باشد.

ماتریس اطلاع فیشر: اطلاع فیشر را می‌توان برای θ موقعی که چندبعدی باشد تعمیم داد و یک ماتریس به نام ماتریس اطلاع فیشر بدست آورد. مثلاً برای $\theta = (\theta_1, \theta_2)$ موضوع را شرح می‌دهیم.

فرض کنید X دارای چگالی $f(x;\theta_1,\theta_2)$ باشد. همچنین فرض می‌کنیم تابع f از

نظر مشتق‌گیری و انتگرال‌گیری دارای شرایط لازم باشد. ماتریس ۲ × ۲ زیر را ماتریس اطلاع برای $\theta = (\theta_1, \theta_2)$ می‌نامیم.

$$I_X(\theta_1, \theta_2) = I_X(\theta) = \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{pmatrix}$$

لم ۱. فرض کنید X داری توزیع B۱۲ باشد، آنگاه

$$E(X^r) = \frac{\Gamma(\frac{r}{c}+1)\Gamma(k-\frac{r}{c})}{\Gamma(k)}$$

که در آن $\Gamma(.)$ تابع گاما می‌باشد.

اثبات: با استفاده از (۱) داریم

$$E(X^r) = ck\int_0^\infty x^{c+r-1}(1+x^c)^{-(k+1)}dx$$

که با استفاده از تغییر متغیر $t = 1 + x^c$

$$E(X^r) = k\int_1^\infty (t-1)^{\frac{r}{c}}t^{-(k+1)}dt = k\int_0^1 z^{k-\frac{r}{c}-1}(1-z)^{r/c}dz$$

$$= kB(k-\frac{r}{c}, \frac{r}{c}+1) = \frac{\Gamma(k-\frac{r}{c})\Gamma(\frac{r}{c}+1)}{\Gamma(k)}$$

لم ۲. فرض کنید X داری توزیع B۱۲ باشد، آنگاه

$$E_k(\frac{X^c \log X}{1+X^c}) = \frac{k}{k+1}E_{k+1}(X^c \log X)$$

اثبات: با استفاده از (۱)

$$E_k(\frac{X^c \log X}{1+X^c}) = \int_0^\infty (\frac{x^c \log x}{1+x^c})ckx^{c-1}(1+x^c)^{-(k+1)}dx$$
$$= \int_0^\infty (x^c \log x)c\frac{k}{k+1}(k+1)x^{c-1}(1+x^c)^{-((k+1)+1)}dx$$
$$= \frac{k}{k+1}E_{k+1}(X^c \log X)$$

لم ۳. فرض کنید X داری توزیع B۱۲ باشد، آنگاه

$$E_k[\frac{X^c(\log X)^2}{(1+X^c)^2}] = \frac{k}{k+2}E_{k+2}[X^c(\log X)^2]$$

اثبات:

$$E_k[\frac{X^c(\log X)^{۲}}{(۱+X^c)^{۲}}] = \int_{۰}^{\infty} \frac{x^c(\log x)^{۲}}{(۱+x^c)^{۲}} ckx^{c-۱}(۱+x^c)^{-(k+۱)}dx$$
$$= \int_{۰}^{\infty} x^c(\log x)^{۲} c\frac{k}{k+۲}(k+۲)x^{c-۱}(۱+x^c)^{-((k+۱)+۲)}dx$$
$$= \frac{k}{k+۲}E_{k+۲}[X^c(\log X)^{۲}]$$

لم ۴. فرض کنید X داری توزیع B۱۲ باشد، آنگاه

$$E_k(\frac{\log X}{۱+X^c}) = \frac{k}{k+۱}E_{k+۱}(\log X)$$

لم ۵. فرض کنید X داری توزیع B۱۲ باشد، آنگاه

$$E_k[\frac{X^c \log X}{(۱+X^c)^{۲}}] = \frac{k}{k+۲}E_{k+۲}(X^c \log X)$$

لم ٦. برای تابع چگالی احتمال $f(x;\theta_۱,...,\theta_k)$ تحت شرایط نظم داریم
الف) امید ریاضی متغیر نمره صفر است، یعنی

$$E(\frac{\partial \ell(\theta_۱,\theta_۲,...,\theta_k)}{\partial \theta_i}) = ۰ \qquad i = ۱,۲,...,k$$

ب)

$$Var(\frac{\partial \ell}{\partial \theta_i}) = -E(\frac{\partial^{۲}\ell}{\partial \theta_i^{۲}}) \qquad i = ۱,۲,...,k$$

ج)

$$Cov(\frac{\partial \ell}{\partial \theta_i}, \frac{\partial \ell}{\partial \theta_j}) = -E(\frac{\partial^{۲}\ell}{\partial \theta_i \partial \theta_j}) \qquad i,j = ۱,۲,...,k$$

که در آن

$$\ell(\theta_۱,\theta_۲,...,\theta_k) = \sum_{i=۱}^{n} \log f(x_i;\theta_۱,\theta_۲,...,\theta_k)$$

۳ لگاریتم درستنمایی و مشتقات آن

برای بدست آوردن ماتریس اطلاع فیشر این توزیع، ابتدا لازم است تابع درستنمایی مربوطه را تشکیل داده و مشتقات لگاریتم درستنمایی را محاسبه کنیم. برای این منظور فرض می‌کنیم $X_n, ..., X_۲, X_۱$ یک نمونه تصادفی از توزیع (۱) باشد. می‌دانیم که تابع درستنمائی یک نمونه تصادفی عبارتست از

$$\ell = \sum_{i=۱}^{n} \log f(x_i; c, k)$$

آنگاه از (۱) داریم

$$\ell = n\log(ck) + (c - ۱)\sum_{i=۱}^{n} \log x_i - (k + ۱)\sum_{i=۱}^{n} \log(۱ + {x_i}^c)$$

مشتق جزئی ℓ نسبت به c عبارتست از

$$\frac{\partial \ell}{\partial c} = \frac{n}{c} + \sum_{i=۱}^{n} \log x_i - (k + ۱)\sum_{i=۱}^{n} \frac{{x_i}^c \log x_i}{۱ + {x_i}^c} \qquad (۲)$$

همینطور داریم

$$\frac{\partial \ell}{\partial k} = \frac{n}{k} - \sum_{i=۱}^{n} \log(۱ + {x_i}^c) \qquad (۳)$$

از طرفی بسادگی می‌توان نشان داد که

$$\frac{d}{dc}\sum_{i=۱}^{n} \frac{{x_i}^c \log x_i}{۱ + {x_i}^c} = \sum_{i=۱}^{n} \frac{{x_i}^c (\log x_i)^۲}{(۱ + {x_i}^c)^۲}$$

و لذا مشتقات جزئی مرتبه دوم لگاریتم درستنمایی، به شرح ذیل می‌باشد

$$\frac{\partial^۲ \ell}{\partial k^۲} = -\frac{n}{k^۲} \qquad (۴)$$

$$\frac{\partial^۲ l}{\partial c^۲} = -\frac{n}{c^۲} - (k + ۱)\sum_{i}^{n} \frac{{x_i}^c (\log x_i)^۲}{(۱ + {x_i}^c)^۲} \qquad (۵)$$

$$\frac{\partial^۲ \ell}{\partial c \partial k} = -\sum_{i}^{n} \frac{{x_i}^c (\log x_i)^۲}{(۱ + {x_i}^c)^۲} \qquad (۶)$$

۴ امید ریاضی توابع پایه‌ای

بدست آوردن امیدهای مشتقات جزئی مرتبه اول و دوم لگاریتم درستنمایی، مستلزم محاسبه امید ریاضی توابع پایه‌ای زیر می‌باشد

$$\log X, \log(1+X^c), \frac{X^c(\log X)}{1+X^c}, \frac{X^c(\log X)^2}{(1+X^c)^2}$$

فرض کنید X دارای توزیع B۱۲ باشد،در اینصورت با محاسبه امید توابع یاد شده نتایج زیر را داریم.

نتیجه ۱.

$$E(\log X) = -\frac{\gamma+\psi(k)}{c} \qquad (۷)$$

که $\psi = \frac{\Gamma'}{\Gamma}$ و γ ثابت اویلر است.

اثبات : بنا به لم ۱ داریم

$$E(X^r) = \frac{\Gamma(\frac{r}{c}+1)\Gamma(k-\frac{r}{c})}{\Gamma(k)} \qquad (۸)$$

با مشتق گیری از طرفین (۸) نسبت به r داریم

$$E(X^r \log X) = \frac{\Gamma'(\frac{r}{c}+1)\Gamma(k-\frac{r}{c}) - \Gamma(\frac{r}{c}+1)\Gamma'(k-\frac{r}{c})}{c\Gamma(k)} \qquad (۹)$$

با فرض $r = 0$ در (۹)

$$\begin{aligned} E(\log X) &= \frac{\Gamma'(1)\Gamma(k) - \Gamma(1)\Gamma'(k)}{c\Gamma(k)} \\ &= \frac{\Gamma'(1) - \frac{\Gamma'(k)}{\Gamma(k)}}{c} = -\frac{\gamma+\psi(k)}{c} \end{aligned}$$

نتیجه ۲.

$$E[\log(1+X^c)] = \frac{1}{k} \qquad (۱۰)$$

اثبات : بدیهی است هر گاه X دارای توزیع (۱) باشد، آنگاه $Y = ۱ + X^c$ از یک توزیع پارتو با تابع چگالی احتمال $f(y) = ky^{-(k+۱)}, \quad y \geq ۱$ پیروی می‌کند. همینطور $Z = \log Y$، توزیع نمایی منفی با میانگین $k^{-۱}$ دارد. بنابراین نتیجه می‌شود که

$$E[\log(۱ + X^c)] = \frac{۱}{k} \qquad (۷)$$

نتیجه ۳.

$$E[(\log X)^۲] = \frac{\frac{\pi^۲}{۶} + \gamma^۲ + ۲\gamma\psi(k) + (\psi(k))^۲ + \psi'(k)}{c^۲} \qquad (۱۱)$$

اثبات : با مشتق‌گیری از طرفین (۹) بدست می آوریم

$$E[X^r(\log X)^۲] = \frac{\Gamma''(\frac{r}{c}+۱)\Gamma(k-\frac{r}{c}) - ۲\Gamma'(\frac{r}{c}+۱)\Gamma'(k-\frac{r}{c}) + \Gamma(\frac{r}{c}+۱)\Gamma''(k-\frac{r}{c})}{c^۲\Gamma(k)}$$

با فرض $r = ۰$ در عبارت فوق داریم

$$\begin{aligned} E[(\log X)^۲] &= \frac{\Gamma''(۱)\Gamma(k) - ۲\Gamma'(۱)\Gamma'(k) + \Gamma(۱)\Gamma''(k)}{c^۲\Gamma(k)} \\ &= \frac{\frac{\pi^۲}{۶} + \gamma^۲ + ۲\gamma\psi(k) + (\psi(k))^۲ + \psi'(k)}{c^۲} \end{aligned}$$

که ψ' مشتق تابع ψ می‌باشد.

نتیجه ۴.

$$E_k(\frac{X^c \log X}{۱ + X^c}) = \frac{k}{k+۱} E_{k+۱}(X^c \log X) \qquad (۱۲)$$

(نوشتن امید بصورت E_kبرای تاکید بر نقش k در تابع چگالی احتمال توزیع مورد نظر می‌باشد).

اثبات : با توجه به لم ۲ داریم

$$E_k(\frac{X^c \log X}{۱ + X^c}) = \frac{k}{k+۱} E_{k+۱}(X^c \log X)$$

حال با استفاده از معادله (۹) و جایگذاری $r = c$ و همچنین جایگزین کردن $k + ۱$ به جای k داریم

$$E_k(\frac{X^c \log X}{۱ + X^c}) = \frac{k}{k+۱}[\frac{\Gamma'(۲)\Gamma(k) - \Gamma(۲)\Gamma'(k)}{c\Gamma(k+۱)}]$$

$$= \frac{۱ - \gamma - \psi(k)}{c(k+۱)}$$

نتیجه ۵. برای تابع ψ رابطه بازگشتی زیر را داریم

$$\psi(k+۱) = \psi(k) + \frac{۱}{k} \qquad (۱۳)$$

اثبات: می‌توانیم بنویسیم

$$E_k(\frac{X^c \log X}{۱ + X^c}) = E_k(\log X) - E_k(\frac{\log X}{۱ + X^c})$$

با استفاده از لم ۴ داریم

$$E_k(\frac{X^c \log X}{۱ + X^c}) = E_k(\log X) - \frac{k}{k+۱} E_{k+۱}(\log X)$$

حال با استفاده از معادلات (۷) و (۱۲) داریم

$$\frac{۱ - \gamma - \psi(k)}{c(k+۱)} = -\frac{\gamma + \psi(k)}{c} + \frac{k}{k+۱}(\frac{\gamma + \psi(k+۱)}{c})$$

$$\frac{۱ - \gamma - \psi(k)}{c(k+۱)} = \frac{-\gamma k - \gamma - k\psi(k) - \psi(k) + k\gamma + k\psi(k+۱)}{c(k+۱)}$$

$$۱ = -k\psi(k) + k\psi(k+۱)$$

نهایتاً داریم

$$\psi(k+۱) = \psi(k) + \frac{۱}{k}$$

نتیجه ٦.

$$E_k[\frac{X^c(\log X)^۲}{(۱ + X^c)^۲}] = \frac{k}{c^۲(k+۱)(k+۲)}[\frac{\pi^۲}{٦} + \gamma^۲ - ۲\gamma$$

$$+۲(\gamma - ۱)\psi(k+۱) + (\psi(k+۱))^۲ + \psi'(k+۱)] \qquad (۱۴)$$

اثبات: با توجه به لم ۳ داریم

$$E_k[\frac{X^c(\log X)^۲}{(۱+X^c)^۲}] = \frac{k}{k+۲}E_{k+۲}(X^c(\log X)^۲)$$

یا بصورت زیر داریم

$$E_k[\frac{X^c(\log X)^۲}{(۱+X^c)^۲}] = E_k(\frac{(\log X)^۲}{۱+X^c}) - E_k(\frac{(\log X)^۲}{(۱+X^c)^۲})$$

$$= \frac{k}{k+۱}E_{k+۱}[(\log X)^۲] - \frac{k}{k+۲}E_{k+۲}[(\log X)^۲] \qquad (۱۵)$$

با قرار دادن (۱۱) در (۱۵) و با استفاده از نتیجه ۵ و همچنین رابطه بازگشتی $\psi'(k+) = \psi'(k+۱) - \frac{۱}{(k+۱)^۲}$ داریم (۲

$$\begin{aligned}
E_k[\frac{X^c(\log X)^۲}{(۱+X^c)^۲}] &= \frac{k}{k+۱}E_{k+۱}[(\log X)^۲] - \frac{k}{k+۲}E_{k+۲}[(\log X)^۲] \\
&= \frac{k}{k+۱}\{\frac{\frac{\pi^۲}{۶}+\gamma^۲+۲\gamma\psi(k+۱)+(\psi(k+۱))^۲+\psi'(k+۱)}{c^۲}\} \\
&- \frac{k}{k+۲}\{\frac{\frac{\pi^۲}{۶}+\gamma^۲+۲\gamma\psi(k+۲)+(\psi(k+۲))^۲+\psi'(k+۲)}{c^۲}\} \\
&= \frac{k}{c^۲(k+۱)(k+۲)}[\frac{\pi^۲}{۶}+\gamma^۲+۲\gamma k\psi(k+۱)+۴\gamma\psi(k+۱) \\
&+ (k+۲)(\psi(k+۱))^۲+(k+۲)\psi'(k+۱)-۲\gamma k\psi(k+۲)-۲\gamma\psi(k+۲) \\
&- (k+۱)(\psi(k+۲))^۲-(k+۱)\psi'(k+۲)] \\
&= \frac{k}{c^۲(k+۱)(k+۲)}[\frac{\pi^۲}{۶}+\gamma^۲-۲\gamma+۲(\gamma-۱)\psi(k+۱)+(\psi(k+۱))^۲+\psi'(k+۱)]
\end{aligned}$$

۵ اطلاع فیشر در حالت دو پارامتری

اکنون که امید ریاضی توابعی پایه‌ای از توزیع B۱۲ را محاسبه کرده‌ایم، بسادگی می‌توانیم امید ریاضی مشتقات جزئی دوم تابع درستنمایی را به منظور تعیین درایه‌های ماتریس اطلاع فیشر این توزیع، معرفی کنیم.

۱– میزان اطلاع فیشر توزیع B۱۲ نسبت به پارامتر c

از معادله (۵) داریم

$$E(\frac{\partial^۲\ell}{\partial c^۲}) = -\frac{n}{c^۲} - n(k+۱)E[\frac{X^c(\log X)^۲}{(۱+X^c)^۲}]$$

با استفاده از (۱۴)

$$E(\frac{\partial^۲\ell}{\partial c^۲}) = -\frac{n}{c^۲}\{۱ + \frac{k}{k+۲}[\frac{\pi^۲}{۶} + \gamma^۲ - ۲\gamma + ۲(\gamma-۱)\psi(k+۱) + (\psi(k+۱))^۲ + \psi'(k+۱)]\}$$

بنا به لم ۶.ب می‌دانیم میزان اطلاع نسبت به پارامتر c عبارتست از

$$-E(\frac{\partial^۲\ell}{\partial c^۲})$$

۲– میزان اطلاع فیشر توزیع B۱۲ نسبت به پارامتر k

با استفاده از (۴) امید کمیت معلوم $\frac{\partial^۲\ell}{\partial k^۲}$ عبارتست از

$$E(\frac{\partial^۲\ell}{\partial k^۲}) = E(\frac{-n}{k^۲}) = \frac{-n}{k^۲}$$

بنا به لم ۶.ب در مورد میزان اطلاع نسبت به پارامتر k داریم

$$-E(\frac{\partial^۲\ell}{\partial k^۲}) = \frac{n}{k^۲}$$

۳– میزان اطلاع فیشر توزیع B۱۲ نسبت به پارامترهای c و k

از معادله (۶) و با استفاده از معادله (۱۲) داریم:

$$E(\frac{\partial^۲\ell}{\partial c\partial k}) = -nE(\frac{X^c\log X}{۱+X^c}) = -n[\frac{۱-\gamma-\psi(k)}{c(k+۱)}]$$

و بنا به لم ۶.پ در مورد میزان اطلاع نسبت به پارامترهای c و k داریم؛

$$-E(\frac{\partial^۲\ell}{\partial c\partial k}) = n[\frac{۱-\gamma-\psi(k)}{c(k+۱)}]$$

توجه داریم که علیرغم شکل انعطاف ناپذیر نسبی عناصر ماتریس اطلاع فیشر بدست آمده، به راحتی می‌توان آنها را با استفاده از روشهای عددی و با به عدد درآوردن ψ و ψ' محاسبه کرد.

۶– پارامترسازی

چندین روش برای معرفی یک پارامتر مقیاس در معادله (۱) وجود دارد که عادی‌ترین آن تعریف $Y = \theta X$ به ازای $\theta > ۰$ می‌باشد.در اینصورت Y دارای تابع توزیع تجمعی $F(\frac{y}{\theta}; c, k)$ و تابع چگالی احتمال $\frac{۱}{\theta} f(\frac{y}{\theta}; c, k)$ می‌باشد.این نحوه پارامترگذاری باعث بوجود آمدن رابطه $x_i = \frac{y_i}{\theta}$ می‌شود. لذا برای داده‌های $Y_۱, Y_۲, ..., Y_n$ داریم

$$f(y; c, k) = \frac{۱}{\theta} ck (\frac{y}{\theta})^{c-۱} [(۱ + \frac{y}{\theta})^c]^{-(k+۱)}$$

در این حالت، لگاریتم درستنمایی ℓ^y عبارتست از

$$\ell^y = n \log ck - nc \log \theta + (c - ۱) \sum_{i=۱}^{n} \log y_i - (k + ۱) \sum_{i=۱}^{n} \log(۱ + (\frac{y_i}{\theta})^c)$$

مشتقات اول $\frac{\partial \ell^y}{\partial c}$ و $\frac{\partial \ell^y}{\partial k}$ را برای حالت دو پارامتری بدست آوردیم و اما در مورد $\frac{\partial \ell^y}{\partial \theta}$ داریم

$$\frac{\partial \ell^y}{\partial \theta} = -\frac{nc}{\theta} - (k + ۱) \sum_{i=۱}^{n} \frac{c(\frac{y_i}{\theta})^{c-۱}(-\frac{y_i}{\theta^۲})}{۱ + (\frac{y_i}{\theta})^c}$$

$$= -\frac{nc}{\theta} - \frac{c(k + ۱)}{\theta} \sum_{i=۱}^{n} \frac{{x_i}^c}{۱ + {x_i}^c} \qquad (۱۶)$$

همچنین مشتقات جزئی دوم $\frac{\partial^۲ \ell^y}{\partial c \partial k}, \frac{\partial^۲ \ell^y}{\partial k^۲}, \frac{\partial^۲ \ell^y}{\partial c^۲}$ در حالت دو پارامتری محاسبه شدند و در مورد بقیه مشتقات دوم روابط زیر برقرار است

$$\frac{\partial^۲ \ell^y}{\partial k \partial \theta} = \frac{c}{\theta} \sum_{i=۱}^{n} \frac{{x_i}^c}{۱ + {x_i}^c} \qquad (۱۷)$$

$$\frac{\partial^۲ \ell^y}{\partial c \partial \theta} = -\frac{n}{\theta} + \frac{(k + ۱)}{\theta} \sum_{i=۱}^{n} \frac{{x_i}^c}{۱ + {x_i}^c} + \frac{c(k + ۱)}{\theta} \sum_{i=۱}^{n} \frac{{x_i}^c \log x_i}{(۱ + {x_i}^c)^۲} \qquad (۱۸)$$

$$\frac{\partial^۲ \ell^y}{\partial \theta^۲} = \frac{nc}{\theta^۲} - \frac{c(k+۱)}{\theta^۲}\sum_{i=۱}^{n}\frac{{x_i}^c}{۱+{x_i}^c} - \frac{c^۲(k+۱)}{\theta^۲}\sum_{i=۱}^{n}\frac{{x_i}^c}{(۱+{x_i}^c)^۲} \qquad (۱۹)$$

برای بدست آوردن امیدهای مشتقات لگاریتم درستنمایی علاوه بر امیدهای بدست آمده در حالت دوپارامتری، اکنون به امید عبارات زیر نیز احتیاج داریم

$$\frac{X^c}{۱+X^c}, \frac{X^c}{(۱+X^c)^۲}, \frac{X^c \log X}{(۱+X^c)^۲}$$

این امیدها می‌توانند بر حسب امیدهای موجود و با بکارگیری معادله (۱) بصورت زیر بیان شوند

$$E(\frac{X^c}{۱+X^c}) = E_k(\frac{X^c}{۱+X^c}) = \frac{k}{k+۱}E_{k+۱}(X^c)$$

$$= \frac{k}{k+۱}\frac{\Gamma(k)}{\Gamma(k+۱)} = \frac{۱}{k+۱} \qquad (۲۰)$$

$$E[\frac{X^c}{(۱+X^c)^۲}] = E_k[\frac{X^c}{(۱+X^c)^۲}] = \frac{k}{k+۲}E_{k+۲}(X^c)$$

$$= \frac{k}{(k+۱)(k+۲)} \qquad (۲۱)$$

تذکر: در این رهیافت از $r = c$ در معادله (۸) استفاده شده است. همچنین برای بدست آوردن معادله (۱۹) k را با $k+۱$ و برای بدست آوردن معادله (۲۰) k را با $k+۲$ جایگزین کرده‌ایم.

۳. با توجه به لم ۵

$$E[\frac{X^c \log X}{(۱+X^c)^۲}] = E_k[\frac{X^c \log X}{(۱+X^c)^۲}] = \frac{k}{k+۲}E_{k+۲}(X^c \log X)$$

حال با استفاده از (۹) وقتی که $r = c$ داریم

$$E[\frac{X^c \log X}{(۱+X^c)^۲}] = \frac{k}{k+۲}[\frac{۱-\gamma-\psi(k)}{c(k+۱)}] \qquad (۲۲)]$$

این نتیجه همچنین می‌تواند با نوشتن امید بصورت

$$E_k[\frac{X^c \log X}{(۱+X^c)^۲}] = E_k(\frac{\log X}{۱+X^c}) - E_k(\frac{\log X}{(۱+X^c)^۲})$$

$$= \frac{k}{k+1} E_{k+1}(\log X) - \frac{k}{k+2} E_{k+2}(\log X)$$

و با استفاده از معادله (۷) و با نوشتن امید بصورت

$$E_k[\frac{X^c \log X}{(1+X^c)^2}] = \frac{k}{k+1} E_{k+1}[\frac{X^c \log X}{1+X^c}]$$

و استفاده از معادله (۱۲) بدست آید.

برای بدست آوردن امیدهای مشتقات لگاریتم درستنمایی در حالت سه پارامتری، توجه به این نکته که در بین مشتقات اول، $\frac{\partial \ell^y}{\partial c}$ و $\frac{\partial \ell^y}{\partial k}$ و در بین مشتقات دوم، $\frac{\partial^2 \ell^y}{\partial k^2}$، $\frac{\partial^2 \ell^y}{\partial c \partial k}$ و $\frac{\partial^2 \ell^y}{\partial c^2}$ از حالت دوپارامتری تبعیت می‌کنند، حائز اهمیت است. اما در مورد امید سایر مشتقات در حالت سه‌پارامتری به شرح ذیل داریم

۱. از معادله (۱۶)، امید $\frac{\partial \ell^y}{\partial \theta}$ را با استفاده از معادله (۲۰) و با ساده کردن بدست می‌آوریم

$$E(\frac{\partial \ell^y}{\partial \theta}) = -\frac{nc}{\theta} + \frac{nc(k+1)}{\theta} E(\frac{X^c}{1+X^c}) = 0$$

۲. از معادله (۱۷) و با استفاده از معادله (۲۰) داریم

$$E(\frac{\partial^2 \ell^y}{\partial \theta \partial k}) = \frac{nc}{\theta} E(\frac{X^c}{1+X^c}) = \frac{nc}{\theta(k+1)}$$

۳. از معادله (۱۸) و با استفاده از معادلات (۲۰) و (۲۲) داریم

$$\begin{aligned} E(\frac{\partial^2 \ell^y}{\partial c \partial k}) &= \frac{n}{\theta}\{-1 + (k+1)E(\frac{X^c}{1+X^c}) + c(k+1)E[\frac{X^c \log X}{(1+X^c)^2}]\} \\ &= \frac{nk[1-\gamma-\psi(k+1)]}{\theta(k+2)} \end{aligned}$$

۴. از معادله (۱۹) و با استفاده از معادلات (۲۰) و (۲۱) داریم

$$\begin{aligned} E(\frac{\partial^2 \ell^y}{\partial \theta^2}) &= \frac{nc}{\theta^2}\{1 - (k+1)E(\frac{X^c}{1+X^c}) - c(k+1)E[\frac{X^c}{(1+X^c)^2}]\} \\ &= -\frac{nkc^2}{\theta^2(k+2)} \end{aligned}$$

مراجع

احمدی، ج، (۱۳۷۳)، رکوردها در توزیعهای پیوسته، کارشناسی ارشد، دانشگاه فردوسی مشهد.

Ahmadi, J. and Arghami, NR. (2001), *On the Fisher information in record values*. Metrika 53: 195-206

Hofmann, G. and Nagaraja, N. (2002), *Fisher information in Record Data*. to apaer in Metrika.

Park, S. (1996), *Fisher information in order statisrics*. J. Amer. Stat. As.91, 385-390.

Timr, L. Fry. (1993),*univariate and multivariate Burr distributions*. PJS, A,1-24.

Zheng, G. and Gastwirth, JL. (2000), *Where is the Fisher information in an ordered sample?*. Statistica Sinica 10, 1267-1280.

برآوردگر رگرسیونی عام و کاربرد آن در بررسی هزینه خانوار

حمید بیدرام[۱]، محمد صالحی مرزیجرانی[۲]

$P۱۱۱۲۸$

[۱] دانشکده ریاضی و کامپیوتر خوانسار، وابسته به دانشگاه اصفهان
[۲] دانشکده ریاضی، دانشگاه صنعتی اصفهان

چکیده: در این مقاله، برآوردگر رگرسیونی عام را که برپایه برآوردگر هارویتز — تامپسون بنا شده است، معرفی می‌کنیم. علاقمندیم درهنگام مطالعه یک بررسی علاوه بر متغیرهای کمکی از متغیرهای مشترکی که در هر دو بررسی وجود دارد جهت افزایش دقت استفاده نماییم. زیچانگ (۱۹۹۰) در بررسی هزینه خانوار ایالات متحده به خوبی این کار را به انجام رسانیده‌است. در این مقاله ضمن معرفی روش رنسن و نیونبروک (۱۹۹۷) که روش زیچانگ را تا حد زیادی ساده و کاربردی تر نموده‌است با مقایسه این دو روش یک مدل رگرسیونی ارائه خواهیم داد که همزمان شامل هر دو متغیرهای کمکی و مشترک بوده و خواهیم دید با این روش نحوه اجرای کار به مراتب ساده‌تر خواهد شد.

واژه های کلیدی: برآوردگر رگرسیونی، برآوردگر رگرسیونی عام، برآوردگر هارویتز—تامپسون، متغیر کمکی، نمونه‌گیری با احتمالات نابرابر.

۱ مقدمه

برآوردگر رگرسیونی عام از جمله برآوردگرهایی است که پتانسیل عملی بالایی دارد. برآورد هزینه خانوار هم اینک درتمام ملل دنیا صورت می‌گیرد. که در آن نمونه گیری چند مرحله‌ای انجام می‌شود و در هر مرحله اولا باید وزنهای متفاوتی برای انتخاب زیر دامنه‌ها و واحدهای مختلف در نظر گرفت ثانیا به علت تفاوت سطح فرهنگ در خانوارها روش جمع آوری اطلاعات نیز باید متفاوت باشد. برآوردگر رگرسیونی عام قادر است از متغیرهای مشترک اندازه‌گیری شده در دو یا چند بررسی علاوه بر متغیرهای کمکی استفاده‌نموده و دقت را افزایش دهد. در این مقاله به کاربرد عملی این برآوردگر در بررسی هزینه خانوار اشاره خواهیم نمود. دراینجاحجم جامعه را با N و اعضای آن را $Y_۱, Y_۲, ..., Y_N$ و پارامتر مجموع کل که معمولا نامعلوم وقابل برآورد است را با $t = \Sigma_{i=۱}^{N} Y_i$ نمایش می‌دهیم. به دلیل اهمیت برآوردگر هارویتز—تامپسون در برآوردگر رگرسیونی عام این برآوردگر را معرفی می‌کنیم.

۲ برآوردگر هارویتز—تامپسون

جامعه‌ای با حجم N با اعضای $Y_1, Y_2, ..., Y_N$ را در نظر بگیرید. اگر احتمال انتخاب اعضاء را با π_k نمایش دهیم مجموع کل یعنی $t = \Sigma_{i=1}^{N} Y_i$ رابا

$$\widehat{t_\pi} = \Sigma_{i=1}^{n} \frac{Y_i}{\pi_i}.I_i \qquad (1)$$

برآورد می‌کنیم که آن را برآوردگر هارویتز—تامپسون می‌نامند و در آن

$$I_k = \begin{cases} 1 & \text{اگر عضو } k \text{—ام در نمونه باشد} \\ 0 & \text{در غیر اینصورت} \end{cases}$$

واضح است که اگر در بررسی از نمونه‌گیری تصادفی ساده استفاده‌کنیم یعنی به همه اعضاء احتمال انتخاب یکسانی را تخصیص دهیم به ازای هر k در جامعه، $\pi_k = \frac{n}{N}$ و بنابراین:

$$\begin{aligned} \widehat{t_\pi} &= \Sigma_{i=1}^{n} \frac{Y_i}{\frac{n}{N}}.I_i \\ &= N\Sigma_{i=1}^{n} \frac{Y_i}{n}.I_i \\ &= N\overline{Y_s} \end{aligned} \qquad (2)$$

که $\overline{Y_s}$ میانگین اعضاء نمونه است و (۲) همان برآورد مجموع کل در نمونه‌گیری تصادفی ساده است. این برآوردگر دارای خواص کاربردی زیادی است که در [۱] مفصلا بیان شده‌است.

۳ برآوردگر رگرسیونی عام

فرض کنیم بردارمتغیرهای کمکی برای k—امین عضو به صورت

$$\underline{x_k} = (x_{1k}, x_{2k}, ..., x_{jk})' \quad , \quad k \in S$$

می‌باشد که S مجموعه‌اعضای نمونه است. برآوردگر رگرسیونی عام $t = \Sigma_{i=1}^{N} Y_i$ (مجموع کل نامعلوم جامعه) بصورت زیر فرموله می‌شود.

$$\widehat{t_r} = \widehat{t_\pi} + \Sigma_{j=1}^{J} \widehat{\beta}_j (t_{x_j} - \widehat{t_{x_j \pi}}) \qquad (3)$$

که در آن $\hat{t_\pi}$ برآوردگر هارویتز—تامپسون t و $\widehat{t_{x_j\pi}} = \Sigma_{k=1}^{N} \frac{x_{j_k}}{\pi_k} . I_k$ برآوردگر هارویتز—تامپسون $t_{x_j} = \Sigma_{k=1}^{N} x_{j_k}$ (مجموع کل متغیرهای کمکی) است و همچنین با در نظر گرفتن مدل رگرسیون

$$Y = X'\beta + \epsilon$$

که‌در آن $Y = (Y_1, Y_2, ..., Y_N)'$ یک بردار $N \times 1$، X یک ماتریس $J \times N$ از متغیرهای کمکی، β یک بردار j تایی بصورت $\beta = (\beta_1, \beta_2, ..., \beta_j)$ و ϵ یک بردار $N \times 1$ است. β با روش کمترین مربعات خطا بصورت زیر برآورد می‌شود:

$$\hat{\beta} = (X\Sigma X')^{-1} X\Sigma^{-1} Y \qquad (۴)$$

که در آن

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_N{}^2 \end{bmatrix}$$

ماتریس واریانس—کوواریانس Y است.

۴ استفاده از برآوردگر رگرسیونی عام بعنوان یک تعدیل کننده وزنی در بررسی هزینه خانوار

نمونه‌ای شامل n خانوار را در نظر بگیرید. بطوریکه بردار $\underline{S} = (s_1, s_2, ..., s_n)'$ که در آن $s_k = \pi_k{}^{-1}$ و π_k احتمال انتخاب خانوار k—ام‌می باشد، معلوم و مشخص باشد. به بردار S اصطلاحا بردار وزنهای آغازین (ابتدایی) می‌گویند. همچنین فرض کنیم J حجره طبقه‌بندی پسین وجود داشته‌باشد بطوریکه تعداد افراد جامعه (N_j) برای هر حجره معلوم باشد. برای مثال در بررسی هزینه مصرف کننده ایالات متحده تعداد $J = ۴۸$ حجره مطابق با ترکیبات ۲ جنس، ۲ نژاد (سیاه و غیر سیاه) و ۱۲ رده سنی وجود دارد.
در واقع بردار $\underline{N} = (N_1, N_2, ..., N_j)'$، بردار متغیرهای کمکی است و N_j—ها تعداد افراد در حجره j—ام را نشان می‌دهد. ترکیب خانوارهای نمونه توسط ماتریس $X = (x_{k_j})$، ($1 \le k \le n$, $1 \le j \le J$) مشخص می‌گردد که x_{k_j} تعداد اشخاص نمونه در j—امین حجره از خانوار k—ام می‌باشد. بطور کلی ماتریس X را بصورت زیر نمایش

می‌دهیم:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1J} \\ x_{21} & x_{22} & \cdots & x_{2J} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nJ} \end{bmatrix}_{n \times J}$$

واضح است که اگر برداری از ماتریس فوق بصورت $(2, 1, 0, ..., 0)$ باشد بیانگر آن است که خانوار مذکور شامل ۲ نفر در حجره اول و یک نفر در حجره دوم خواهدبود.
با استفاده از وزنهای آغازین می‌توان تعداد افراد در حجره j—ام را از رابطه

$$\widehat{N_j} = \Sigma_k x_{k_j} s_k$$

یا بطور کلی از رابطه

$$\underline{\widehat{N}} = X' \underline{S} \qquad (۵)$$

برآورد کرد.
مسلما $\underline{\widehat{N}} \neq \underline{N}$. بنابراین دنبال بردار تعدیل کننده وزنی هستیم تا بواسطه آن $\underline{\widehat{N}} = \underline{N}$. زیچانگ (۱۹۹۰ و ۱۹۸۶) به روش GLS (General Least Square) بردار جدیدی از وزنها یعنی $\underline{W} = (w_1, w_2, ..., w_n)'$ بدست آورد بطوریکه $\underline{N} = X'\underline{W}$. روش به اینصورت است که با در نظر گرفتن قید $\underline{N} = X'\underline{W}$ معیار

$$\Sigma_k (w_k - s_k)/s_k = (\underline{W} - \underline{S})' \Lambda^{-1} (\underline{W} - \underline{S}) \qquad (۶)$$

را که در آن $\Lambda_{n \times n} = diag(\underline{S})$، کمینه می‌کنند بنابراین بردار $\underline{W}$ از کمینه کردن معیار (۶)، بصورت زیر بدست می آید:

$$\underline{W} = \underline{S} + \Lambda \underline{X} (\underline{X}' \Lambda \underline{X})^{-1} (\underline{N} - \underline{X}' \underline{S}) \qquad (۷)$$

حال اگر فرض کنیم Y یک بردار $n \times 1$ از صفت مورد مطالعه در نمونه باشد با داشتن بردار $\underline{W}$ بسادگی برآوردگر رگرسیونی عام مجموع کل صفت موردمطالعه یعنی $t = \Sigma_k Y_k$ بصورت زیر بدست می آید:

$$\begin{aligned} \widehat{t_r} &= \underline{Y}' \underline{W} \\ &= \underline{Y}' (\underline{S} + \Lambda \underline{X} (\underline{X}' \Lambda \underline{X})^{-1} (\underline{N} - \underline{X}' \underline{S})) \\ &= \underline{Y}' \underline{S} + Y' \Lambda \underline{X} (\underline{X}' \Lambda \underline{X})^{-1} (\underline{N} - \underline{X}' \underline{S}) \end{aligned} \qquad (۸)$$

که با مقایسه با رابطه (۳) برای بر آوردگر رگرسیونی عام ملاحظه می‌کنیم:

$$\underline{N} = \underline{X'W} = t_x$$
$$\widehat{\beta'} = Y'\Lambda^{-1}\underline{X}(\underline{X'}\Lambda\underline{X})^{-1}$$
$$\widehat{t_\pi} = \underline{Y'S}$$
$$\underline{X'S} = \widehat{t_{x_\pi}}$$

بنابراین بنظر می‌رسد با داشتن وزنهای $\underline{W}$ براحتی با ضرب بردار نمونه Y در بردار وزنهای $\underline{W}$، برآوردگر رگرسیونی عام بدست آید.

۵ وجود متغیرهای مشترک دو بررسی

زیچانگ (۱۹۹۰ و ۱۹۸۶) در بررسی هزینه خانوار جهت افزایش دقت جمع‌آوری اطلاعات نمونه را به دو طریق اعمال کرد. یک نمونه را با روش مصاحبه‌ای (پرسش و پاسخ بین مامور و خانوار) و نمونه دیگر را با روش یادداشتی (ارسال پرسشنامه به خانوارها و تکمیل آنها توسط خانوارها) انجام داد.ایشان به این نتیجه مهم دست یافت که جهت افزایش دقت برآورد بهتر است علاوه بر استفاده از متغیرهای کمکی دو بررسی از متغیرهای مشترکی (مانند ناحیه نمونه‌گیری، مالکیت، نوع خانوار و...) که در هر دو بررسی وجود دارد استفاده کرد. انتظار داریم که برآورد مجموع کل توسط هر دو طرح یکسان باشد. زیچانگ با مشخص کردن بررسی اول و دوم بترتیب با اندیسهای ۱ و ۲ رابطه (۸) را با بکار بردن نمادهای زیر، تعمیم داد.

$$\underline{S} = \begin{bmatrix} \underline{S_1} \\ \underline{S_2} \end{bmatrix} \quad , \quad \Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$
$$N_X = \begin{bmatrix} N_{X^\circ} \\ N_{X^\circ} \\ \circ \end{bmatrix}$$
$$X = \begin{bmatrix} X_1^\circ & \circ & X_1^c \\ \circ & X_2^\circ & -X_2^c \end{bmatrix} \qquad (۹)$$

که در آنها:

$\underline{S_1}$: بردار وزنهای آغازین بررسی اول.

$\underline{S_2}$: بردار وزنهای آغازین بررسی دوم.

Λ_{11} و Λ_{22} ماتریسهای قطری بصورت $\Lambda_{11} = diag(\underline{S_1})$ و $\Lambda_{22} = diag(\underline{S_2})$ و $\Lambda_{12} =$

$\Lambda_{۲۱} = ۰$

$X_۱^۰$ و $X_۲^۰$: ماتریسهای کمکی بررسی اول و دوم.

$X_۱^c$ و $X_۲^c$: ماتریسهای مشترک بررسی اول و دوم.

$N_{X^۰}$: مجموع عناصر ستونهای ماتریس X در جامعه است.

بنابراین بردار $\underline{W}$ در رابطه (۷) با جایگذاری ماتریسها و بردارهای (۹) محاسبه می‌شود. مثال کاربردی از این روش در [۳] بیان شده‌است. همانطور که ملاحظه می‌کنیم استفاده از متغیرهای مشترک به فرم ماتریس‌مشکل و بعضا با اشتباه همراه است. روشی را در زیر جهت استفاده بهتر متغیرهای مشترک بیان می‌کنیم.

٦ استفاده از متغیرهای مشترک بعنوان متغیرهای رگرسور اضافی

فرض کنید دو بررسی در نمونه دارای متغیرهای مشترکی باشند اگر مجموع کل این متغیرها معلوم باشند آنگاه متغیرهای مذکور را بعنوان متغیرهای کمکی در برآورد رگرسیونی هر کدام از بررسیها در نظر می‌گیریم ولی اگر مجموع این متغیرها نامعلوم باشد ممکن است برآورد مجموع کل آنها را همراه با متغیرهای کمکی به عنوان متغیرهای توضیحی اضافی در برآورد رگرسیونی عام هر دو بررسی بکار ببریم. با مشخص شدن وزنهای برآورد بوجود آمده، مجموع کل متغیرهای کمکی را دوباره برآورد می‌کنیم. در اینجا برآورد رگرسیونی جدیدی بدست می‌آوریم که به آن برآورد رگرسیونی تعدیل شده می‌گوییم و بر اساس دو نمونه $S_۱$ و $S_۲$ به معرفی آن خواهیم پرداخت. در واقع با معرفی این روش وجود متغیرهای مشترک را بعنوان یک جمله اضافی در مدل رگرسیون بکار می‌بریم.

فرض کنید در اینجا برای هر عضو u_k یک بردار n—تایی $\underline{z_k}$ از متغیرهای مشترک داشته باشیم و

$t_z = Z'l$ مجموع کل متغیرهای مشترک در جامعه‌باشد که در آن

$$l_{N\times ۱} = \begin{bmatrix} ۱ \\ ۱ \\ \vdots \\ ۱ \end{bmatrix} \quad , \quad Z_{N\times q} = \begin{bmatrix} z_۱' \\ z_۲' \\ \vdots \\ z_N' \end{bmatrix}$$

برآوردگر هارویتز—تامپسون برای t_z را بصورت $\widehat{t_{z\pi}} = Z_s'\Lambda_s^{-۱}l_s$ نشان می‌دهیم که در آن Z_s مقادیر مشاهده شده متغیرهای مشترک نمونه است و Λ_s و l_s قبلا معرفی شده‌اند. رنسن و نیونبروک (۱۹۹۷) برآوردگر رگرسیونی تعدیل شده را با توجه به نمادهای فوق بصورت زیر معرفی کردند:

$$\widehat{t_{Ar}} = \widehat{t_\pi} + \widehat{B'_A}(t_x - \widehat{t_{x\pi}}) + \widehat{D'_A}(\widehat{t_z} - \widehat{t_{z\pi}}) \qquad (۱۰)$$

که $\widehat{B'_A}$ و $\widehat{D'_A}$ از رابطه زیر بدست می آید.

$$\begin{pmatrix} \widehat{B_A} \\ \widehat{D_A} \end{pmatrix} = ((X_s Z_s)' \Lambda_s (X_s Z_s))^{-1} (X_s Z_s)' \Lambda_s Y_s$$

و $\widehat{t_z}$ یک بردار q—تایی از برآوردهای t_z است. در واقع رابطه (۱۰) را با توجه به تعریف برآوردگر رگرسیونی عام به صورت زیر نیز می‌توان نوشت:

$$\begin{pmatrix} t_x - \widehat{t_{x\pi}} \\ \widehat{t_z} - \widehat{t_{z\pi}} \end{pmatrix} \widehat{t_{Ar}} = \widehat{t_\pi} + \begin{pmatrix} \widehat{B_A} \\ \widehat{D_A} \end{pmatrix} \qquad (۱۱)$$

حال به خاصیت اساسی برآوردگر رگرسیونی تعدیل شده می‌پردازیم. در واقع این خاصیت تکمیل کننده روش رنس و نیونبروک می‌باشد.فرض کنید:

$$\begin{aligned}
&\widehat{t_r} = \widehat{t_\pi} + \widehat{\beta}'(t_x - \widehat{t_{x\pi}}) \\
&\widehat{\beta}' = (X'_s \Lambda_s X_s)^{-1} X'_s \Lambda_s Y'_s \\
&\widehat{t_{zr}} = \widehat{t_{z\pi}} + \widehat{L}'(t_x - \widehat{t_{x\pi}}) \\
&\widehat{L} = (X'_s \Lambda_s X_s)^{-1} X'_s \Lambda_s Z_s
\end{aligned}$$

برآوردهای رگرسیونی عام مجموع کل t_Y و t_Z باشند. فرض می‌کنیم X_sپر رتبه $J \geq 1$ باشد اگر $(X_s Z_s)$پر رتبه $J + q$ باشد برآوردگر رگرسیونی تعدیل شده (۱۰) یعنی $\widehat{t_{Ar}}$ بصورت زیر نوشته می‌شود:

$$\widehat{t_{Ar}} = \widehat{t_r} + \widehat{D'_A}(\widehat{t_z} - \widehat{t_{zr1}}) \qquad (۱۲)$$

که‌در آن $\widehat{t_{zr1}}$براوردگر رگرسیونی عام متغیرهای مشترک بررسی اول است و ضریب رگرسیون جزئی $\widehat{D_A}$ بصورت زیر نوشته می‌شود:

$$\widehat{D_A} = (Z'_s R_s Z_s)^{-1} Z'_s R_s Y_s$$

$$R_s = \Lambda_s - \Lambda_s X_s (X_s' \Lambda_s X_s)^{-1} X_s' \Lambda_s$$

اثبات (۱۲) در [۴] آمده است.

از (۱۲) ملاحظه می‌کنیم که برآوردگر رگرسیون تعدیل شده برابراست با برآوردگر رگرسیونی عام باضافه یک جمله تعدیل کننده. این جمله تعدیل کننده یک عامل اصلاح کننده برآوردگر رگرسیونی عام و درعین حال عاملی برای ایجاد سازگاری بین برآورد متغیرهای مشترک از دو بررسی می‌باشد. برآوردگر رگرسیونی تعدیل شد. در (۱۲) بردارهای وزنی n—تایی زیر را نتیجه می‌دهد.

$$W_{As} = W_s + R_s Z_s (Z_s' R_s Z_s)^{-1} (\hat{t_z} - \widehat{t_{zr}}) \qquad (۱۳)$$

که در آن:
$$W_s = \underline{S} + \Lambda_s X_s (X_s' \Lambda_s X_s)^{-1} (t_x - \widehat{t_{x\pi}})$$

یک بردار n—تایی از وزنهایی است که مطابق با برآورد رگرسیون عام $\hat{t_r}$ بدست می آمد. دقت کنیم در اینجا $X_s' R_s = 0$ و بنابراین

$$X_s' W_{As} = X_s' W_s = t_x$$
$$Z_s' W_{As} = \hat{t_z}$$

یعنی بردار وزنی W_{As}—ای پیدا کردیم که علاوه بر اینکه با حاصلضرب آن در X_s' مجموع کل متغیرهای کمکی در جامعه را می‌دهد، با حاصلضرب در Z_s' (که از نمونه مشاهده می‌شود) مجموع کل متغیرهای مشترک در جامعه را برآورد می‌کند. (چون $\hat{t_z}$ برآورد t_z مجموع کل متغیرهای مشترک در جامعه بود.) مثال کلی از هر دو روش زیچانگ و روش مطرح شده فوق در [۴] آمده است.

مراجع

C.E.Sandral ,B.Swenson and J.Wretman, Model Assisted Survey sampling, New York :Springer-Verlag(1990)

K.D. Zieschang,"Generalized Least Squares: An Alternative to Principal person weighting" Washington, DC. Bureau of Labor Satistics (1986).

K.D. Zieschang,"Sample Weighting Methods and Estimation of totals in the consumer Expenditure Surveys", Journal of the American Statiscal Association, No:85, PP. 986-1001,(1990).

حمید بیدرام، : برآوردگر رگرسیونی عام و کاربرد آن در بررسی هزینه خانوار: پایان نامه کارشناسی ارشد، (دانشگاه صنعتی اصفهان— ۱۳۷۹).

برآورد پارامترهای مدل رگرسیونی با رهیافت ماکسیمم آنتروپی

عین‌اله پاشا[۱]، محسن محمدزاده[۲]، علی آقامحمدی[۲]

[۱] گروه آمار، دانشگاه تربیت معلم
[۲] گروه آمار، دانشگاه تربیت مدرس

چکیده: در استفاده از مدل رگرسیون که در اکثر مطالعات اقتصادی و اجتماعی به کار می‌رود، معمولاً با دو مشکل مواجه می‌شویم. نخست اینکه در مسئله تعداد پارامترهای مجهول که باید برآورد شوند بیش از تعداد مشاهدات است و دوم فقدان صحّت شرایط لازم جهت کاربرد مدل. این مشکلات اغلب هنگامی که فرایند جمع‌آوری داده‌ها خوب طراحی نمی‌شوند، اتفاق می‌افتد. یک مسئله در این حالت همخطی داده‌ها است. یعنی دو یا چند متغیر همبستگی خیلی زیادی با هم دارند. البته همخطی کامل به ندرت رخ می‌دهد، ولی همخطی غیر کامل نیز منجر به افزایش واریانس در برآورد پارامترها به روش LS شده، و این مسئله باعث افت توان آزمون در آزمون فرض پارامترهای مدل می‌شود.

اخیراً رهیافتهای متعددی برای مواجه با این گونه مسائل پیشنهاد شده که یکی از آنها رهیافت ماکسیمم آنتروپی تعمیم یافته است که در سال ۱۹۹۳ توسط گالن و جودجی ارائه شده است. در این مقاله ضمن مطالعه این روش کارائی آن را با یک شبیه سازی بررسی می‌کنیم.

۱ مقدمه

هدف از علم آمار حصول اطلاعات از داده‌های نمونه‌ای می‌باشد. در این راستا یکی از پر کاربردترین ابزارهای آماری رگرسیون می‌باشد، که هدف از آن تبیین تغییرات اندازه‌های آزمایشی یک متغیر از روی متغیرهای وابستهٔ دیگری که مقادیر آنها در جریان آزمایش تغییر می‌کند، می‌باشد. با شرکت دادن صریح داده‌های این متغیرهای نافذ در تجزیه و تحلیل آماری، اغلب امکان دارد که از طبیعت رابطهٔ بین این متغیرها مطلع شده، آنگاه از این اطلاع برای توصیف و استنباطهای مربوط به متغیر اولیه مورد علاقه بهره‌گیری کرد. این واقعیت در درجهٔ اول اهمیّت است که بدانیم تجزیه و تحلیل رگرسیون، با برازش یک

مدل بوسیلهٔ کوچکترین توانهای دوم خطا، به دست آوردن فاصله های اطمینان و بالاخره با آزمون فرضهای متعدد کامل نمی‌شود. این گامها فقط بیان نیمی از این مطلب است که استنباطهای آماری را وقتی می توان انجام داد که مدل مفروض، مناسب باشد. در بیشتر مطالعات مربوط به علوم اجتماعی و علوم اقتصادی، روابط بین متغیرها شکل تجربی دارند، به نحوی که هیچ وقت نمی‌توانیم مطمئن باشیم که مدل خاصی صحیح است. در اکثر اوقات فرضها و شرایط لازم برای کاربرد رگرسیون موجود نیست. این موضوع به نوع جمع آوری داده‌ها و... برمی‌گردد. اگرچه روشهای نمونه گیری نیز بسط و توسعه یافته‌اند، اما بیشتر داده‌ها که توسط آماردانان به کار گرفته می‌شوند، محدود به خطاهای نمونه‌ای و غیر نمونه‌ای نظیر خود جامعه، تبدیل داده‌ها، کدبندی و استفاده از تقریب می‌باشد. استفاده از مدل رگرسیون که در اکثر مطالعات اقتصادی و اجتماعی به کار می‌رود، معمولاً با دو مشکل مواجه هست. نخست اینکه در مسئله، تعداد پارامترهای مجهول که باید برآورد شوند بیش از تعداد مشاهدات است و دوم اینکه پارامترها در مسئله ثابت نیستند. این مشکلات اغلب هنگامی که فرایند جمع آوری داده‌ها خوب طراحی نمی‌شوند، اتفاق می‌افتد. یک مسئله در این حالت همخطی داده‌ها است. یعنی دو یا چند متغیر همبستگی خیلی زیادی با هم دارند. در این حالت ماتریس طرح ستونی رتبه کامل نبوده، در نتیجه معکوس ماتریس، حاصل ضرب ماتریس طرح در ترانهادهٔ آن، موجود نیست. البته باید توجه داشت که همخطی کامل بندرت رخ می‌دهد، ولی همخطی غیر کامل نیز منجر به افزایش واریانس در برآورد پارامترهای LS شده، و این مسئله باعث افت توان آزمون در آزمون فرض پارامترهای مدل می‌شود.

اخیراً رهیافتهای متفاوتی برای مواجه با این گونه مسائل به وجود آمده که یکی از آنها رهیافت ماکسیمم آنتروپی تعمیم یافته است. اگر چه ایدهٔ اصلی ماکسیمم آنتروپی به سالهای ۱۹۵۷ برمی‌گردد، ولی مشکلاتی در کاربرد این مفهوم به شکل اصلی آن برای برآورد پارامترهای یک مدل آماری موجود بوده است. مثلاً کمیتهای مجهول باید خاصیت توزیع احتمال داشته باشند، ولی در اکثر اوقات پارامترهای مدلها این گونه نیستند. اگر برای مثال هدف، برآورد پارامترهای مدل رگرسیونی $\underline{Y} = \underline{X}\underline{\beta} + \underline{E}$ باشد، از آنجائیکه $\underline{\beta}$ بردار ثابتی است، توزیع احتمال برای آن بی‌معنی است. اصل ماکسیمم آنتروپی طوری تعمیم داده شده که برای این منظور نیز بتوان از آن استفاده کرد.

۲ ماکسیمم آنتروپی تعمیم یافته

با در نظر گرفتن محدودیتهای فوق جودجی و گُلن (۱۹۹۴) ماکسیمم آنتروپی را به وسیلهٔ تعریف پارامترهای مجهول بر حسب توزیعهای احتمال گسسته تعمیم دادند. برای

توصیف کاربردهای ماکسیمم آنتروپی تعمیم یافته[۱] (GME) مدل رگرسیونی استاندارد زیر را در نظر بگیرید.

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{E} \qquad (۱)$$

که در آن $\underline{Y}$، یک بردار $(T \times ۱)$ بعدی، $\underline{X}$ ماتریس طرح $(T \times K)$ بعدی، $\underline{\beta}$ یک بردار $(K \times ۱)$ بعدی از پارامترهای ناشناخته و $\underline{E}$ نیز یک بردار $(T \times ۱)$ بعدی از خطاها می‌باشد. جودجی و گُلن (۱۹۹۴) فرض کردند که β_kها، مؤلفه‌های بردار $\underline{\beta}$ متغیرهای تصادفی گسسته‌ای باشند، که M مقدار، $۲ \leq M < \infty$ را از یک بازهٔ بسته با احتمالهای متفاوت اختیار می‌کنند. اگر z_{kl} و z_{km} به ترتیب کران پایین و بالای این بازه باشند، آنگاه β_k را می‌توان به عنوان یک ترکیب خطی محدب از کرانها نوشت. به طور مشابه e_tها، مؤلفه‌های بردار $\underline{E}$ را هم می‌توان به عنوان یک متغیر تصادفی گسسته که J مقدار، $۲ \leq J < \infty$ را از یک بازهٔ بسته با احتمالهای متفاوت اختیار می‌کند، در نظر گرفت. در این حالت رابطهٔ (۱) را می‌توان به صورت زیر باز نویسی کرد.

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{E} = \underline{X}Z\underline{P} + V\underline{W}$$

که در آن $\underline{\beta} = Z\underline{P}$ و $\underline{E} = V\underline{W}$، به طوری که Z یک ماتریس $(K \times KM)$ بعدی، از مقادیری که $\underline{\beta}$ اختیار می‌کند و $\underline{P}$ یک بردار $(KM \times ۱)$ بعدی، از احتمالهای ناشناخته است، به طوری که

$$Z\underline{P} = \begin{pmatrix} \underline{Z}'_۱ & ۰ & \cdots & ۰ \\ ۰ & \underline{Z}'_۲ & \cdots & ۰ \\ \vdots & \vdots & \ddots & \vdots \\ ۰ & ۰ & \cdots & \underline{Z}'_k \end{pmatrix} \begin{pmatrix} \underline{P}_۱ \\ \underline{P}_۲ \\ \vdots \\ \underline{P}_k \end{pmatrix}$$

$$\underline{Z}'_۱ = (z_{k۱}, z_{k۲}, \ldots, z_{km}), \qquad \underline{P}'_k = (p_{k۱}, p_{k۲} \ldots, p_{km})$$

$$\underline{Z}'_k\underline{P}_k = \sum_{m=۱}^{M} z_{km}p_{km} = \beta_k, \qquad k = ۱, ۲, \ldots, K$$

$$p_{km} > ۰, \qquad k = ۱, ۲, \ldots, K, \qquad m = ۱, ۲, \ldots, M$$

$$\sum_{m=۱}^{M} p_{km} = ۱, \qquad k = ۱, ۲, \ldots, K$$

[۱] Generalized Maximum Entropy

به طور مشابه V یک ماتریس $(T \times TJ)$ بعدی، از مقادیری که $\underline{E}$ اختیار می‌کند و $\underline{W}$ نیز یک بردار $(TJ \times ۱)$ بعدی، از وزنهای احتمالی است، به طوری که

$$V\underline{W} = \begin{pmatrix} \underline{V}'_۱ & ۰ & \cdots & ۰ \\ ۰ & \underline{V}'_۲ & \cdots & ۰ \\ \vdots & \vdots & \ddots & \vdots \\ ۰ & ۰ & \cdots & \underline{V}'_T \end{pmatrix} \begin{pmatrix} \underline{W}_۱ \\ \underline{W}_۲ \\ \vdots \\ \underline{W}_T \end{pmatrix}$$

$$\underline{V}'_t = (v_{t۱}, v_{t۲}, \ldots, v_{tJ}),\ \underline{W}'_t = (w_{t۱}, w_{t۲}, \ldots, w_{tJ}),\ t = ۱, ۲, \ldots, T$$

$$\underline{V}'_t\underline{W}'_t = \sum_{j=۱}^{J} v_{tj}w_{tj} = e_t, \qquad t = ۱, ۲, \ldots, T$$

$$w_{tj} > ۰, \qquad t = ۱, ۲, \ldots, T, \qquad j = ۱, ۲, \ldots, J$$

$$\sum_{j=۱}^{J} w_{tj} = ۱, \qquad t = ۱, ۲, \ldots, T$$

با فرضیات بالا می‌توان از مدل ماکسیمم آنتروپی جهت برآورد p_kها و w_tها به شکل زیر استفاده کرد.

$$MAX\ H(\underline{P}, \underline{W}) = -\sum_{k=۱}^{K}\sum_{m=۱}^{M} p_{km} \ln p_{km} - \sum_{t=۱}^{T}\sum_{j=۱}^{J} w_{tj} \ln w_{tj} \qquad (۲)$$

تحت شرایط

$$\sum_{m=۱}^{M} p_{km} = ۱, \qquad k = ۱, ۲, \ldots, K \qquad (۳)$$

$$\sum_{j=۱}^{J} w_{tj} = ۱, \qquad t = ۱, ۲, \ldots, T \qquad (۴)$$

$$\underline{Y} = X\underline{\beta} + \underline{E} = XZ\underline{P} + V\underline{W} \qquad (۵)$$

معادلهٔ رابطهٔ (۲) تابع هدف و رابطهٔ (۵) قید مدلی و روابط (۳) و (۴) قیدهای نرمال ساز هستند. در چنین ساختاری معادلهٔ هدف با مجموع آنتروپی شانون روی پارامترها و خطاها متناسب است. باید توجه داشت که برآورد $\underline{\beta}$ طوری به دست می‌آید که شامل اطلاعات داده‌ها، مدل و مجموعهٔ قیود باشد. کرانهای پایین و بالای بازه‌های شامل پارامترها

بسته به نوع مسئله می‌توانند متقارن باشند. همچنین هنگامی که توزیع خطاها نرمال باشد مقادیر ماتریس V را می‌توان از یک توزیع یکنواخت گسسته که حول صفر متقارن است، در نظر گرفت. محدودهٔ بازهٔ مربوط به خطا را می‌توان از روی مقادیر y_tها نیز تعیین کرد. پالکشیم (۱۹۹۴) نشان داد، هنگامی که مقادیر $\underline{V}_t$ها بصورت $J = \{-\alpha, ۰, \alpha\}$ یعنی ۳، در نظر گرفته شوند، آنگاه α را می‌توان به عنوان واریانس نمونه‌ای y_tها در نظر گرفت. درتعیین بازه‌های مربوط به مقادیر Z نیز بازه باید طوری انتخاب شود که مجموعه شدنی، یعنی مجموعهٔ شامل $\underline{P}$ها و $\underline{W}$هایی که در شرایط (۳) تا (۵) صادق هستند، تهی نباشد تا در داخل این مجموعهٔ شدنی بتوان آنتروپی را برحسب $\underline{P}$ و $\underline{W}$ ماکسیمم کرد. همچنین تعیین تعداد نقاط یعنی مقادیر M و J نیز مشکلاتی ایجاد می‌کنند اگر چه جودجی و گُلن پیشنهاد می‌کنند که در یک بازه مقادیر را به فاصلهٔ مساوی از یکدیگر انتخاب کنیم ولی کیتامورا و همکارانش (۱۹۹۸) روشی پیشنهاد می‌کنند که در آن توزیع احتمالها به صورت پیوسته در نظر گرفته شده‌اند.

تابع لاگرانژ را می‌توان به صورت زیر تشکیل داد.

$$L = -\sum_{k=۱}^{K}\sum_{m=۱}^{M} p_{km}\ln p_{km} - \sum_{t=۱}^{T}\sum_{j=۱}^{J} w_{tj}\ln w_{tj} - \sum_{k=۱}^{K}\alpha_k(\sum_{m=۱}^{M} p_{km} - ۱)$$
$$- \sum_{t=۱}^{T}\gamma_t(\sum_{j=۱}^{J} w_{tj} - ۱) - (\lambda_۱, \lambda_۲, \ldots, \lambda_T)(\underline{Y} - X\underline{Z}\underline{P} - V\underline{W}).$$

حال از L نسبت به p_{km} و w_{tj} مشتق گرفته، مساوی صفر قرار می‌دهیم

$$\frac{\partial L}{\partial p_{km}} = -۱ - \ln p_{km} - \alpha_k - z_{km}(\sum_{t=۱}^{T}\lambda_t x_{tk}) = ۰$$

اگر قرار دهیم $\underline{\lambda} = (\lambda_۱, \lambda_۲, \ldots, \lambda_T)$ و $\underline{X}_k = (x_{۱k}, x_{۲k}, \ldots, x_{Tk})$، از طرفی چون $\sum_{m=۱}^{M} p_{km} = ۱$، خواهیم داشت:

$$p^*_{km} = \frac{\exp(z_{km}\underline{\lambda}'\underline{X}_k)}{\sum_{m=۱}^{M}\exp(z_{km}\underline{\lambda}'\underline{X}_k)}, \qquad k = ۱, ۲, \ldots, K, m = ۱, ۲, \ldots, M$$

و همچنین

$$\frac{\partial L}{\partial w_{tj}} = -۱ - \ln w_{tj} - \gamma_t - v_{tj} = ۰$$

بنابراین

$$w_{tj} = \exp(۱ + \gamma_t + \lambda_t v_{tj})$$

از طرفی چون $\sum_{j=۱}^{J} w_{tj} = ۱$، پس

$$w^*_{tj} = \frac{exp(\lambda_t v_{tj})}{\sum_{j=۱}^{J} exp(\lambda_t v_{tj})} \qquad t = ۱, ۲, \ldots, T, \quad j = ۱, ۲, \ldots, J$$

در نتیجه مقادیر β_kها و e_tها را می‌توان به صورت زیر برآورد کرد:

$$\beta^*_k = \sum_{m=۱}^{M} z_{km} p^*_{km}, \qquad k = ۱, ۲, \ldots, K$$

$$e^*_t = \sum_{j=۱}^{J} v_{tj} w^*_{tj}, \qquad t = ۱, ۲, \ldots, T$$

البته در عمل برای برآورد پارامترهای مدل از روش تحلیلی به دلیل پیچیده بودن معادلات نمی‌توان استفاده کرد. به همین خاطر از روشهای بهینه سازی که برای کمینه و بیشینه کردن توابع هدف تحت قیود مشخص بکار می‌روند، استفاده می‌شود. بیشتر عملیات بهینه سازی در این گونه مسائل با استفاده از نرم‌افزار $GAMS$ انجام می‌شود.

۳ نتایج یک شبیه سازی

برای بررسی کارائی روش گفته شده پارامترهای مدل رگرسیونی زیر را با استفاده از روش ماکسیمم آنتروپی برآورد می‌کنیم.

فرض کنید که یک مدل رگرسیونی به صورت

$$Y_i = \beta_۰ + \beta_۱ x_{i۱} + \beta_۲ x_{i۲} + E_i \qquad (۶)$$

را می‌خواهیم به داده‌ها ببرازانیم. برای تولید داده‌ها فرض می‌کنیم

$$\underline{\beta} = (\beta_۰, \beta_۱, \beta_۲) = (۰/۲, ۰/۳, ۰/۶)$$

باشد. مقادیر E_iها را از توزیع نرمال با میانگین صفر و واریانس ۰/۰۹ تولید می‌کنیم. مقادیر $x_{i۱}$ و $x_{i۲}$ که به ترتیب، $x_{i۱}$: تعداد گرافیتهای شبه کره‌ای موجود در هر میلیمتر مربع از سطح قالب آهن و $x_{i۲}$: درصد ترکیب آهن و کربن موجود در قالب آهن است، را از کتاب باتاچاریا جلد دوم صفحهٔ ۴۳۲ در نظر می‌گیریم. حال برای تولید مقادیر y_tها، ابتدا e_t را از توزیع نرمال به طور تصادفی انتخاب کرده و از طریق معادلهٔ (۶) مقادیر y_t را به ازای $x_{t۱}$ و $x_{t۲}$ و $\underline{\beta}$ مشخص شده، به دست می‌آوریم.

ابتدا ۱۰۰ نمونه به حجم‌های ۶، ۱۰ و ۲۰ تولید کرده و برای هر نمونه مقادیر $\underline{\beta}$ را از طریق ماکسیمم آنتروپی به روش زیر برآورد می‌کنیم

$$\underline{Z}_۱ = (۰, ۰/۵, ۱)', \qquad \underline{P}_۱ = (p_{۱۱}, p_{۱۲}, p_{۱۳})', \qquad \beta_۱ = \sum_{d=۱}^{۳} z_{۱d} p^*_{۱d}$$

$$\underline{Z}_۲ = (۰, ۰/۵, ۱)', \qquad \underline{P}_۲ = (p_{۲۱}, p_{۲۲}, p_{۲۳})', \qquad \beta_۲ = \sum_{d=۱}^{۳} z_{۲d} p^*_{۲d}$$

$$\underline{Z}_۳ = (۰, ۰/۵, ۱)', \qquad \underline{P}_۳ = (p_{۳۱}, p_{۳۲}, p_{۳۳}), \qquad \beta_۳ = \sum_{d=۱}^{۳} z_{۳d} p^*_{۳d}$$

مقادیر $\underline{V}_t$ها را برای کلیهٔ tها، $t = ۱, ۲, \ldots, T$، که T حجم نمونه است، برابر با $\underline{V}_t = (-۰/۹, ۰, +۰/۹)'$ در نظر گرفته، در نتیجه $\underline{W}_t = (w_{t۱}, w_{t۲}, w_{t۳})'$ و e_t برابر است با:

$$e_t = \sum_{j=۱}^{۳} v_{tj} w^*_{tj}, \qquad t = ۱, ۲, \ldots, T$$

حال برای یک نمونه T تایی تابع هدف به صورت زیر است،

$$-\sum_{k=۱}^{۳}\sum_{d=۱}^{۳} p_{kd} \ln p_{kd} - \sum_{t=۱}^{T}\sum_{j=۱}^{۳} w_{tj} \ln w_{tj}$$

که باید تحت قیود

$$\sum_{j=۱}^{۳} w_{tj} = ۱, \qquad t = ۱, ۲, \ldots, T$$

$$\sum_{d=۱}^{۳} p_{kd} = ۱, \qquad k = ۱, ۲, ۳$$

$$\begin{pmatrix} y_۱ \\ y_۲ \\ \vdots \\ y_T \end{pmatrix} = \begin{pmatrix} ۱ & x_{۱۱} & x_{۱۲} \\ ۱ & x_{۲۱} & x_{۲۲} \\ \vdots & \vdots & \vdots \\ ۱ & x_{T۱} & x_{T۲} \end{pmatrix} \begin{pmatrix} \underline{Z}'_۱ & ۰ & ۰ \\ ۰ & \underline{Z}'_۲ & ۰ \\ ۰ & ۰ & \underline{Z}'_۳ \end{pmatrix} \begin{pmatrix} \underline{P}_۱ \\ \underline{P}_۲ \\ \underline{P}_۳ \end{pmatrix}$$

$$+ \begin{pmatrix} \underline{V}'_۱ & ۰ & \cdots & ۰ \\ ۰ & \underline{V}'_۲ & \cdots & ۰ \\ \vdots & \vdots & \ddots & \vdots \\ ۰ & ۰ & \cdots & \underline{V}'_T \end{pmatrix} \begin{pmatrix} \underline{W}_۱ \\ \underline{W}_۲ \\ \vdots \\ \underline{W}_T \end{pmatrix}$$

ماکسیمم شود، تا مقادیر بردارهای $\underline{p}_k$، $k = ۱, ۲, ۳$ و $\underline{w}_t$، $t = ۱, ۲, \ldots, T$ مشخص شوند. در نتیجه مقادیر β_k، $k = ۱, ۲, ۳$ از روابط (۷) به دست می آیند.

مقادیر میانگین ضرایب رگرسیونی $\beta_۱$، $\beta_۲$ و $\beta_۳$ روی ۱۰۰ نمونه و همچنین واریانسهای آنها طبق جدول ۱ می‌باشد.

میانگین	واریانس	$N = ۶$		$N = ۱۰$		$N = ۲۰$	
$\sum_{n=۱}^{۱۰۰} \frac{\beta^*_{۱n}}{۱۰۰}$	$Var(\beta^*_۱)$	۰/۴۲	۰/۰۰۳	۰/۳۵	۰/۰۰۹	۰/۲۸	۰/۰۱۰
$\sum_{n=۱}^{۱۰۰} \frac{\beta^*_{۲n}}{۱۰۰}$	$Var(\beta^*_۲)$	۰/۳۵	۰/۰۰۴	۰/۳۸	۰/۰۰۵	۰/۳۷	۰/۴۰
$\sum_{n=۱}^{۱۰۰} \frac{\beta^*_{۳n}}{۱۰۰}$	$Var(\beta^*_۳)$	۰/۵۲	۰/۰۰۸	۰/۴۹	۰/۰۰۹	۰/۵۴	۰/۰۱۰

جدول ۱: میانگین و واریانس ضرایب رگرسیون در ۱۰۰ تکرار

مراجع

Golan, A., Judge, G. G. and Karp, L. (1996a), A Maximum Entrop Approch to Estimation and Inference in Dynamic Models or Counting Fish in the Sea Using Maximum Entropy, *J. of Economic Dynamics and Control*, 20, 559-582.

Golan, A., Judge, G. G. and Robinson, S. (1994), Recovering Information from Incomplete or Partial Multisectoral Economic Data, *Review of Economic and Statistics*, 76, 541-549.

Judge, G. G. and Golan, A. (1993), Recovering Information in the Case of Ill-Posed Inverse Problem with Noise, *Mimeo Department of Agrichltural and Natural Resources, University of Colifonia, Berkely, Ca.*.

Kitamura, Y. and Stutzer, M. (1998), An Information-Theoretic Alternative to GME Estimation, *Econometrica*, 65, 861-874.

Pukelsheim, F. (1994), The There Sigma Rule, *Teh American Statistician*, 48, 88-91.

Shannon, C. E. (1948), A Mathematical Theory of Communication, *Bell System Technical Journal*, 27, 379-423.

آشنایی با داده‌کاوی

علی‌اصغر حائری، دکتر حمیدرضا نواب‌پور

P۱۳۰۱۵

گروه آمار، دانشگاه علامه طباطبایی

چکیده: با گسترش روز افزون فناوری رایانه‌ای و مجهز شدن سازمانها، شرکتهای دولتی و خصوصی و ... به این ابزار، امکان جمع آوری اطلاعات دقیق و جدید با قیمت ارزان برای آنها فراهم آمده، به طوری که این امکان منجر به ایجاد مجموعه‌های بسیار وسیعی از داده‌ها شده است، که شامل مقادیر زیادی از سوابق ثبت شده هستند. این سوابق مدنظر مدیران، برنامه‌ریزان، پژوهشگران و ... بوده و از آنها برای تهیهٔ گزارشهای هفتگی، ماهانه، پرسنلی، خرید، فروش، موجودی انبار و ... استفاده می‌شود. اخیراً داده‌کاوی[۱] موضوع تعداد زیادی از مقالات در تجارت و مجله‌های آماری و نرم‌افزاری شده است. ولی تا چند سال قبل فقط تعداد کمی از مردم عبارت داده‌کاوی را شنیده بودند، اگرچه داده‌کاوی تکامل یافته رشته‌ای با تاریخ قدیمی است ولی عبارت داده‌کاوی اخیراً و در دهه ۱۹۹۰ مطرح شده است. داده‌کاوی به استخراج ساختار مورد علاقه از پایگاه داده‌ها به‌ویژه انبار وسیع داده‌ها مربوط می‌شود. در این مقاله سعی شده است مفهوم کلی داده‌کاوی، ضرورت استفاده از داده‌کاوی، روشهای تحلیل داده‌ها و کاربرد آن مدنظر قرار گیرد.

واژه‌های کلیدی: داده انبار[۲]، پردازش تحلیلی آنی (OLAP)[۳]، رده‌بندی[۴]، پردازش معاملات آنی (OLTP)[۵]، پایگاه داده‌ها[۶] ساختار داده‌ها[۷].

۱ مقدمه

گسترش روزافزون فناوری رایانه‌ای امکان جمع آوری اطلاعات دقیق، جدید و لحظه به لحظه با قیمت ارزان را فراهم آورده است به طوری که این امکان منجر به افزایش انفجارآمیز منابع اطلاعاتی در دهه‌های اخیر شده و ضرورت گردآوری و پردازش منابع اطلاعاتی

[۱] Data Mining
[۲] Data Warehouse
[۳] On-Line Analytic Processing
[۴] Classification
[۵] On-Line Transactional Processing
[۶] Database
[۷] Data Structure

را آشکارتر نموده است. استخراج و کشف سریع و دقیق اطلاعات با ارزش از این منابع اطلاعاتی از جمله اموری است که هر سازمان، شرکت و کشوری به منظور توسعه علمی، فنی و اقتصادی خود به آن نیاز دارد. هم‌اکنون در هر کشوری اغلب سازمانها و شرکتها برای پرسنل، بازرگانی و حتی مراجعه کنندگان، پایگاه داده‌ها ایجاد یا خریداری نموده‌اند به طوری که برای مدیران، برنامه‌ریزان و پژوهشگران و ... جهت گرفتن تصمیمهای راهبردی، گزارشهای مختلف و توصیف وضعیت کنونی خود و ... می‌تواند حیاتی باشد. بعضی مواقع اطلاعات موردنیاز در بیش از یک پایگاه داده‌ها می‌باشد. اغلب این پایگاه‌های داده‌ها همگن نیستند؛ بنابراین به داده انبار نیاز خواهیم داشت که به تفصیل به آن خواهیم پرداخت. پس از تشکیل داده انبار برای استخراج الگوها و اطلاعات با ارزش به داده‌کاوی خواهیم پرداخت. در ادامه به مفهوم داده انبار و داده‌کاوی، روشهای داده‌کاوی، مراحل، کاربرد و ضرورت داده‌کاوی خواهیم پرداخت.

۲ مفهوم داده انبار

داده انبار یکی از ابزارهای فناوری مدیریت داده‌ها برای حمایت از داده‌کاوی است و خیلی شبیه به یک یا ترکیبی از چند پایگاه داده‌ها می‌باشد و با زمان تغییر می‌کند. اغلب پایگاه‌های داده‌های مختلف همگن نبوده و تشکیل داده‌انبار مشکل می‌باشد. به نظر اینمان داده‌انبارها موضوعات جهت داده شده‌ای هستند که تا حد زیادی به کاربرد داده‌انبار بستگی دارد[۱]. این داده انبارها ممکن است از منابع داده‌های ناهمگن و مختلف درست شوند. ایده‌ای که در پشت مفهوم داده انبار وجود دارد این است که چندین مدل پردازش برای کمک به یک درخواست اطلاعات از پایگاه‌های داده‌ها در محیطهای ناهمگن با هر فرایندی لازم است. بنابراین داده انبار به دلیل وجود پایگاه داده‌های ناهمگن ایجاد می‌شود. اطلاعات، درون این پایگاه‌های داده‌ها ترکیب شده و در یک داده انبار قرار می‌گیرند. راههای مختلفی برای ترکیب اطلاعات وجود دارد:

یک روش، تکرار ساده پایگاه داده‌ها می‌باشد. این روش هیچ مزیتی نسبت به تعیین پایگاه داده‌های ناهمگن ندارد.

دومین روش، تکرار اطلاعات می‌باشد؛ اما هر اطلاعات ناسازگار و اضافه‌ای حذف می‌شود. این روش چند مزیت دارد. مثل اهمیت به دست آوردن یک تصویر سازگار از پایگاه داده‌ها.

[۱] Thuraisingham, Bahavan (1999)

سومین روش، انتخاب زیرمجموعه‌ای از اطلاعات پایگاه داده‌ها و انتقال آن به یک انبار می‌باشد. چندین پیامد در اینجا وجود دارد. چگونه زیرمجموعه‌ها انتخاب می‌شوند؟ زیرمجموعه‌های انتخابی تصادفی هستند یا روشهایی برای انتخاب داده‌ها به کار برده می‌شود؟ (برای مثال، فرد هر سطری را در یک رابطه می‌تواند انتخاب کند و این سطرها را در انبار ذخیره کند «با فرض اینکه پایگاه داده‌ها رابطه‌ای است»).

چهارمین روش که تغییر جزئی روش سوم می‌باشد، انواع سؤالاتی را که کاربران ممکن است مطرح کنند تعیین نموده و سپس به تجزیه و تحلیل داده‌ها و صرفاً به ذخیره‌سازی داده‌هایی که موردنیاز کاربر است می‌پردازد. این روش «پردازش تحلیلی آنی» (OLAP) نامیده می‌شود که در مقابل «پردازش معاملات آنی» (OLTP) که عمل تولید داده را انجام می‌دهد، قرار دارد.

داده‌ها اغلب ممکن است به صورت متفاوت و برای کاربردهای مختلفی منظور شوند که آنها را داده‌های چندبُعدی گویند. بنابراین می‌توان گفت «داده انبار» سیستمی برای ذخیره و ارائهٔ توده‌ای از مقادیر داده‌ها است. روش‌شناختی ویژه‌ای که برای داده انبار طرح شده است باید برای استفاده در کشور بررسی شود. اگرچه تشابهاتی وجود دارد لیکن برخی از امور در روش‌شناختی داده انبار با گسترش کاربرد امور مرسوم مغایرت دارد.

روش‌شناختی داده انبار در اموری که در پروژه‌های داده انبار گنجانده شده یکپارچگی ایجاد می‌کند. کشف دانش در پایگاه داده‌ها[۱] (KDD) در جهت کشف اطلاعات مفید از مجموعهٔ بزرگ داده‌ها می‌باشد. دانش کشف شده می‌تواند قاعده‌ای باشد تا ویژگیهای داده‌ها، الگوهایی که به طور متناسب رخ می‌دهند، خوشه‌بندی موضوعات در درون پایگاه داده‌ها و غیره را توصیف کند.

یک کاربر سیستم KDD بایستی درک بالایی از قلمرو داده‌ها به منظور انتخاب زیرمجموعهٔ صحیحی از داده‌ها، رده مناسبی از الگوها و معیار خوبی برای الگوهای جالب داشته باشد. بنابراین سیستم KDD باید ابزارهایی با اثر تعاملی داشته باشد نه سیستمهای تجزیه و تحلیل خودکار. لذا کشف دانش از پایگاه داده‌ها باید مثل یک فرایند شامل گامهای زیر باشد:

۱) درک قلمرو
۲) آماده کردن مجموعه داده‌ها
۳) کشف الگوها (داده‌کاوی)
۴) پردازش بعد از کشف الگو

[۱] Knowledge Discovery of Database (KDD)

۵) استفاده از نتایج

۳ مفهوم داده‌کاوی

عبارت داده‌کاوی مترادف با یکی از عبارتهای استخراج دانش، برداشت اطلاعات، بررسی داده‌ها و حتی لایروبی کردن داده‌ها[۱] است که در حقیقت کشف دانش در پایگاه داده‌ها را توصیف می‌کند. بنابراین این ایده که مبنای داده‌کاوی است، یک فرایند با اهمیتی از شناخت الگوهای بالقوه مفید، تازه و در نهایت قابل درک در داده‌ها می‌باشد. واژه کشف دانش در پایگاه داده‌ها در سال ۱۹۸۹ در مراجعه به مفهوم کلی، گسترده و سطح بالا و به دنبال جستجوی دانش در اطلاعات شکل گرفته است. عبارت داده‌کاوی فنون کاربردی ابزارهای به‌کار برده شده برای بیان و تجزیه و تحلیل داده‌ها و برای نشانه‌های داوری می‌باشد. این اصطلاح (داده‌کاوی) را آمارشناسان، تحلیلگران داده‌ها و انجمن سیستمهای اطلاعات مدیریت به کار برده‌اند. در حالیکه محققین یادگیری ماشین و هوش مصنوعی از KDD ها بیشتر استفاده می‌کنند. **بنابراین داده‌کاوی چیست؟** در زیر چند تعریف ارائه می‌شود.

«داده‌کاوی در حقیقت کشف ساختارهای جالب توجه، غیرمنتظره و با ارزش از روی مجموعهٔ وسیعی از داده‌ها می‌باشد و فعالیتی است که اساساً با آمار و تحلیل دقیق داده‌ها منطبق است.» David Hand (1999)

«داده‌کاوی فرایند کشف رابطه‌ها، الگوها و روندهای جدید معنی‌داری است که به بررسی حجم وسیعی از اطلاعات ذخیره شده در انبارها با فناوریهای تشخیص الگو (مانند ریاضی و آمار) می‌پردازد.» Gartner Group

«داده‌کاوی یا به تعبیر دیگر کشف دانش در پایگاه داده‌ها، استخراج غیر بدیهی اطلاعات بالقوه مفید از روی داده‌هاست که قبلاً ناشناخته مانده‌اند. این مطلب برخی از روشهای فنی مانند خوشه‌بندی، خلاصه‌سازی داده‌ها، فراگیری قاعده‌های رده‌بندی، یافتن ارتباط شبکه‌ها، تحلیل تغییرات و کشف بی‌قاعدگی‌ها را شامل می‌شود.» William. J. Frawly, Gregory Piatetsky, Shapiro and Christopher. J. Matheus

بنابراین معمولی‌ترین فنون به‌کار رفته در داده‌کاوی عبارتند از:

۱) درختهای تصمیم[۲]

---

[۱] data dredging
[۲] Decision Trees

۲) شبکه‌های عصبی[۳]
۳) قاعده‌های استقراء[۱]
۴) نزدیکترین همسایگی[۲]
۵) الگوریتم‌های ژنتیک[۳]
و ... .

۴ اختلافها و تشابهات داده‌کاوی با آمار

داده‌کاوی که کشف دانش در پایگاه داده‌ها (KDD) نیز نامیده می‌شود، همپوششی زیادی با آمار دارد به طوری که ممکن است برخی از مردم KDD را زیرمجموعه‌ای از آمار تلقی کنند. این رویکرد برای برخی از آمارشناسان محافظه‌کار مشترک است ولی واقع‌گرایانه نیست. داده‌کاوی از ایده‌ها، ابزارها و روشهایی از دیگر سطوح به ویژه یادگیری ماشین، نظریه و فناوری پایگاه داده‌ها استفاده می‌کند و ضرورتی ندارد که به برخی از سطوحی که مدنظر آمار است مربوط شود. در حالت کلی، داده‌کاوی با مجموعه‌های داده‌های بزرگ سروکار دارد. برای آمارشناسان مجموعه داده‌ها با ۱۰۰۰ رکورد ممکن است خیلی بزرگ تلقی شود اما به نظر می‌رسد که در مقایسه با پایگاه داده‌های با یک میلیون بایت کاملاً کوچک باشد. واضح است که برای بررسی مقادیر داده‌های این چنینی شخص بایستی از رایانه استفاده کند، چرا که داده‌کاوی پیش‌زمینهٔ مهمی در علوم رایانه‌ای به ویژه در نظریهٔ پایگاه داده‌ها دارد.

با این حال عبارت داده‌کاوی برای فرد آماری جدید نیست و اغلب در یک مفهوم جزئی مانند فرایند تولید مدلهای زیادی به منظور به دست آوردن برازش خوب به کار برده شده است.

در آمار تمرکز روی استنباط است و وظیفهٔ اصلی آمار وقتی نمونه‌ای از یک جامعه مشاهده می‌شود استنباط دربارهٔ جامعه است و فرضیات آزمون می‌شوند ولی در داده‌کاوی با کل داده‌ها (جامعه) سروکار داریم. فرضیات از روی داده‌ها ساخته می‌شوند در چنین مواردی به آزمون معنی‌داری احتیاجی نیست و مقدار مشاهده شده آماره همان مقدار پارامتر است.

در آمار مدل سیمای دیگری داشته و نقش مهمی را ایفا می کند، هر چند که ممکن

[۳] Neural Networks
[۱] Induction Rules
[۲] Nearest Neighbur
[۳] Genetic Algorithms

است معانی مختلفی داشته باشد، به طوری که همه مکانیسم‌ها در آمار به مدلسازی ختم می‌شوند. محاسبات، معیار انتخاب و نمونه‌گیری و غیره جزئیات کمکی برای ساختن یک مدل خوب هستند. ضرورتی ندارد که مدل برای داده‌کاوی درست باشد از این رو الگوریتم‌ها برای جستجوی الگو خیلی مهم هستند هر چند بعضی از این الگوریتم‌ها در آمار خلق شده‌اند ولی برای مجموعهٔ داده‌های بزرگ به تدبیر جدیدتر، سریعتر، مؤثرتر و همچنین وجود KDD احتیاج است.

داده‌کاوی کشف اطلاعات با ارزش ولی ناشناخته است و فرایند آن به طور ضروری برخلاف تجزیه و تحلیل آماری، صرفاً اکتشافی است.

۵ مراحل داده‌کاوی

در اینجا فرایندی برای استخراج دانش نهفته از داده انبار، فایل اطلاعات مشتری یا هرپایگاه داده‌های شرکتهای دیگر وجود دارد.

۱) تعیین موضوع: قبل از شروع باید روشن کرد که با تجزیه و تحلیل خود انتظار داریم به چه چیزی دست یابیم، از قبل بایستی هدف از داده‌کاوی را تعیین کرده و برای خود ثابت کنیم که این هدف قابل اندازه‌گیری است یا نه؟ برخی از اهداف ممکن عبارتند از:
- یافتن رابطه فروش بین محصولات ویژه یا خدمات
- تعیین الگوهای خرید خاص و فوق العاده
- تعیین انواع پتانسیلهای مشتریها
- یافتن روند فروش محصولات

۲) انتخاب داده‌ها: در مرحله قبل هدف تعیین گردید. گام بعدی، انتخاب داده برای رسیدن به این هدف می باشد. این امر می تواند زیر مجموعه‌ای از داده‌انباری که در اختیار داریم یا از فروشگاه داده‌ها[۱] که حاوی اطلاعات خاص یا فایل اطلاعات مشتری است، باشد. برای این کار تا حد امکان گستره داده‌هایی که بایستی کاویده شوند را بخش‌بندی کنیم.
چند نتیجه کلیدی وجود دارد.
- آیا داده‌ها مناسب برای توصیف پدیده‌هایی هستند که تحلیل داده‌کاوی تلاش

[۱] data mart

می‌کند برای آنها مدلسازی کند؟
— آیا می‌توانیم سوابق مشتری داخلی را با شیوه زندگی بیرونی داده‌های جمعیتی مطابقت دهیم؟
— آیا داده‌ها ثبات دارند— آیا بعد از تحلیل ویژگیها ثابت می‌ماند؟
— اگر پایگاه داده‌ها را ادغام نماییم آیا می‌توانیم زمینه مشترکی برای پیوند دادن آنها بیابیم؟
— داده‌ها چه رابطه‌ای با هدف کاری دارند؟

۳) آماده کردن داده‌ها: در مرحله قبل داده‌ها گردآوری شد. حال باید تصمیم بگیریم کدامیک از ویژگیهای موجود در قالبها قابل استفاده هستند. سهم متخصصین، تولیدکنندگان و کاربران را از قلمرو داده‌ها ملاحظه کنیم.
— اتخاذ تدابیری برای بررسی داده‌های گمشده، نوفه خارجی و داده‌های دورافتاده
— تعیین متغیرهای اضافی در مجموعه داده‌ها و تصمیم‌گیری درباره اینکه کدام فیلد باید خارج شود
— در صورت لزوم تصمیم‌گیری درباره تبدیل لگاریتمی، توان دوم، یا ...
— بازرسی شهودی مجموعه داده‌ها برای درک پایگاه داده‌ها
— تعیین توزیع فراوانی داده‌ها

می‌توانیم برخی از این تصمیم‌ها را تا انتخاب ابزار داده‌کاوی به تعویق بیندازیم. مثلاً اگر ما به یک شبکه عصبی یا شبکه چندبُعدی نیاز داشته باشیم ممکن است مجبور شویم برخی از فیلدها را تغییر دهیم.

۴) بازرسی داده‌ها: ارزیابی ساختار داده‌ها به منظور تعیین ابزارهای مناسب
— نسبت صفات رسته‌ای یا دو حالتی در پایگاه داده‌ها چیست؟
— طبیعت و ساختار پایگاه داده‌ها چیست؟
— توزیع مجموعه داده‌ها چیست؟

بایستی ارزیابی عینی ساختار داده‌ها را در مقابل نیاز کاربران برای درک یافته‌ها توازن بخشیم. برای مثال شبکه‌های عصبی نتایج را نمی‌توانند توصیف کنند.

۵) انتخاب ابزارها: اهداف کاری و ساختار داده‌ها دو رهنمون مهم برای انتخاب ابزار داده‌کاوی مناسب می‌باشند. هر دو بایستی ما را به ابزار یکسان هدایت کنند. موقعی که یک مجموعه از ابزارهای بالقوه را ارزیابی می‌کنیم، این پرسش‌ها را ملاحظه کنیم.
— آیا مجموعه داده‌ها صرفاً رسته‌ای است؟
— چه برنامه‌ای برای حمایت از ابزارهای منتخب خود داریم؟
— آیا ابزارهای انتخابی سازگار با ODBC[۱] یک رابط برنامه‌ای از شرکت مایکروسافت

[۱] Open Database Connectivity (ODBC)

است که ارتباط بین برنامه‌های کاربردی ویندوز را برای دسترسی به پایگاههای داده‌های روی شبکه فراهم می‌سازد. هستند؟

— چه قالبی از داده‌ها می‌تواند برای ابزارها مفهوم داشته باشد؟

ابزارهای متعددی وجود دارد که به طور مشابه برای پروژه داده‌کاوی تهیه می‌کنند. تعدادی از ابزارها چندین فناوری را در یک مجموعه از برنامه‌های تحلیل آماری، شبکه عصبی و رده‌بندی کننده نمادین جمع می‌کنند.

۶) قالب پاسخ: به همراه بازرسی داده‌ها، هدف کاری و انتخاب ابزار، قالب پاسخ را تعیین می‌کنند. سوالات کلیدی عبارتند از:

— قالب بهینه حل درخت تصمیم— کد C، دستور SQL— چیست؟

— قالب مناسب گزینه چیست؟

— هدف از حل چیست؟

— کاربر به چه چیزی احتیاج دارد— نمودارها، گزارشها، برنامه—؟

۷) طرح‌ریزی مدل: در این هنگام است که فرایند داده‌کاوی آغاز می‌شود. معمولاً اولین گام، استفاده از چند رده تصادفی برای جداکردن داده‌ها در مجموعه تولیدی، مجموعه آزمون و ایجاد و ارزیابی یک مدل می‌باشد. تولید قاعده‌های رده‌بندی، درختهای تصمیم، خوشه‌بندی زیرگروه‌ها، امتیازها، برنامه‌ها، وزن‌ها و ارزیابی داده‌ها یا نسبت خطاها در این مرحله قرار دارند.

تفکیک کردن این پیامدها:

— آیا نسبت خطاهادر سطح قابل قبولی است؟ آیا می‌توانیم آنها را بهبود دهیم؟

— چه صفات خارجی را یافته‌ایم؟ آیا می‌توانیم آنها را حذف کنیم؟

— آیا داده‌های اضافی یا روش‌شناختی مختلفی لازم است؟

— آیا مجموعه داده‌های جدیدی تولید و آزمون خواهیم کرد؟

۸) ارزیابی یافته‌ها: نتایج تجزیه و تحلیل را با مشتری کاری یا متخصص قلمرو به بحث و تبادل نظر گذاشته و مطمئن می‌شویم که یافته‌هابرای اهداف کاری صحیح و مناسب هستند.

— آیا یافته‌هاحساسیت‌ساز هستند؟

— آیا برای هر گام قبلی و بهبود نتایج برمی‌گردیم؟

— آیا می‌توانیم ابزارهای دیگر داده‌کاوی را برای تکرار یافته‌ها به‌کار ببریم؟

۹) ارایه یافته‌ها: یک گزارش نهایی برای واحد کاری یا مشتری تهیه می‌کنیم . این گزارش باید مستند به فرایند داده‌کاوی بی‌عیب و نقص شامل آماده‌سازی داده‌ها، ابزارهای به‌کار رفته، نتایج آزمون، کد منابع و قاعده‌ها باشد.

بعضی از این نتایج عبارتند از:

— داده‌های اضافی تجزیه و تحلیل را بهبود خواهند داد؟
— چه بینش اساسی‌ای پوشش دهیم و چگونه قابل کاربرد است؟
— از تجزیه و تحلیل داده‌کاوی چه نتایجی می‌تواند پیشنهاد شود؟
— آیا یافته‌ها مناسب هدف کاری هستند؟

۱۰) هماهنگ کردن پاسخ‌ها: یافته‌ها را با کلیه علایق کاربران در واحدهای کاری مناسب تقسیم می‌کنیم. ما می‌توانیم به‌طور کامل نتایج تجزیه و تحلیل در روشهای کار شرکت را به هم پیوند دهیم، بعضی از حلهای داده‌کاوی ممکن است موارد زیر را در برگیرد.
—دستورهای SQL برای توزیع مصرف کنندگان نهایی
— تولید سیستم با همکاری کدهای C
— تجمیع قاعده‌ها در نظام تصمیم‌گیری

اگر چه ابزارهای داده‌کاوی، پایگاه داده‌ها را به صورت خودکار تجزیه و تحلیل می‌کنند، ولی اگر مراقب نباشیم ممکن است به یافته‌های نادرست و نتایج غلط منجر شود. در نظر داشته باشید که داده‌کاوی یک فرایند کاری با هدف مشخص است، که برای استخراج دیدگاههای رقابتی از سوابق تاریخی در پایگاه داده‌ها می‌باشد.

٦ روشهای داده‌کاوی

الف) گزارش‌دهی و پردازش تحلیل آنی

موقعی که یک تحلیلگر در زمان تجزیه و تحلیل با OLAP کار می‌کند در مورد چیزهایی که به دنبال آن است پیش‌زمینه‌ای دارد و با فرضیاتی که مورد بررسی قرار گرفته شروع می‌کند . در حالیکه در مورد داده‌کاوی، تحلیلگر هیچ شناخت قبلی در رابطه با نتایج احتمالی ندارد. کاربران از طریق OLAP به دنبال پاسخ به پرسشهایشان هستند. هر جستجو برای دست‌یابی به پرسش خاص ما را با مرحلهٔ پیچیده‌تری روبرو می‌کند، و کاربر به دانش قبلی از نتایج مورد انتظار نیاز دارد. فرایند در داده‌کاوی کاملاً متفاوت است. در حالیکه OLAP برای تجزیه و تحلیل گذشته و کسب بینش‌های تازه به کاربر کمک می‌کند، داده‌کاوی برای پیش‌گویی آینده به کاربر کمک می‌کند.

ب) مدلسازی نظریه رهنمون

همبستگی
آزمون t
تحلیل واریانس

رگرسیون خطی
رگرسیون لوژستیک
تحلیل تشخیصی
مدلهای پیش‌بینی

ج) مدلسازی داده رهنمون

تحلیل خوشه‌ای
درختهای تصمیم[۱]
تصویر کردن داده‌ها[۲]
شبکه‌های عصبی
قاعده‌های پیوند[۳]
قاعده استقراء[۴]

۷ کاربردهای داده‌کاوی

داده‌کاوی در رشته‌های مختلفی کاربرد دارد که در زیر به برخی از این موارد اشاره می‌کنیم:

۱) بازاریابی/خرده‌فروشی
- تشخیص الگوهای خرید مشتریها
- یافتن پیوند میان مشخصه‌های جمعیت‌شناختی مشتری
- پیش‌گویی پاسخ برای (پست کردن) اوراق مبارزهٔ انتخاباتی
- تجزیه و تحلیل سبد بازار

۲) بانکداری
- کشف الگوهای استفاده از کارتهای اعتباری قلابی
- تشخیص مشتریهای ثابت (وظیفه‌شناس)
- پیش‌گویی احتمالی مشتریها برای تغییر کارتهای اعتباری پیوسته آنها
- تعیین پرداخت کارتهای اعتباری برای گروههای مشتریان
- یافتن ارتباط مخفی بین نشانگرهای مالی مختلف
- تشخیص قواعد بازرگانی موجود از داده‌های بازاریابی

[۱] Decision Trees
[۲] Data Visualization
[۳] Association Rules
[۴] Rule Induction

۳) بیمه و مراقبت بهداشتی
- تجزیه و تحلیل ادعاها ـ یعنی اینکه کدامیک از روشهای پزشکی با هم ادعا شده‌اند
- پیش‌گویی اینکه مشتریها بیمه‌نامهٔ جدید بخرند
- تشخیص الگوهای رفتاری مشتریان خطری
- تشخیص رفتارهای قلابی

۴) حمل و نقل (انتقال)
- تعیین توزیع فهرست مقدار فروش
- تجزیه و تحلیل الگوهای بارگیری

۵) پزشکی
- تعیین الگوی رفتار بیماران برای پیش‌بینی مراجعات به مطب (ویزیتهایی که در مطب صورت می‌گیرد)
- تشخیص معالجات موفق پزشکی برای بیماریهای مختلف

۶) متن‌کاوی[۱]

۷) شبکه‌کاوی[۲]

۸) صوت‌کاوی[۳]

۹) تصویرکاوی[۴]

۱۰) سیستم هواشناسی

۱۱) سیستم ثبت احوال کشور

۱۲) صنعت

مراجع

Thuraisingham, Bahavan (1999), *Data Mining Technologies, Techniques, Tools, and Trend*, CRC press LLC, Florida.

Michael. J. A. Berry Gordon Linoff (1997), *Data Mining Techniques*, John Wiley, New York.

[۱] Text Mining
[۲] Web Mining
[۳] Audio Mining
[۴] Video Mining

David Hand (1999), *Why data mining is more than statistics writ large*, Imprial College of Science, Technology, and Medicine, Department of Mathematics.

Gartner Group, *http://www.spss.com*

Han Jiawei, Kamber Micheline (2001), *Data Mining concepts and techniques*, Morgan Kaufmann.

Heikki Mannila (1996), *Data Mining: machine learning, statistics, and databases*, 8th International Conference on Scientifice and Statistical Database Management, Stockholm, June 18-20, 1996, pp 1-8.

Paulraj Ponniah(2001), *Data Warehousing Fundamentals*, John Wiley, New York.

Rud, Olivia Parr (2001), *Data Mining Cookbook*,John Wiley, New York.

Data Mining vs. Statistics (1997), *http://www.pmsi.fr/dminita.htm*, Article and figure originally published in 1997 in Science Tribune.

William. J. Frawly, Gregory Piatetsky, Shapiro and Christopher. J. Matheus, *http://pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_6.html*

http://www.sas.com

http://www.kdnuggets.com/fad/data-mining.html

کاهش اریبی انتخاب در آزمایشهای دنباله ای

آرزو حبیبی راد

P۱۲۰۳۱

گروه آمار، دانشگاه فردوسی مشهد

چکیده برای مقایسه اثرات دو یا چند تیمار در یک طرح تحقیقی، در بسیاری از موارد این مشکل وجود دارد که به افراد (واحدهایی) که آزمایشها روی آنها باید انجام شود، در یک زمان دسترسی نداریم بلکه به این افراد (واحدها) در طول زمان دسترسی پیدا می‌کنیم، به طوری که پس از دسترسی به هر واحد لازم است تیمار مناسب بلافاصله به آنها تخصیص داده شود. حال اگر آزمایشگر آگاه باشد و یا حدس بزند که چه تیماری توسط آن آماردان به واحد تخصیص داده می‌شود، می‌تواند آگاهانه یا ناآگاهانه بااعمال نفوذ درانتخاب واحدها برای آزمایش، باعث ایجاد اریبی در نتایج آزمایش شود، که اریبی حاصل از این اعمال نفوذ را اریبی انتخاب می‌نامیم. از این رو دراین مقاله به دنبال ارائه طرحهایی هستیم که در آن با توجه به استراتژی‌های استفاده شده توسط آزمایشگر متوسط تعداد حدسهای درست او (آزمایشگر) را تا حدامکان کاهش دهیم.

واژه‌های کلیدی: اریبی انتخاب، استراتژی همگرا، استراتژی احتمال متناسب، طرح سکه اریب Efron، طرح کاسه‌ای، طرح سکه اریب تطبیقی.

۱ مقدمه

فرض کنید آزمایشگر (E) علاقمند به مقایسه اثرات دوتیمار A و B بر روی یک موضوعی در جامعه باشد، اما ممکن است این مشکل وجود داشته باشد که وی به واحدها (افراد) در یک زمان دسترسی نداشته باشد بلکه به این واحدها (افراد) در طول زمان دسترسی پیدا کند بطوریکه پس از دسترسی ،آزمایشگر (E) باید تصمیم بگیرد که آیا این واحدها برای آزمایش مناسب است یا خیر ،و در صورت مثبت بودن پاسخ وی ،آنگاه آمارداران (S)تصمیم می گیرد که تیمار A یا تیمار B را به واحد مورد نظر انتساب دهد و این روند تا انتخاب نمونه ای به حجم $2n$ ادامه پیدا می کند .
اما اگر آزمایشگر (E) تمایل به برتری تیمار A نسبت به B داشته باشد می تواند آگاهانه یا نا آگاهانه براساس حدس خود واحدهائی را برای تیمار A انتخاب کند که متوسط پاسخ مورد انتظار آن $\mu+\Delta$ (Δ اریبی در هر آزمایش می باشد) و برعکس اگر حدس بزند که

تیمار B توسط آماردان انتخاب می شود واحدهائی را انتخاب کند که متوسط پاسخ مورد انتظار آنها $\mu - \Delta$ است .

تعریف ۱: روشی را که‌طبق آن در هر مرحله آزمایشگر حدس می زند که آماردان چه تیماری را به مریض بعدی اختصاص می دهد استراتژی آزمایشگر می نامیم .

تعریف ۲:روشی را که در آن آماردان ،برای انتخاب تیمارها اتخاذ می کند طرح آماردان نامیده می شود .

تعریف ۳:کمیت $۲\Delta(E(G) - n)$،اریبی انتخاب نامیده می شود که در آن G تعداد حدسهای درست آزمایشگر است .

۲ طرحهای آماردان (S)

۱.۲ طرح سکه اریب (EFRON)

فرض کنیدrآزمایش از $۲n$(معلوم)آزمایش انجام شده ،برای آزمایش $(r + ۱)$ ام اگر تیمار A ، D_r تا با گروه تیمار B اختلاف داشته باشد ،انتخاب تیمار برای آزمایش $(r + ۱)$ ام توسط آماردان به قرار زیر صورت می گیرد

۱— اگر $D_r > ۰$ باشد ،با احتمال q تیمار A و با احتمال p تیمار B را انتخاب می کند .

۲— اگر $D_r = ۰$باشد ،با احتمال $\frac{۱}{۲}$ تیمار A و با احتمال $\frac{۱}{۲}$ تیمارB را انتخاب می کند .

۳— اگر$D_r < ۰$ باشد ،با احتمال p تیمار A و با احتمال q تیمار B را انتخاب می کند .

بطوریکه همواره $p + q = ۱, p \geq q$.

طرح فوق را طرح سکه اریب با اریبی p می نامیم و با علامت اختصاری $BCD(p)$ نمایش می دهیم .

در حالتیکه $p = \frac{۲}{۳}$ فرض شود اریبی انتخاب و متعادل بودن طرح(برابری حجم نمونه برای دو تیمار A, B) نسبت به سایر حالتها در وضعیت مطلوبتری قرار دارد . از این رو در این مقاله مقدار $p = \frac{۲}{۳}$ فرض می شود .

۲.۲ طرح کاسه ای (The Urn Design)

ظرفی را در نظر بگیرید که از ابتدا α توپ قرمز و α توپ سفید را در خود جای داده ،یک توپ به طور کاملاً تصادفی از ظرف بیرون کشیده می شود ،اگر توپ سفید باشد ،آماردان تیمار A را برای مریض بعدی انتخاب می کند و اگر قرمز بود تیمار B را انتخاب می کند .و هر توپ که بیرون آورده شود نه تنها به ظرف بر می گردد بلکه β توپ از رنگ مخالف آن نیز به ظرف برگردانده می شود ،این طرح به طرح کاسه ای معروف است و با علامت اختصاری $UD(\alpha,\beta)$ نمایش داده می شود .در حالت خاص $\alpha = ۰, \beta > ۰$ با احتمال $\frac{۱}{۲}$ یکی از دو تیمار را انتخاب می کند و β توپ رنگ مخالف را به ظرف اضافه می کند و فرایند انتخاب تیمارهارا مشابه بالا ادامه می دهد.

اما دیده می شود که $UD(۰,\beta)$ دارای خواص جالبی است از جمله اینکه $UD(۰,\beta)$ از مقدار β مستقل است و دیگراینکه به راحتی قابل اجرا است زیرا اگر تا مرحله r ام ،مریض به تیمار A و j مریض به تیمار B نسبت داده باشیم آنگاه تیمار $A(B)$ را با احتمال $\frac{j}{i+j}$ (یا $\frac{i}{i+j}$) به مریض $r+۱$ ام نسبت می دهیم .

۳.۲ طرح سکه اریب تطبیقی (The AdaPtive Biased Coin Design)

فرض کنید که در مرحله rام $(r = ۰, ۱, ۲, \ldots, ۲n)$یک آزمایش دنباله‌ای قرار داریم، بطوریکه تا این مرحله i واحد به تیمار A و j واحد را به تیمار B نسبت داده باشیم .

در این طرح آماردان برای انجام مرحله $(r+۱)$ ام آزمایش با احتمال $p = p(\frac{D_r}{r})$ تیمار A و با احتمال $q = q(\frac{D_r}{r})$ تیمار B را انتخاب می کند .(که $D_r = i - j$ است و p یک تابع غیر صعودی از $(\frac{D_r}{r})$ است بطوریکه مقادیر p در بازه $[۰, ۱]$ تغییر می کند)از طرفی $p + q = ۱$ فرض شده است و همچنین $p(x) = q(-x)$ برای هر $x \in [-۱, ۱]$. توجه کنید که طرح سکه اریب $BCD(p)$ حالت خاصی از طرح فوق است اگر $p(x)$ مقدار ثابت (p) بزرگتر از $\frac{۱}{۲}$ فرض شود .

و اگر $p(x) = \frac{۱-x}{۲}$ فرض شود طرح فوق معادل طرح $UD(۰,\beta)$ است . Wei(۱۹۷۷a)

۳ استراتژیهای آزمایشگر (E)

۱.۳ استراتژی همگرا (Convergent Strategy)

متداولترین استراتژی برای آزمایشگر در قبال طرحهای ارائه شده توسط آماردان استراتژی همگرا $(\theta_۱)$ می باشد .بطوریکه تحت استراتژی همگرا $(\theta_۱)$ آزمایشگر تیماری را حدس

می زند که تا آن مرحله کمتر انتخاب شده است و در نتیجه واحدی (فردی)را انتخاب می کند تا به تیمار مورد نظر وی پاسخ مطلوبتری را بدهد در صورتیکه تعداد تیمارهای انتخاب شده تا آن مرحله برابرباشند بااحتمال $\frac{1}{2}$ حدس می زند که چه تیماری انتخاب شود .

۲.۳ استراتژی احتمال متناسب (Probability Convergent Strategy)

تحت استراتژی احتمال متناسب (θ_2) آزمایشگر (E) در هر مرحله با احتمال γ تیمار A و با احتمال $1-\gamma$ تیمار B را حدس می زند ،سپس متناسب با حدس خود واحدها (افراد)را انتخاب می کند (یعنی اگر نتیجه حدس وی تیمار A باشد واحد(فردی) با متوسط پاسخ $\mu+\Delta$ و اگر نتیجه حدس وی تیمار B باشد واحد (فردی)با متوسط پاسخ $\mu-\Delta$ انتخاب می کند). مقدار $\gamma = \frac{1}{2}-\frac{i-j}{2}$ فرض می شود (i,j به ترتیب تعداد تیمارهای A,B هستند تا مرحله r ام)

۴ خاصیت مجانبی تعادلی طرح

تعریف ۴:طرحی را به طور مجانبی متعادل گوییم که درآن $\frac{D_{2n}}{n}$ در احتمال به صفر همگرا باشد وقتی که $n \to \infty$.(که $D_{2n} = N_A - N_B$ و N_B, N_A به ترتیب تعداد تیمارهای B, A در پایان آزمایش هستند)
Efron(۱۹۷۱)و Wei(۱۹۷۷b)نشان دادند که دو طرح $BCD(\frac{2}{3})$ و طرح $\cup D(0,\beta)$ (که هر دو حالت خاصی از طرح سکه اریب تطبیقی هستند)به طور مجانبی متعادل هستند.

۵ اریبی انتخاب طرح آماردان (S)

۱.۵ تحت استراتژی همگرا (θ_1)

اریبی انتخاب برای دو طرح $\cup D(\alpha,\beta)$ و $BCD(p)$ تحت استراتژی (θ_1) توسط Wei (۱۹۷۷a) محاسبه شده است او همچنین نشان داده که به ازاء $n > 4$، طرح $\cup D_{\theta_1}(0,\beta)$ اریبی انتخاب کمتری از طرح $BCD_{\theta_1}(\frac{2}{3})$ دارد (شکل ۱).

۲.۵ تحت استراتژی احتمال متناسب (θ_2)

۱.۲.۵ اریبی انتخاب طرح سکه اریب EFRON تحت استراتژی احتمال متناسب (θ_2)

می دانیم اریبی انتخاب برابر است با $\Delta(E(G)-n)$۲، پس $E(G)$ معیار خوبی برای اندازه اریبی انتخاب است. برای محاسبه $E(G)$ فرض کنید در مرحله rام اختلاف بین دوتیمار d باشد $(D_r = d)$،اگر G_r پیشامد حدس درست آزمایشگر برای مرحله $(r+۱)$ ام باشد و $X_{r+۱} = A$ پیشامد انتخاب تیمار A در مرحله $(r+۱)$ ام و $X_{r+۱} = B$ پیشامد انتخاب تیمار B در مرحله $(r+۱)$ باشد داریم ،

$$P(G_r = ۱|D_r = d) = P(G_r = ۱|D_r = d, X_{r+۱} = A)P(X_{r+۱} = A)$$

$$+P(G_r = ۱|D_r = d, X_{r+۱} = B)P(X_{r+۱} = B)$$

تحت استراتژی $\theta_۲$ ،احتمال آنکه آزمایشگر درست حدس بزند،به شرط آنکه تیمار $A)B)$ انتخاب شده برابرγ $(\gamma - ۱)$ است ،پس می توان نوشت ،

$$P(G_r = ۱|D_r = d) = \gamma P(X_{r+۱} = A) + (۱ - \gamma)P(X_{r+۱} = B)$$

حال اگر آماردان از طرح سکه اریب EFRON استفاده کند ،احتمال انتخاب تیمار $A)B)$ در مرحله $(r+۱)$ ام بستگی به $(i - j = d)d$ دارد ،به طوریکه اگر $d > ۰$ باشد این احتمال برابر $(q)p$ است .واگر $d < ۰$ باشد ، با احتمال$q(p)$ واگر$d = ۰$ باشد با احتمال $\frac{۱}{۲}(\frac{۱}{۲})$ تیمار $A)B)$ را انتخاب می کند . پس می توان نوشت ،

$$P(G_r = ۱) = \sum_{d=-r}^{r} P(G_r|D_r = d)P(D_r = d)$$

در نتیجه ،

$$E(G) = \sum_{r=۰}^{۲n-۱} P(G_r)$$

$$= \frac{۱}{۲} + \sum_{r=۱}^{۲n-۱} \sum_{d=-r}^{-۱} (p\gamma + (۱ - \gamma)q)P(D_r = d) + \frac{۱}{۲} \sum_{r=۱}^{۲n-۱} P(D_r = d)$$

$$+ \sum_{r=۱}^{۲n-۱} \sum_{d=۱}^{r} (q\gamma + (۱ - \gamma)p)P(D_r = d)$$

بنا به فرض داریم $\gamma = \frac{۱}{۲} - \frac{d}{۲r}$ و $۱ - \gamma = \frac{۱}{۲} + \frac{d}{۲r}$ پس

$$\Rightarrow E(G) = \frac{۱}{۲} + \sum_{r=۱}^{۲n-۱} \sum_{d=-r}^{r} (\frac{۱}{۲} - \frac{|d|}{۲r}(q-p))P(D_r = d)$$

$$\frac{۱}{۲} + \sum_{r=۱}^{۲n-۱} (\frac{۱}{۲} - \frac{(q-p)E|D_r|}{۲r})$$

اگر در طرح EFRON $p = q = \frac{۱}{۴}$(طرح کاملاً تصادفی)فرض شود ،داریم

$$E(G) = \frac{۱}{۲} + \sum_{r=۱}^{۲n-۱} \frac{۱}{۲} = \frac{۱}{۲} + \frac{۲n-۱}{۲} = n$$

پس اریبی انتخاب طرح فوق ،صفر است .ولی می دانیم که این طرح در حجم نمونه ای کم نامتعادل است .
در حالتیکه $p = \frac{۲}{۳}$ است ،اریبی انتخاب و متعادل بودن طرح ،نسبت به بقیهٔ حالتها در وضعیت مطلوبتری قرار دارد . در نتیجه تحت استراتژی احتمال متناسب ،مقدار $E(G)$ را برای حالت $p = \frac{۲}{۳}$ محاسبه می کنیم پس با فرض $p = \frac{۲}{۳}$ و$q = \frac{۱}{۳}$ داریم ،

$$E(G) = \frac{۱}{۲} + \sum_{r=۱}^{۲n-۱} (\frac{۱}{۲} + \frac{E|D_r|}{۶r})$$

۲.۲.۵ اریبی انتخاب طرح سکه اریب تطبیقی وقتی $p(x) = \frac{۱-x}{۲}$ تحت استراتژی احتمال متناسب

اگر آماردان از طرح سکه اریب تطبیقی با تابع $p(x) = \frac{۱-x}{۲}$ که همان $\cup D(۰, \beta)$ است استفاده کند اریبی انتخاب طرح فوق تحت استراتژی $\theta_۲$ به صورت زیر محاسبه می شود

$$P(G_r) + \sum_{d=-r}^{r} (p(\frac{d}{r})\gamma + (۱-\gamma)q(\frac{d}{r}))P(D_r = d)$$

اگر $p(x) = \frac{۱-x}{۲}$ فرض شود .پس $p(\frac{d}{r}) = \frac{۱}{۲} - \frac{d}{۲r}$ و به همین صورت $q(\frac{d}{r}) = \frac{۱}{۲} + \frac{d}{۲r}$ ،در نتیجه

$$E(G) = \frac{۱}{۲} + \sum_{r=۱}^{۲n-۱} \sum_{d=-r}^{r} ((\frac{۱}{۲} - \frac{d}{۲r})^۲ + (\frac{۱}{۲} + \frac{d}{۲r})^۲)P(D_r = d)$$

و

$$= \frac{۱}{۲} + \sum_{r=۱}^{۲n-۱} \sum_{d=-r}^{r} (\frac{۱}{۲} + \frac{d^۲}{۲r^۲}) P(D_r = d)$$

$$= \frac{۱}{۲} + \sum_{r=۱}^{۲n-۱} (\frac{۱}{۲} + \frac{E(D_r^۲)}{۲r^۲})$$

شکل (۱) ارریبی انتخاب طرحهای $BCD(\frac{۲}{۳}), \cup D(۰, \beta)$ را تحت استراتژیهای $\theta_۱, \theta_۲$ برای مقادیر مختلف n نشان می دهد.

شکل ۱: آزمایش $۲n$ در $(E(G) - n)$ ارریبی انتخاب

بنا به شکل واضح است که تحت استراتژی $\theta_۱$، طرح $\cup D(۰, \beta)$ در مقایسه با طرح $BCD(\frac{۲}{۳})$ به ازاء $n > ۴$، ارریبی انتخاب کمتری دارد، اما تحت استراتژی $\theta_۲$، ارریبی انتخاب $\cup D(۰, \beta)$ به ازاء هر n بیشتر از ارریبی انتخاب طرح $BCD_{\theta_۲}(\frac{۲}{۳})$ است، از طرفی

مشاهده می‌شود که $BCD_{\theta_2}(\frac{2}{3})$ به ازاء هر $n > 3$، ارایی انتخاب کمتری نسبت به طرح $BCD_{\theta_1}(\frac{2}{3})$ دارد.

مراجع

BLACKWEEL, D. & HODGES,J.L.(1957).Design for the control of selection bias. Ann.Math.statist.28,449-60.

STIGLER ,STEPHEN(1969).The use of random allocation for the control of selection bias.Biometrika .56,553-60.

EFRON, BRADLEY(1971). Forcing a sequential experiment to be balanced. Biometrika. 58, 403-17.

WEI,L.J.(1977a),A class of designs for sequential clinical trials. J.Amer. Statist. Assoc. Vol. 72, No.358, 382-386.

WEI,L.J.(1977b),The adaptiv biased coin design for sequential experiments .The Annals of Statistics .Vol;No.1,92-100.

بررسی بعضی روش های رگرسیون جایگزین LS و مقایسهٔ آنها با یکدیگر

محمد رضا ربیعی [۱]، مجتبی گنجعلی [۲]

$P۱۱۲۰۹$

[۱] دانشگاه علوم پزشکی و خدمات درمانی گلستان
[۲] دانشگاه شهید بهشتی

چکیده: تعدادی شیوه‌های رگرسیون به عنوان روش‌های جایگزین برای کمترین توان‌های دوم وقتی نقاط نافذ[۱] یا دورافتاده[۲] در داده‌ها وجود دارند، مورد مطالعه قرار گرفته است. هیچ رگرسیون استواری نمی‌تواند در حالت کلی به عنوان بهترین، پذیرفته شود. روش‌های مختلف نقاط ضعف و و قوت متفاتی را با توجه به درصد داده‌های دورافتاده و موقعیت متغیرهای توضیحی و پاسخ، از خود نشان می‌دهند.هدف این مقاله مقایسه روش‌های استوار با سه خاصیت گوناگون الف)نقطهٔ فروریزش ب)کارایی ج)کراندار بودن نقاط موثر است.برآورد ضرایب به دست آمده از روش‌های گوناگون استوار با روش کمترین توان های دوم در دو مثال مقایسه شده‌اند.

واژه‌های کلیدی: کمترین توان های دوم، رگرسیون استوار[۳]، کارایی، کراندار بودن نقاط موثر[۴]، نقطهٔ فروریزش[۵].

۱ مقدمه

وقتی رگرسیون کمترین توان‌های دوم را با n مشاهده برای مدل p پارامتری $Y = X\beta + \epsilon$ به کار می‌بریم فرض‌های بخصوصی را در مورد بردار خطاها (ϵ) داریم ، یکی از این فرض هاآن است که بردار ϵ دارای توزیع $N(o, I\sigma^۲)$ است. به‌هرحال در عمل انحرافاتی از این فرضها وجود دارد. اگر این انحرافات جدی باشند امیدواریم آن‌ها را در رفتار باقیمانده‌ها تشخیص دهیم. بنابراین این انحرافات منجر می‌شود که تعدیلات مناسبی را در مدل و یا در ماتریس متغیرهای کمکی وارد کنیم . مثلاً ممکن است تبدیلی روی یک متغیر یا بیش از یک متغیر وارد کرده یا مدل را با اضافه کردن عبارتهای مرتبهٔ بالاتر تعدیل کنیم.

[۱] Leverage Pointes
[۲] Outlier
[۳] Robust Regression
[۴] Bounded Influence
[۵] Breakdown Point

اگر در تحلیل این نکته به نظر برسد که خطاها توزیع غیرنرمال دارند، مخصوصاً در مواقعی که توزیع خطا دمهای پهنتری نسبت به توزیع نرمال داشته باشد، یعنی از احتمالهای بیشتری در دمها نسبت به توزیع نرمال برخوردار باشد، بایستی، یک روش رگرسیون استوار را در نظر بگیریم. چنین توزیعهای دم پهن محتمل هستند که خطاهای بزرگتری را نسبت به حالت نرمال تولید کنند. روش کمترین توان‌های دوم در به‌دست آوردن برآورد پارامترها به هر مشاهده وزن یکسان می‌دهد. روش‌های استوار قادرند به مشاهدات وزن‌های نابرابر اختصاص دهند. به‌طور کلی مشاهداتی که باقیمانده‌های بزرگی را تولید می‌کنند به‌وسیلهٔ برآورد استوار کم وزنترند. تعدادی از این روش‌های استوار می توانند معرفی و به‌کار روند.

۲ روش رگرسیون کمترین توان‌های دوم (LS)

فرض کنید مدل مورد نظر به‌صورت زیر باشد:

$$Y = X\beta + \varepsilon \qquad (۱.۲)$$

که در آن Y،بردار$(n \times ۱)$مشاهدات و X،ماتریس $(n \times (p+۱))$ به‌صورت معلوم است. β بردار $(p+۱) \times ۱$ پارامترها وε بردار $(n \times ۱)$ خطاهاست. به‌قسمی که $E(\varepsilon) = ۰$ و$= V(\varepsilon)$ $I\sigma^۲$ بنابراین مولفه های ε ناهمبسته اند. برآورد کمترین توان‌های دوم β بردار b است که مقدار $\varepsilon'\varepsilon$ را مینیمم می کند:

$$b = (X'X)^{-۱}X'Y \qquad (۲.۲)$$

اگر خطاها مستقل باشند $\varepsilon_i \sim N(۰, \sigma^۲)$ آنگاه b برآورد حداکثر درستنمایی β است.

وضعیت مقادیر نقش مهمی در برازش کمترین توان‌های دوم بازی می کنند . در حالتی که همه نقاط در تعیین ارتفاع خط ، وزن یکسانی دارند، شیب خط قویاً تحت تاثیر مقادیر دور افتاده است . مشاهدات موثر را می توان به دور افتاده‌ها و نقاط نافذ تقسیم کرد.

دور افتاده‌ها: یک دور افتاده در بین مانده ها مانده ای است که از نظر قدر مطلق خیلی بزرگتر از بقیه است ، و شاید به فاصله بیشتر از سه یا چهار برابر انحراف معیار از میانگین مانده ها قرار دارد . به عبارت دیگر عنصر دور افتاده معمولاً به داده‌ای اطلاق می شود که دارای مقدار منتهی الیهی برای متغیر پاسخ باشد.

نقطهٔ نافذ: نقطه داده‌ای که مقدار منتهی الیهی برای یکی از متغیرهای توضیحی را داراست یک نقطهٔ نافذ نامیده می شود.نقاط نافذ اثری نامناسب در برآوردهای ضرایب

رگرسیونی دارد، زیرا که زمانی نقاط نافذ وجود دارند این نقاط هیچ تبعیتی از بقیه داده‌ها در مدل نمی‌کنند و اثر نامطلوبی در مدل باقی می‌گذارند.

نقطهٔ فروریزش: فرض کنید X عبارت است از یک نمونهٔ n تایی (x_i, y_i) و همچنین برآوردگر رگرسیون، T باشد. فرض کنید $\beta(m; T, X)$، سوپریمم $||T(X') - T(X)||$ برای تمام نمونه‌های آشفته X' باشد. وقتی این نمونه‌های آشفته، با جایگزین کردن هر m تا از نقاط اصلی نمونه با مقادیر دلخواه به‌دست آیند، آنگاه نقطهٔ فروریزش T برای X عبارتست از:

$$\varepsilon^*(T, X) = \min\{\frac{m}{n}; \text{ یی که نامتناهی هستند } \beta(m; T, X)\ \} \qquad (۳.۲)$$

به بیان دیگر، کوچکترین مقدار اغتشاش که سبب می‌شود برآوردگر، فاصلهٔ دلخواهی تا $T(X)$ داشته باشد نقطهٔ فروریزش نام دارد. توجه کنید که این تعریف شامل هیچ توزیع احتمالی نیست. برای کمترین توان‌های دوم $\varepsilon^*(T, X) = \frac{۱}{n}$ است، چون یک مشاهدهٔ بد می‌تواند سبب فروریزش شود.

کارایی: نسبت میانگین مربع خطا برای کمترین توان‌های دوم به میانگین توان‌های دوم خطا برای هر رگرسیون استوار دیگر را کارایی گویند.

خاصیت کراندار بودن نقاط موثر: روش کمترین توان‌های دوم به هر مشاهده وزن یکسان می‌دهد در صورتی که روش‌های استوار قادرند به مشاهدات وزن‌های نابرابر اختصاص داده و اثر آن‌ها را در مدل کم کنند.

ما در این جا به مقایسهٔ روش‌های رگرسیون استوار با رگرسیون کمترین توانهای دوم با سه خاصیت نقطهٔ فروریزش، کارایی، خاصیت کراندار بودن نقاط موثر می‌پردازیم.

۳ رگرسیون حداقل قدر مطلق انحرافات (LAD)

برآوردهای کمترین قدرمطلق انحرافات[۱] α و β مقادیر a و b ای هستند که عبارت زیر را کمینه می کنند:

$$\sum_i |y_i - (a + bx_i)| \qquad (۱.۳)$$

تفاضل $y_i - (a + bx_i)$ انحراف نقطهٔ (x_i, y_i) از خط $\hat{y} = a + bX$ نامیده می‌شود.

هرگاه هر یک از موارد زیر رخ دهد روش رگرسیون LAD نسبت به روش رگرسیون LS ترجیح داده می‌شود:

[۱] Least Absolute Deviations

الف) خطاها دارای توزیع متقارنی باشد که میانه آن برآوردگر کاراتری از نمونه باشد. ب) خطاها دارای توزیع دم پهن باشند.پ) دادهٔ دورافتاده در متغیر پاسخ داشته باشیم. ت) هم‌خطی چندگانه در بین متغیرهای مستقل وجود داشته باشد. ث) تابع زیان قدرمطلق خطا مناسب‌تر از تابع زیان درجهٔ دوم باشد.

خواص سه گانه

۱— چون در مقابل نقاط نافذ عاجز است لذا نقطهٔ فروریزش برای نمونه‌های متناهی همچنان برابر $\frac{۱}{n}$ است که با افزایش نمونه به‌سمت صفر میل می‌کند. ۲— با توجه به (۱) این روش از خاصیت کرانداری نقاط نافذ برخوردار نیست. ۳— اگر این خطاها دارای یک توزیع لاپلاس باشند، آنگاه کارایی برابر $۰/۷۰۷ = \frac{۲}{\sigma} < ۱$ درصد است.

۴ M- برآوردگرها

فرض کنید که $\rho(u)$ تابعی تعریف شده از u است و s برآورد مقیاس باشد (نه لزوماً به روش LS). یک برآوردگر را استوار گویند اگر عبارت:

$$\sum_{i=۱}^{n} \rho(\frac{e_i}{s}) = \sum_{i=۱}^{n} \rho\left\{\frac{Y_i - x_i'\beta}{s}\right\} \qquad (۱.۴)$$

را کمینه کند. برای کمینه کردن رابطهٔ فوق با مشتق گیری از این رابطه نسبت به پارامترهای β_j، $j = ۰, ۱, ۲, \cdots, k$ به $p = k + ۱$ معادله به شکل:

$$\sum_{i=۱}^{n} x_{ij}\psi\left\{\frac{Y_i - x_i'\beta}{s}\right\} = ۰ \qquad j = ۰, ۱, ۲, \cdots, k \qquad (۲.۴)$$

می‌رسیم.به دلیل عدم جواب صریح این معادلات، لازم است جوابها از روش تکراری محاسبه شوند. پس از محاسبه با استفاده از شکل ماتریسی داریم:

$$X'W_\beta X\beta = X'W_\beta Y \qquad (۳.۴)$$

که در آن $W_\beta = diagonal(w_{۱\beta}, w_{۲\beta}, \cdots, w_{n\beta})$ است، که

$$W_{i\beta} = \frac{\psi\{(Y_i - x_i'\beta)/s\}}{(Y_i - x_i'\beta)/s} \qquad i = ۱, ۲, \cdots, n \qquad (۴.۴)$$

و در صورتی که دقیقاً $Y_i - x_i'\beta = ۰$ باشد آن را یک تعریف می‌کنیم. می‌توان جواب تکراری را به‌صورت زیر نوشت :

$$\hat{\beta}_{q+۱} = (X'W_qX)^{-۱}X'W_qY \qquad q = ۰, ۱, ۲, \cdots \qquad (۵.۴)$$

این روش تکراری را کمترین توان‌های دوم با تکرار تجدید وزن[۱] (IRLS) می‌نامند.

:زا تستراب‌ه س‌ایقم ل‌ماع ی‌ارب راوتسا دروآرب ک‌ی

$$s = median|e_i - median(e_i)|/۰/۶۷۴۵ \qquad (۶.۴)$$

که وقتی n بزرگ و خطاها به‌طور نرمال توزیع شده‌باشند تقریباً برآوردگری نااریب برای $sd(y_i) = \sigma$ است.

۵ رگرسیون کمترین میانهٔ توان‌های دوم(LMS)

رسو[۱] (۱۹۸۴) در این روش مجموع توان‌های دوم مانده‌ها را با میانهٔ توان‌های دوم مانده‌ها جایگزین کرد، در نتیجه این برآورد می‌تواند در مقابل اغتشاش ۵۰ درصدی داده‌ها مقاومت کند. برآورد کمترین میانهٔ توان‌های دوم[۲] (LMS) به‌وسیلهٔ مینیمم کردن میانهٔ $e_i^۲$ ها یعنی

$$\underset{\hat{\beta}}{minimize}\ \ \underset{i}{med}\ \ e_i^۲ \qquad (۱.۵)$$

به دست می‌آید. این پیشنهاد اساس ایدهٔ همپل[۳] (۱۹۷۸، صفحهٔ ۳۸۰) است.

برای برآورد پارامتر مقیاس، σ، آن را به‌صورت زیر برآورد می‌کنیم:

$$s = ۱/۴۸۲۶(۱ + \frac{۵}{n-p-۱})\sqrt{\underset{i}{med}\ \ e_i^۲} \qquad (۲.۵)$$

که ۱/۴۸۳ یک عمل تصحیح مجانبی برای حالت خطاهای نرمال است.

تعریف : n مشاهدهٔ $(۱, x_{i۱}, ..., x_{ip}, y_i) = (x_i, y_i)$ متعلق به فضای خطی بردارهای سطری $p+۱$ هستند و پارامتر مجهول β یک بردار ستونی $p+۱$ بعدی $(\beta_۰, \beta_۱, ..., \beta_p)$ است. گوییم مشاهدات در موقعیت کلی هستند، هرگاه هر $p+۱$ تای آن‌ها یک β ی منحصر بفردی را تضمین کنند. مثلاً در حالتی که $p = ۲$ است جملهٔ بالا به این معنی است که هر جفت مشاهده $(۱, x_{i۱}, x_{i۲}, y_i)$ و $(۱, x_{j۱}, x_{j۲}, y_j)$ نمی‌توانند هم خط باشند.

قضیه : اگر $p > ۱$ و مشاهدات در موقعیت کلی باشند آنگاه نقطهٔ فروریزش روش LMS برابر است با

$$([\frac{n}{۲}] - p + ۲)/n \qquad (۳.۵)$$

---

[۱] Iteratively Reweighted Least Squares

[۱] Rousseeuw

[۲] Least Median of Squares

[۳] Hample

اغلب حد عبارت بالا را وقتی n به سمت بی‌نهایت میل می‌کند در نظر می‌گیرند. بنابراین می‌توان گفت که (با p ثابت) روش LS دارای نقطهٔ فروریزش صفر است. درحالی‌که نقطهٔ فروریزش روش LMS برابر با ۵۰ درصد است.

خواص سه گانه

۱– دارای نقطهٔ فروریزش ۵۰٪ است. ۲– از کارایی کمی برخوردار است. ۳– دارای خاصیت کراندار بودن نقاط موثر نیست.

۶ رگرسیون کمترین توان‌های دوم پیراسته (LTS)

رسو(۱۹۸۴) روش کمترین توان‌های دوم پیراسته[۱] (LTS) را به صورت زیر معرفی می‌کند :

$$\underset{\hat{\beta}}{minimize} \sum_{i=1}^{h} (e^2)_{i:n} \qquad (۱.۶)$$

که $(e^2)_{1:n} \leq \ldots \leq (e^2)_{n:n}$ توان دوم مانده‌های مرتب شده است.

قضیه: نقطهٔ فروریزش روش LTS با $h = [n/2] + [(p+1)/2]$ برابر است با:

$$\varepsilon_n^* = ([(n-p)/2] + 1)/n \qquad (۲.۶)$$

با انتخاب h تقریباً برابر $n/2$ نقطهٔ فروریزش به ۵۰٪ می رسد. در روش LTS یک قاعده برای $\hat{\sigma}$ به صورت :

$$\hat{\sigma}_{LTS} = C\sqrt{\frac{1}{n}\sum_{i=1}^{h}(e^2)_{i:n}} \qquad (۳.۶)$$

است که C یک عامل تصحیح نمونه متناهی و بزرگتر از یک است. در هر یک از موارد فوق زمانی یک مشاهده دورافتاده است که فقط وفقط نسبت $|e_i/\hat{\sigma}|$ بزرگ باشد.

خواص سه گانه

۱– دارای نقطهٔ فروریزش ۵۰٪ است. ۲– کارایی LTS از LMS بیشتر ونرخ همگرایی آن سریعتر است. ۳– دارای خاصیت کراندار بودن نقاط موثر است.

[۱] Least Trimmed of Squares

۷ S— برآورد

S— برآوردگر توسط رسو و یوهای[۲](۱۹۸۴) و رسو و لروی[۳](۱۹۸۷) به صورت زیر تعریف شده است :

$$\underset{\hat{\beta}}{minimize}\ S(e_1(\beta),\cdots,e_n(\beta)) \qquad (۱.۷)$$

و برآورد مقیاس نهایی آن $\hat{\sigma}_{S-est} = s(e_1(\hat{\beta}),\cdots,e_n(\hat{\beta}))$ است. انحراف معیار $s(e_1(\beta),\cdots,e_n(\beta))$ به عنوان جوابی از :

$$\frac{1}{n}\sum_{i=1}^{n}\rho(\frac{e_i}{s}) = K \qquad (۲.۷)$$

تعریف می شود و K اغلب برابر $E_\Phi[\rho]$ است که Φ تابع توزیع نرمال استاندارد است.تابع ρ با شروط زیر مشخص می‌شود: ۱) ρ متقارن و به طور پیوسته مشتق پذیر است. ۲) $c > 0$ وجود دارد به طوری که ρ روی فاصلهٔ $[0,c]$ اکیداً افزیشی و روی فاصلهٔ $[c,\infty]$ ثابت است. ۳) $\frac{K}{\rho(c)} = \frac{1}{2}$

اگر بیش از یک جواب برای معادلهٔ (۲.۷) وجود داشته باشد آنگاه:

$$S(e_1,\cdots,e_n) = sup\{s;\,(1/n)\sum\rho(e_i/s) = K\} \qquad (۳.۷)$$

اگر هیچ جوابی برای معادلهٔ (۲.۷)وجود نداشت آنگاه $S(e_1,\cdots,e_n)$ را برابر صفر قرار می دهیم.

با توجه به مطالب فوق $S(\beta) = S(e_1(\beta),\cdots,S(e_n(\beta))$ یک نوع M— برآوردگر استوار از انحراف معیار مانده هاست. به علاوه هدف S— برآوردگرها همانند M— برآوردگرها کمینه کردن انحراف معیار مانده‌هااست.

قضیه: یک S— برآوردگر ساخته شده از تابع ρ که در شرایط (۱) تا (۳) صدق می کند دارای نقطهٔ فروریزش:

$$\varepsilon_n^* = ([n/2] - p + 2)/n \qquad (۴.۷)$$

است، اگر برای هر نمونهٔ $\{(x_i,y_i); i = 1,\cdots,n\}$ در موقعیت کلی باشد.

همان طور که مشاهده می‌شود در حد نقطهٔ فروریزش به سمت ۰/۵ (۵۰ درصد) میل می‌کند.

---

[۲] Yohai
[۳] Leroy

خواص سه گانه

۱– دارای نقطهٔ فروریزش ٪۵۰ است. ۲– کارایی S– برآورد از LTS و LMS بیشتر است. ۳– دارای خاصیت کراندار بودن نقاط موثر نیست.

۸ M– برآوردگرهای تعمیم یافته(GM)

مالوز[۱](۱۹۷۵) توابع وزنی w را در روش هوبر به صورت زیر به کار می برد:

$$\sum_{i=1}^{n} w(x_i)\psi(e_i/\hat{\sigma})x_i = 0 \qquad (1.8)$$

: ت‌فرگ رظن رد ریز ت‌روصه‌با ر ق وف ل ومرف [۲]پ وِوشُ ش ور زا ن اوت ی م ایِ

$$\sum_{i=1}^{n} w(x_i)\psi(e_i/w(x_i)\hat{\sigma})x_i = 0 \qquad (2.8)$$

که این M– برآوردگرهای تعمیم یافته[۳] تک تک مشاهدات موثر را کراندار می کنند.

قضیه: نقطهٔ فروریزش روش GM برابر

$$\varepsilon_n^* = \frac{1}{p+1} \qquad (3.8)$$

که تابعی نزولی ازتعداد متغیرهای توضیحی (p) است (مارونا، باستاس و یوهای[۴] ۱۹۷۹). با توجه به تعریف GM در معادلهٔ (۲.۸) چند انتخاب ممکن را در جدول ۱ خلاصه کرده و در زیر یک مورد از آنها را به فرم الگوریتم در می‌آوریم. در ضمن خواص سه گانه در جدول خلاصه شده است.

جدول ۱: چند GM– برآورد خاص و خواص سه گانهٔ آنها

[۱] Mallows
[۲] Schweppe
[۳] Generalized M-estimators
[۴] Maronna, Bustos and Yohai

اجزاGM	GM1	GM2	GM3	GM4	GM5	GM6
فرمول GM	شُووپ	شُووپ	شُووپ	شُووپ	شُووپ	شُووپ
برآورد اولیه	LTS	M— برآورد همپل	S— برآورد	S— برآورد	S— برآورد	LS
برآورد مقیاس	$\hat{\sigma}_{LMS}$	$\hat{\tau}_k$	$\hat{\sigma}_{S-est}$	$\hat{\sigma}_{S-est}$	$\hat{\sigma}_{S-est}$	$\hat{\tau}_k$
تابع وزن w	$min(۱, b/RD^۲)$	$۱/\|z\|$	$med\|z\|/\|z\|$	$med\|z\|/\|z\|$	$med\|z\|/\|z\|$	$w_{i,k}$
تابع ψ	هوبر	هوبر	هوبر	هوبر	توکی[۶]	هوبر[۵]
ثابت‌های تنظیم کننده	۱٫۳۴۵	$۱٫۰۵p^{۱/۲}$	۱٫۳۴۵	۱٫۳۴۵	۴٫۶۸۵	$۲\sqrt{(p+۱)/n}$
خواص سه‌گانه						
کارایی بالا	دارد	دارد	دارد	دارد	دارد	دارد
فرورِیزش بالا	دارد	ندارد	ندارد	دارد	دارد	دارد
خاصیت کرانداری	دارد	دارد	دارد	دارد	دارد	دارد

الگوریتم GM1 در $K = ۰$ امین تکرار برآورد LTS پارامترها را محاسبه کرده، سپس به صورت زیر عمل می‌کنیم. ۱—مانده های $e_{i,k}$ را یافته و با استفاده از رابطهٔ (۲.۶) مقدار $\hat{\sigma}_{LMS}$ را بیابید.
۲— وزن های :

$$w_{i,k} = min\{۱, [\frac{b}{(x_i - m_x)'C^{-۱}(x_i - m_x)}]^{a/۲}\} \qquad (۴.۸)$$

را به دست آورید،که در آن $m_x.a = ۲, b = \chi^۲_{۰/۹۵}(p+۱)$ و C برآورد کننده‌های مکانی و کوواریانس همراه با متغیرهای کمکی در بیضی گون با کمترین حجم (MVE) هستند که می توانند با تابع $cov.mve()$در $Splus$ محاسنه شوند.توجه کنید که در جدول ۱ داریم $RD^۲ = (x_i - m_x)'C^{-۱}(x_i - m_x)$ و $z_i = ۱ - h_i$.
۳—وزن ها را استفاده کرده و برآوردهای جدید را با استفاده از رابطهٔ زیر محاسبه کنید:

$$\hat{\beta}_{k+۱} = \hat{\beta}_k + (X'BX)^{-۱}X'W\psi(\frac{e_{i,k}}{w_{i,k}\hat{\sigma}_{LMS}})\hat{\sigma}_{LMS} \qquad (۵.۸)$$

که در آن $B = diag(\psi'(e_i/w_i\hat{\sigma}_{LMS})), W = diag(w_i)$ و ψ' مشتق تابع ψ هوبر است. ۴—گام های ۱، ۲ و ۳ را تا همگرایی تکرار کنید (آنقدر ادامه دهید تا پارامترها در دو تکرار متوالی به هم نزدیک باشند).

خلاصه: جدول ۲ مقایسه‌ای از روش‌های رگرسیونی استوار را با سه مشخصهٔ ذکر شده خلاصه می‌کند.

جدول ۲: خواص سه گانهٔ روش‌های رگرسیون استوار

معیار	روش محاسبه	ε^*	شیوه
بهترین برآوردگرهای نااریب خطی	ساده	۰٪	LS
کارا	تکراری	۰٪	M— برآورد
کراندار کردن نقاط نافذ و کارا	تکراری با توابعی وزنی از x_iها	بالا ولی با افزایش p به سمت صفر میل می‌کند	GM
نقطهٔ فروریزش بالا	تکنیک‌های تعاقب تصویری	ثابت و به سمت ٪۵۰ میل می‌کند	LMS LTS S— برآورد

۹ کاربردهای عملی

مثال ۱ (وزن مخصوص چوب): بیست نمونه از ورقه‌های نازک مقاطع عرضی چوب کاج به ضخامت ۳۰cm تهیه شده‌اند.این مقاطع در مایع کلروزول رنگ شده و مقادیر زیر تعیین شده‌اند.

جدول ۳: داده‌های مربوط به عاملهای آناتومی و وزن مخصوص برشهای چوب کاج (دراپر و اسمیت (۱۹۸۷))

وزن مخصوص چوب	جذب نور چوب تابستانه ٪	جذب نور چوب بهاره ٪	چوب بهاره ٪	تعداد رشته‌ها در یک میلیمتر مربع از چوب تابستانه ٪	تعداد رشته‌ها در یک میلیمتر مربع از چوب بهاره
۰٫۵۳۴	۸۴٫۱	۵۳٫۸	۴۶٫۵	۱۰۵۹	۵۷۳
۰٫۵۳۵	۸۸٫۷	۵۴٫۵	۵۲٫۷	۱۳۵۶	۶۵۱
۰٫۵۷۰	۹۲٫۰	۵۲٫۱	۴۹٫۴	۱۲۷۳	۶۰۶
۰٫۵۲۸	۸۷٫۹	۵۰٫۳	۴۸٫۹	۱۱۵۱	۶۳۰
۰٫۵۴۸	۹۱٫۵	۵۱٫۹	۵۳٫۱	۱۱۳۵	۵۴۷
۰٫۵۵۵	۹۱٫۴	۵۵٫۲	۵۴٫۹	۱۲۳۶	۵۵۷
۰٫۴۸۱	۸۲٫۴	۴۵٫۵	۵۶٫۲	۱۲۳۱	۴۸۹
۰٫۵۱۶	۹۱٫۳	۴۴٫۳	۵۶٫۶	۱۵۶۴	۶۸۵
۰٫۴۷۵	۸۵٫۴	۴۶٫۴	۵۹٫۲	۱۱۸۲	۵۳۶
۰٫۴۸۶	۹۱٫۴	۵۶٫۴	۶۳٫۱	۱۵۶۴	۶۸۵
۰٫۵۵۴	۸۶٫۷	۴۸٫۱	۵۰٫۶	۱۵۸۸	۶۶۴
۰٫۵۱۹	۸۱٫۲	۴۸٫۴	۵۱٫۹	۱۳۳۵	۷۰۳
۰٫۴۹۲	۸۹٫۲	۵۱٫۹	۶۲٫۵	۱۳۹۵	۶۵۳
۰٫۵۱۷	۸۸٫۹	۵۶٫۵	۵۰٫۵	۱۱۱۴	۵۸۶
۰٫۵۰۲	۸۸٫۹	۵۷٫۰	۵۲٫۱	۱۱۴۳	۵۳۴
۰٫۵۰۸	۹۱٫۹	۶۱٫۲	۵۰٫۵	۱۳۲۰	۵۲۳
۰٫۵۲۰	۹۵٫۴	۶۰٫۸	۵۴٫۶	۱۲۴۹	۵۸۰
۰٫۵۰۶	۹۱٫۸	۵۳٫۴	۵۲٫۲	۱۰۲۸	۴۴۸
۰٫۵۹۵	۹۲٫۹	۵۳٫۲	۴۲٫۹	۱۰۵۷	۴۷۶
۰٫۵۶۸	۹۰٫۰	۵۶٫۶	۴۲٫۴	۱۰۵۷	۵۲۸

با مشاهدهٔ خروجی Splus مدل رگرسیونی به روش LS به صورت زیر به‌دست می آید:

$$\hat{Y} = ۰/۴۴۲۱ + ۰/۰۰۰۱X_۱ - ۰/۰۰۵۳X_۳ - ۰/۰۰۱۸X_۴ + ۰/۰۰۴۴X_۵ \quad (۱.۹)$$

جهت تشخیص نقاط موثر می‌توان از ماتریس کلاه‌دار H، آماره کوک وهمچنین باقیمانده‌های استاندارد شده می‌توان استفاده کرد. مقادیر فوق در جدول زیر خلاصه شده است.رسو و ون‌زومرن[۱] (۱۹۹۰) استفاده از برآوردکننده‌های LMS را برای تشخیص دورافتاده‌ها پیشنهاد کرده‌اند. با استفاده از روش های فوق مشاهدات ۴، ٦، ۸ و ۱۸ را به عنوان یک نقطهٔ موثر نشان می‌دهد.

جدول ۴: مقادیر مانده‌های استاندارد شده و استیودنت شده
و آمارهٔ کوک برای مشاهدات مشکوک

مشاهدات	h_{ii}	cook's	std.res	stud.res
۴	۰٫۲۵٦۴	۰٫۱۰۱۳	−۱٫۳۲۷۸	−۱٫۳٦۸۵
٦	۰٫۰۹۴۲	۰٫۰۹۱۸	۲٫۳۰۰۴	۲٫۸۱۰٦
۸	۰٫۵۰۹۷	۰٫۷۲۱۲	−۲٫۰۴۰۳	−۲٫۳۴۵۴
۱۸	۰٫۲۷۴۴	۰٫۰۷۳۱	−۱٫۰۷۷۰	−۱٫۰۸۳۷
min	۰٫۰۹۴۳	۰٫۰۰۰۳	−۲٫۰۴۰۳	−۲٫۳۴۵۵
max	۰٫۵۵۷۴	۰٫۷۲۱۲	۲٫۳۰۰۴	۲٫۸۱۰٦

جدول ۵: برآورد ضرایب مدل با استفاده روش‌های مختلف

روش	$\beta_۰$	$\beta_۱$	$\beta_۲$	$\beta_۳$	$\beta_۴$	$\beta_۵$
LS	۰٫۴۴۲۱	۰٫۰۰۰۱	۰٫۰۰۰۰	−۰٫۰۰۵۳	−۰٫۰۰۱۸	۰٫۰۰۴۴
LAD	۰٫۳۸۷۱	۰٫۰۰۰۲	۰٫۰۰۰۰	−۰٫۰۰۵۵	−۰٫۰۰۳۸	۰٫۰۰٦۱
$S-est$	۰٫۳۳٦٦	۰٫۰۰۰۱	۰٫۰۰۰۰	−۰٫۰۰۵٦	−۰٫۰۰۵۱	۰٫۰۰۷۵
LMS	۰٫۳۵۳۷	۰٫۰۰۰۱	۰٫۰۰۰۰	−۰٫۰۰۵۴	−۰٫۰۰۴۹	۰٫۰۰۷۲
LTS	۰٫۴۱۱٦	۰٫۰۰۰۲	۰٫۰۰۰۰	−۰٫۰۰۵۸	−۰٫۰۰۳۷	۰٫۰۰۵۷
GM1	۰٫۴۱۲٦	۰٫۰۰۰۲	۰٫۰۰۰۰	−۰٫۰۰۵۸	−۰٫۰۰۳۲	۰٫۰۰۵۵
GM6	۰٫۳۹۹۹	۰٫۰۰۰۲	۰٫۰۰۰۰	−۰٫۰۰۵۷	−۰٫۰۰۳٦	۰٫۰۰۵۸

انتخاب بهترین مدل

رسو (۱۹۸۴) پیشنهاد می کند در مسائلی که با متغیرهای متعددی مواجه هستیم یک برآوردگر بسیار استوار شبیه LMS و LTS را می‌توان برای پیدا کردن داده های دورافتاده به‌کار گرفت. در حالت کلی در مسائل عملی به نظر می رسد که بهتر است هم روش LMS و هم روش LS را در نظر بگیریم. اگر آن‌ها با هم توافق داشتند و نتایجشان با هم نسبتاً یکی بود آنگاه به نتیجه LS می‌توان اعتماد کرد. اما اگر بین آن‌ها اختلاف معنی داری

[۱] Van Zomeren

وجود داشت آنگاه باید به مطالعه باقیمانده هایی که از روش LMS به‌دست می‌آیند، داده‌هایی که بحران ایجاد کرده‌اند را پیدا کرد. پس از یافتن این داده‌های مؤثر، پیشنهاد می‌کنیم در M ـ برآورد گرها یا GM— برآوردگرهای مختلف محاسبه شده، روشی مناسبتر و کاراتر است که به این داده های مؤثر و دور افتاده وزن کمتری را نسبت دهد. به نظر میرسد از M ـ برآورد گرها توکی و GM— برآوردگرها GM6 و GM1 بهترین باشند.

مثال ۲(شبیه سازی): مدل زیر ر در نظر گیرید:

$$y = ۱ + x + \varepsilon \qquad (۲.۹)$$

که در آن $x \sim N(۰, ۱)$ و$\varepsilon \sim N(۰, ۱)$ است. پس از محاسبهٔ y با استفاده از مقادیر تولید شدهٔ x وε در اولین و دومین مطالعهٔ شبیه سازی $(n = ۵۰, n = ۱۰۰)$ برای آزمودنی‌های دوم و سی و پنجم مقادیر x را به ترتیب با ٦ و ٦— تعویض می‌کنیم تا عملکرد LS،GM۱ و GM٦ را مورد بررسی قرار دهیم. در مطالعهٔ سوم شبیه سازی $(n =$ ۱۵) مقادیر x برای آزمودنی‌های ۲ و ۱۳ را به ترتیب با ٦ و ٦— تغییر می‌دهیم

جدول ٦: خلاصهٔ برآورد پارامترهای مدل با استفاده از روش‌های LS، GM1 و GM6

	$n = ۱۰۰$		$n = ۵۰$		$n = ۱۵$	
روش	$\beta_۰$	$\beta_۱$	$\beta_۰$	$\beta_۱$	$\beta_۰$	$\beta_۱$
LS	۰٫۹۴٦	۰٫۵۴۰	۱٫۱۳٦	۰٫۲٦۸	۱٫۰۲۰	−۰٫۱٦۴
GM1	۱٫۰۲۳	۱٫۰۹۸	۰٫۹۸۹	۰٫۸۱۰	۱٫۱۷۸	۰٫۹۵۹
GM6	۱٫۰۳٦	۱٫۰۵۱	۱٫۱۲۰	۰٫٦٦۳	۱٫٦۷۷	۰٫۰۳۷

همان طور که جدول نشان می‌دهد برای نمونه‌های بزرگ از آنجا که مقدار واقعی $\beta_۰ =$ ۱ و $\beta_۱ =$ ۱ است، GM1 و GM6 بهتر از LS هستند ولی برای نمونه‌های کوچک تنها GM1 مقادیری نزدیک به مقادیر واقعی پارامترها تولید می‌کند.

۱۰ نتیجه گیری

با توجه به مطالب بیان شده:

۱— اگر در داده‌ها نقاط موثر(دورافتاده و نافذ) وجود نداشته باشد روش LS از کارایی لازم برخوردار است.

۲— بهتر است ابتدا نتایج روش LS را با یکی از روش‌های LTS یا LMS مقایسه کرده اگر منطبق نبودند به دنبال روش‌های دیگر استوار برای به دست آوردن مدل می‌گردیم.

۳— اگر تنها در مشاهدات، دادهٔ دورافتاده وجود داشته باشد از روش M— برآوردها استفاده می‌کنیم.

۴– اگر در مشاهدات، علاوه بر دورافتاده‌ها نقاط نافذ نیز مشاهده شود GM– برآوردها کاراترین روش هستند.

مراجع

Draper, N. and H. Smith (1987). *"Applied Regression Analysis"*,John Wiley and Sons.

Draper, N. and H. Smith (1998). *"Applied Regression Analysis"*,Second Edition,John Wiley and Sons.

Hampel, F.R. (1978). *"Optimally Bounding the Gross-Error Sensitivity and the Influence of Position in Factor Space"*, Proceedings of The Statistical Computing Section of The American Statistical Association, Washington, D.c., PP.59-64.

Huber, P.J.(1981). *"Robust Statistics"*, New York: John Wiley.

Mallows, C.L. (1975). *"On Some Topics in Robustness"*, Unpublished Memorandum, Bell Telephone Laboratories: Murrvay Hill, NJ.

Marronna,R. A., O. Bustos and V. Yohai(1979).*"Bias and Efficiency-Robustness of General M-estimators for Regression with Random Carriers"*, in Smoothing Techniques for Curve Estimation, eds. T.gasser and M.Rosenblatt, New York: Springer Verlag, pp. 91-116.

Rousseeuw, P.J. (1984). *"Least Median of Squares Regression"*, Journal of The American Statistical Association December, Vol.79, PP.388,871-880.

Rousseeuw, P.J., and A.J. Leroy (1987). *"Robust Regression and Outlier Detection"*. Wiley,New York.

Rousseeuw, P.J., and V.J. Yohai (1984). *"Robust Regression by Means of S-estimators"*,Robust and Nonlinear Time Series Analysis, eds. J. Franke, W. HHardle, and D.Martin, Springer-Verlag: Heidelberg,Germany, pp. 256-272.

Rousseeuw, P.J., and B.C. Van Zomeren (1990). *"Unmasking Multivariate Outliers and Leverage Points"*, Journal of the American Statistical Association, Vol. 85, pp. 633-651.

تقریب خطی اثرات متغیر های کمکی تغییر پذیر در مدل کاکس توسط مدلهای پویا

علیرضا رضایی [۱]، سیامک نور بلوچی [۲]

$A۱۵۲۲۶$

[۱] مرکز آمار ایران
[۲] دانشگاه مینسوتا

چکیده: در مدل کاکس [۱] یک فرض این است که اثر متغیرهای کمکی [۲] در طول زمان ثابت است. در بسیاری از مسائل واقعی به نظر می‌رسد این فرض ضعیف است و بهتر است که این اثرات به عنوان تابعی از زمان در نظر گرفته شوند. مسالهٔ اصلی، برآورد یک تابع است که نمی تواند با استفاده از روشهای معمولی با تعداد محدودی مشاهده برآورد شود. گامرمان [۳] (۱۹۹۱) طول کل زمان مطالعه را به N بازه افراز نمود سپس تقریب ساده $\beta(t)$ را در هر کدام از این بازه‌ها در نظر گرفت و آنها را با استفاده از مدلهای پویا برآورد نمود. یک راه برای بهبود تقریب $\beta(t)$، استفاده از تقریب خطی به جای تقریب ساده در هر بازه می باشد. ما دراین مقاله از تقریب خطی $\beta(t)$ در هر بازه استفاده کرده‌ایم و با استفاده از مدلهای پویا این تقریبها را برآورد نموده‌ایم.

واژه‌های کلیدی: تحلیل بقا، تحلیل بیزی، فرآیند دیریکله، مدل خطر متناسب، مدلهای پویا.

۱ مقدمه

در طول چهار دههٔ اخیر مطالعات زیادی روی آنالیز داده‌های بقا انجام گرفته است. از توابع بسیار مهم در آنالیز بقا تابع خطر است. که عبارت است از:

$$\lambda(t) = \lim_{\delta \to ۰} \frac{P(t \leq T < t+\delta | T \geq t)}{\delta}$$

در بسیاری از مسائل اتفاق می افتد که در کنار دادهٔ بقای بیمار اطلاعات اضافی دیگری از قبیل سن، جنس، میزان فشار خون، سطح هموگلوبین خون و... وجود دارد که می توانند

[۱] Cox
[۲] Covariates variables
[۳] Gamerman

به تنهایی یا در کنار همدیگر بر روی تابع خطر بیمار تاثیر داشته باشند. به این اطلاعات اضافی متغیرهای کمکی گفته می شود. کاکس(۱۹۷۳) مدل زیر را برای تابع خطر ارائه داد که وابستگی اندازهٔ تابع خطر بیمار به مقادیر متغیرهای کمکی را نشان می دهد.

$$\lambda(t) = \lambda_{\circ}(t)exp(Z'\beta) \qquad (۱)$$

$Z' = (Z_1, ..., Z_p)$ بردار p بعدی متغیرهای کمکی، $\beta = (\beta_1, ..., \beta_p)'$ بردار p بعدی اثرات متغیرهای کمکی و $\lambda_{\circ}(t)$ تابع خطر پایه [۱] است.

از زمانیکه کاکس مدل بالا را ارائه داد تا کنون کوششهای زیادی نسبت به بهبود این مدل و قویتر کردن فرضهایی که در آن به کار رفته انجام گرفته است. از فرضهایی که در این مدل شده این است که اثرات متغیرهای کمکی در طول زمان ثابت هستند اما در بسیاری از حالات به نظر می رسد این فرض بسیار ضعیف است و واقع بینانه تر این است که آنرا به عنوان یک تابع از زمان؛ $\beta(t)$ در نظر گیریم. بنابراین رابطهٔ ۱ به صورت زیر تغییر می کند.

$$\lambda(t) = \lambda_{\circ}(t)exp(Z'\beta(t)) \qquad (۲)$$

که $\beta(t) = (\beta_1(t), ..., \beta_p(t))'$ بردار p بعدی اثرات متغیرهای کمکی است. حال در اینجا مسألهٔ اصلی برآورد $\beta(t)$ که یک تابع است، می باشد که با روشهای معمولی نمی توان آن را با تعداد محدودی مشاهده برآورد نمود. در اینجا دو روش عمده برای برآورد وجود دارد. روش اول این است که تغییرات در طول زمان، وابسته به تعدادی پارامتر در نظر گرفته شود یا اصطلاحاً پارامتری شوند، آنگاه با برآورد پارامترها یک برآورد برای $\beta(t)$ حاصل می شود. روش دوم این است که زمان کل مطالعه به تعداد محدودی بازه افراز شود، آنگاه $\beta(t)$ در این بازه‌ها تقریب شود و با بدست آوردن برآوردی برای این تقریبها، برآوردی برای $\beta(t)$ بدست آید. در روش اول فرضهای زیادی برای شکل تغییرات $\beta(t)$ مورد نیاز است، اما روش دوم با ساخت یک تقریب « خوب » مواجه است. گامرمان(۱۹۹۱) از تقریب ثابت $\beta(t)$ در هر کدام از این بازه‌ها استفاده نمود و این تقریب ثابت در هر بازه را با استفاده از مدلهای پویا برآورد نمود. برای بهبود تقریب می توان از تقریب خطی استفاده نمود. ما در این مقاله از تقریب خطی $\beta(t)$ استفاده کرده‌ایم و آنها را در هر بازه برآورد نموده‌ایم.

[۱] Baseline hazard function

۲ تقریب خطی اثرات تغییر پذیر متغیرهای کمکی

در مدل کاکس می توان $\lambda_۰(t)$ را به صورت $\beta_۰(t) = ln(\lambda_۰(t))$ نوشت و Z و $\beta(t)$ را به صورت $Z = (۱, Z_۱, ..., Z_p)$ و $\beta(t) = (\beta_۰(t), \beta_۱(t), ..., \beta_p(t))^T$ تغییر داد. آنگاه رابطه ۲ به صورت زیر ساده‌تر می شود.

$$\lambda(t) = exp(Z\beta(t)) \qquad (۳)$$

فرض کنید n دادهٔ بقا، $t_۱, t_۲, ...t_n$ مشاهده شده‌اند که بعضی از آنها سانسور[۱] شده‌اند و کل زمان مطالعه با استفاده از نقاط $\tau_۱, ..., \tau_N$ به صورت زیر افراز شده است: $(\tau_N \geq t_n)$

$$\begin{cases} I_۱ = [۰, \tau_۱] \\ I_۲ = (\tau_۱, \tau_۲] \\ . \\ . \\ . \\ I_N = (\tau_{N-۱}, \tau_N] \end{cases}$$

با این فرض که تغییرات $\beta(t)$ داخل هر کدام از این بازه‌ها تقریباً به صورت خطی است، داریم:

$$\beta(t) \sim \begin{cases} \underline{\theta}_{۱۰} + \underline{\theta}_{۱۱} t \ \ , \tau_۰ \leq t \leq \tau_۱ \\ \underline{\theta}_{۲۰} + \underline{\theta}_{۲۱} t \ \ , \tau_۱ < t \leq \tau_۲ \\ . \\ . \\ . \\ \underline{\theta}_{N۰} + \underline{\theta}_{N۱} t \ , \tau_{N-۱} < t \leq \tau_N \end{cases} \qquad (۴)$$

که در آن $\underline{\theta}_{i۰} = (\theta_{i۰۱}, \theta_{i۰۲}, ..., \theta_{i۰p})$ و $\underline{\theta}_{i۱} = (\theta_{i۱۱}, \theta_{i۱۲}, ..., \theta_{i۱p})$، $i = ۱, ..N$، با توجه به روابط ۲ و ۳ تابع خطر j امین بیمار در i امین بازه برابر است با

$$\lambda_i^{(j)}(t) = exp\{Z_j(\underline{\theta}_{i۰} + \underline{\theta}_{i۱})\}, \quad i = ۱, ..., N \quad j = ۱, ..., n \qquad (۵)$$

---

[۱] Censored

با تغییر متغیر $\rho_{ij} = exp\{Z_j\underline{\theta}_{i0}\}$ و $\gamma_{ij} = exp\{Z_j\underline{\theta}_{i1}\}$ رابطهٔ ۵ به شکل زیر ساده تر می شود.

$$\lambda_i^{(j)}(t) = \rho_{ij}\gamma_{ij}^t, \quad i = 1,...,N \quad j = 1,...,n \qquad (٦)$$

بدون کاستن از کلیت مساله و برای سادگی در نمادگذاری می توان از اندیس j صرفنظر نمود و شکل کلی تابع خطر تجمعی برای بیماران را به صورت زیر بدست آورد.

$$\Lambda(t) = \begin{cases} \frac{\rho_1}{ln\gamma_1}[\gamma_1^t - \gamma_1^{\tau_0}] & , \quad \tau_0 \le t \le \tau_1 \\ . & \\ . & \\ . & \\ \sum_{k=1}^{i}\{\frac{\rho_k}{ln\gamma_k}[\gamma_k^{\tau_k} - \gamma_k^{\tau_{k-1}}]\} + \frac{\rho_i}{ln\gamma_i}[\gamma_i^t - \gamma_i^{\tau_{i-1}}] & , \tau_{i-1} < t \le \tau_i \\ . & \\ . & \\ . & \\ \sum_{k=1}^{N}\{\frac{\rho_k}{ln\gamma_k}[\gamma_k^{\tau_k} - \gamma_k^{\tau_{k-1}}]\} & , t \ge \tau_N \end{cases} \qquad (٧)$$

با استفاده از روابطی که تابع بقا و خطر با یکدیگر دارند، شکل تابع بقای هر بیمار در بازه i ام برابر است با

$$S(t) = exp\{-\sum_{k=1}^{i}\{\frac{\rho_k}{ln\gamma_k}[\gamma_k^{\tau_k} - \gamma_k^{\tau_{k-1}}]\} - \frac{\rho_i}{ln\gamma_i}[\gamma_i^t - \gamma_i^{\tau_{i-1}}]\} \qquad (٨)$$

بنابراین اگر B_{i-1} پیشامد زنده ماندن تا زمان t_{i-1} باشد، آنگاه براحتی با استفاده از تعریف احتمال شرطی و رابطه (٦) می‌توان نتیجه گرفت که $S(t|B_{i-1})$ و $f(t|B_{i-1})$ برابرند با

$$\begin{cases} S(t|B_{i-1}) = exp\{-\frac{\rho_i}{ln\gamma_i}[\gamma^t - \gamma^{\tau_{i-1}}] \\ f(t|B_{i-1}) = \rho_i\gamma_i^t exp\{-\frac{\rho_i}{ln\gamma_i}[\gamma^t - \gamma^{\tau_{i-1}}] \end{cases} \qquad (٩)$$

۳ تجزیهٔ زمانی تابع درستنمایی

با توجه به روابط بالا پارامترهای مجهولی که در هر بازهٔ iام، $i = 1,...,N$ وجود دارند عبارتند از: $\underline{\theta}_{i0}, \underline{\theta}_{i1}$ واضح است که تابع خطر هر بیمار در هر بازه به مقدار آن در بازه های قبلی وابسته است، از اینرو مقادیر پارامترهای $\underline{\theta}_{i0}, \underline{\theta}_{i1}$ نیز به مقادیر پارامترهایی که

در بازه های قبل وجود دارند وابسته‌اند. در استفاده از مدلهای پویا باید این وابستگی با یک رابطه بیان شود که اصطلاحاً به آن معادلهٔ سیستم گفته می شود. استنباط با استفاده از مدلهای پویا شامل تلفیق اطلاعات نهفته از مشاهدات تحت عنوان تابع درستنمایی و اطلاعات موجود در معادلهٔ سیستم به عنوان پیشین و تشکیل تابع پسین می باشد، به این صورت که در هر مرحله به این طریق استنباطی برای پارامترهای $\underline{\theta}_{i0}, \underline{\theta}_{i1}$ بدست می آید، با استفاده از این اطلاعات به عنوان پیشین برای مرحله بعد و همچنین مشاهدات مرحلهٔ بعد و معادلهٔ سیستم، این استنباطها بهنگام می شود. حال مسالهٔ اصلی در این قسمت، بدست آوردن تابع درستنمایی در هر مرحله است زیرا تابع درستنمایی در کل زمان مطالعه شامل تمام پارامترهای مورد بررسی در هر بازه می باشد و نمی توان از آن برای انجام استباط در i امین بازه $i = 1, ..., N$ استفاده نمود. بنابراین باید تابع درستنمایی را در هر بازه تجزیه کرد به طوریکه تابع درستنمایی تجزیه شده در هر بازه شامل پارامترهای مجهول همان بازه باشد تا بتوان بر اساس آن استنباطی برای پارامترهای آن بازه بدست آورد و از آن برای مرحلهٔ بعد استفاده کرد. به این علت این نوع تجزیه، تجزیهٔ زمانی نام گرفته است که تابع درستنمایی بین بازه‌هایی تجزیه شده است که زمان مطالعه با استفاده از آنها افراز شده است. گامرمان (۱۹۹۱) با در نظر گرفتن تقریب سادهٔ $\beta(t)$ درون هر بازه، تابع درستنمایی را درون آنها تجزیه نمود. ما برای تجزیه تابع درستنمایی به صورت زیر عمل نموده‌ایم که تقریباً مشابه روش تجزیهٔ گامرمان است.

لم ۱ : فرض کنید $T_1, ..., T_n$، n متغیر تصادفی مستقل زمان بقای بیماران باشند که تابع بقای آنها در i امین بازه به صورت رابطهٔ ۷ است. به علت سهولت در نوشتن، بدون کاستن از کلیات مساله فعلاً پارامترهای توزیع برای همهٔ آنها یکسان و برابر با $i = \rho_i, \theta_i$ $1, ..., N$ در نظر می گیریم. همچنین قرار دهید: $\mathcal{P} = (\rho_1, \gamma_1, ..., \rho_N, \gamma_N)$. تابع درستنمائی که از کل مشاهدات بدست می آید را می توان به صورت زیر بین بازه ها تجزیه نمود.

$$
\begin{aligned}
l(T|\mathcal{P}, D_0) &= \prod_{i=1}^{n} l_i(T^{(i)}|\mathcal{P}, D_{i-1}) \\
&= l_i(T^{(i)}|\Lambda, D_{i-1}) = \prod_{w \in F(I_i)} f(t_{iw}|\mathcal{P}, D_{i-1}) \\
&= \times \prod_{q \in R(I_i)} S(t_{iq}|\mathcal{P}, D_{i-1}), \qquad i = 1, ..., N
\end{aligned}
$$

که $T^{(i)} = \{min(T_k, t_i),\ k \in F(I_i) \bigcup R(I_i)\}$ اگر $w \in F(I_i)$ و $t_{iw} = t_w$ بازای $q \in R(I_i)$ و $t_{iq} = min\{t_q, t_i\}$ و D_i عبارت از کلیه اطلاعات موجود تا زمان τ_i.

اثبات: روابط فوق با استفاده از تعاریف احتمال شرطی و قضیه بیز و روابط موجود بین تابعهای بقا، خطر و چگالی در آنالیز بقا بدست می آیند. برای اثبات کامل رجوع کنید به [۲۱].

ملاحظه می شود که تابع درستنمایی تجزیه شده در هر بازه شامل پارامترهای همان بازه می باشد.

۴ معادلات تکامل زمانی پارامترها (معادلهٔ سیستم)

همانطور که اشاره شد آنالیز یک مسأله با استفاده از مدلهای پویا نیاز به مشخص بودن روابط زیر دارد:
۱) مدل نمونه گیری در هر بازه : عبارت است از تابع چگالی مشاهدات بدست آمده در هر بازه که همان عوامل تجزیه شده تابع درستنمایی داخل بازه های زمانی می‌باشند.
۲) معادلهٔ سیستم:

$$\begin{aligned} \underline{\theta}_{\circ i} &= \underline{\theta}_{\circ i-1} + \underline{\epsilon}_i \\ \underline{\theta}_{1 i} &= \underline{\theta}_{1 i-1} + \underline{\epsilon}_i \end{aligned} \qquad (10)$$

که در آن فرض بر این است که $\underline{\epsilon}_i \sim N(\circ, b_i \sigma_i I_p)$، $i = 1 \ldots N$ و b_i عبارت از طول بازهٔ i ام باشد.

ما برای بیان تکامل $\underline{\theta}$ در طول زمان از یک مدل گام برداری تصادفی استفاده کرده‌ایم. زیرا در مسائل آنالیز بقا که در آنها اثرات متغیرهای کمکی در طول زمان ثابت نیست، مقدار یک اثر در زمان t به شدت به مقدار آن در زمان قبل از آن وابسته است. هر چه از زمان t جلوتر و یا عقب تر برویم این وابستگی کمتر می‌شود. از اینرو به نظر می رسد که مدل گام برداری تصادفی برای $\underline{\theta}$ در طول زمان مناسب باشد. هر چه طول یک بازه کوتاهتر باشد، پارامترهایی که در بازهٔ بعد هستند، بیشتر به پارامترهایی که در این بازه هستند وابسته‌اند و اختلاف کمی بین آنهاست. در نتیجه واریانس ϵ_i کاهش می‌یابد و به همین ترتیب برای حالتی که طول بازه زیاد است، وابستگی پارامترها در بازه‌ها کمتر می شود و واریانس ϵ_i افزایش می یابد. برای نشان دادن این موضوع از ضریب b_i در توزیع ϵ_i ها استفاده شده است.

همچنین در ابتدای مطالعه توزیع پیشینی که برای $\underline{\theta}_1 = (\underline{\theta}_{\circ 1}, \underline{\theta}_{11})$ در نظر می گیریم یک توزیع نرمال $2p$ بعدی با میانگین صفر و واریانس $1000 I_p$ است. واریانس

زیاد تقریبا حالت بی اطلاعی را می رساند، اما علت انتخاب مقدار صفر برای میانگین این است که در مسائل آنالیز بقا به خصوص در آزمون اثرات متغیرهای کمکی، همیشه فرض اولیه یا فرض صفر این است که متغیر کمکی بدون اثر است مگر این که خلاف آن مشاهده شود. از اینرو از این فرض برای انتخاب میانگین $\underline{\theta}$ استفاده کرده‌ایم. بنابراین می توان دریافت که مدل و فرضهایی که در بالا برای $\underline{\theta}$ شده است، منطقی می باشند.

۵ تحلیل روش پیشنهادی

با توجه به روابط (۹) به نظر می رسد که می توان با استفاده از روشی که وست و هریسون و میگان در سال ۱۹۸۵ ارائه دادند برای این حالت نیز استفاده نمود. در روش پیشنهادی آنها اولین فرض این بود که مدل نمونه گیری نمایی باشد، یعنی مشاهداتی که بدست آمده‌اند از متغیر تصادفی باشند که توزیع آن متعلق به خانواده نمایی باشد. مدل نمونه گیری در روش پیشنهادی ما نمایی نیست زیرا شکل کلی خانواده نمایی به صورت زیر است:

$$p(Y_t|\eta_t,\phi) = exp[\phi\{Y_t\eta_t - a(\eta_t)\}]b(Y_t,\phi), \qquad (۱۱)$$

که در آن η_t پارامتر طبیعی توزیع است به طوریکه داریم: $E[Y_t|\eta_t,\phi] = \mu_t = a'(\eta_t)$ و ϕ پارامتر مقیاسی توزیع است که $V[Y_t|\eta_t,\phi)] = a''/\phi$. مطابق رابطهٔ (۱۱) در خانواده نمایی فقط یک حاصلضرب جدایی ناپذیر $Y_t\eta_t$ بین مشاهده و پارامتر طبیعی وجود دارد که به هیچ وجه نمی توان آنها را از هم جدا کرد، اما در رابطهٔ (۸) دو حاصلضرب وجود دارد که آنها را نمی توان از هم جدا نمود و به همین دلیل نمی توان تابع $b(Y_t,\phi)$ را تشکیل داد، در نتیجه نمی توان از روش وست و هریسون و میگان در این مساله استفاده نمود.

از طرفی اگر بخواهیم از روشهای معمول بیز استفاده کنیم حجم بسیار زیادی از محاسبات پیچیده در مساله وارد می شود. ما این مسأله را با استفاده از روشهای تولید داده حل کرده‌ایم.

مولر[۱] (۱۹۹۹) روشی را برای انجام تحلیل در اینگونه مسائل ارائه داد. او فرض کرد مدل نمونه‌گیری مشاهدات و معادلهٔ سیستم به ترتیب به صورت زیر باشد:

$$y_t \sim p(y_t|\theta_t) \quad : \text{مدل نمونه گیری مشاهدات}$$

$$\theta_t \sim g_t(\theta_{t-1}) + \omega_t \quad : \text{معادلهٔ سیستم}$$

[۱] Muller

که در آن $p(.)$ معلوم و لزومی ندارد نرمال باشد و $g(.)$ یک تابع معلوم و ω_t یک بردار یا متغیر تصادفی است که لزوماً دارای توزیع نرمال نیست.

روش او مبتنی بر انتگرال مونت کارلو می باشد، به این صورت که یک نمونهٔ مونت کارلو $A_1 = \{\theta_1, ..., \theta_n\}$ از توزیع پیشین که در مرحلهٔ اول برای پارامتر سیستم در نظر گرفته می شود $\pi(\theta_1|D_0)$ تولید می شود، آنگاه با استفاده از روش متروپلیس نمونهٔ اولیه A_1 به یک نمونه مونت کارلو $B_1 = \{\eta_1, ..., \eta_n\}$ از توزیع پسین $\pi(\theta_1|D_1)$ تبدیل می شود. با استفادهٔ مستقیم از معادلهٔ سیستم، B_1 تبدیل به یک نمونهٔ مونت کارلو A_2 از توزیع پیشین $\pi(\theta_2|D_1)$ می شود، به این طریق که ابتدا یک نمونه به اندازه n از توزیعی که برای ω در نظر گرفته شده است تولید می شود $\{\omega_1, ..., \omega_n\}$، آنگاه قرار داده می شود:

$$\theta_i = g(\eta_i) + \omega_i, \quad i = 1, ..., n$$

و این عملیات آنقدر انجام می شود تا فرآیند به نقطهٔ پایانی برسد. برای جزئیات بیشتر رجوع کنید به [۱۶]

در روش پیشنهادی مولر از الگوریتم متروپلیس برای تبدیل نمونهٔ مونت کارلو پیشین به یک نمونه مونت کارلوی پسین استفاده شده است. در بخش زیر اشاره کوتاهی به این الگوریتم شده است.

۱.۵ الگوریتم متروپلیس

فرض کنید A_t یک نمونهٔ مونت کارلو از توزیع پیشین در اختیار است، الگوریتم متروپلیس با این نمونه، یک زنجیر مارکف $x_1, ..., x_M$ تولید می کند به طوری که این زنجیر مارکف دارای توزیع تقریبی پسین است.

فرض کنید $p(.)$ توزیع پسین $p(\theta|data)$ باشد و $\theta_j \in A_i$ ،$i = 1, ..., n$ ،$j = 1, ..., n$ امین نمونه تولید شده در i امین بازه از توزیع پیشین $\pi(\theta_i|D_{i-1})$ باشد. به ترتیب زیر یک زنجیر مارکف تشکیل می شود.

۱) قرار دهید: $x_1 = \theta_i$
۲) y_1 را از توزیع کاندید $g(y_1|x_1)$ تولید کنید.
۳) با احتمال $\theta = min\{1, \frac{p(y_1)}{p(x_1)}\}$ قرار دهید: $x_2 = y_1$ در غیر این صورت $x_2 = x_1$
۴) با x_2 تولید شده به مرحلهٔ ۲ بروید و به همین ترتیب M بار ادامه دهید. x_M را به عنوان یک نمونه تولید شده از توزیع پسین در نظر گیرید.

توزیع کاندید توزیعی است که در شرط $g(y|x) = g(x|y)$ صدق کند و به این علت نام توزیع کاندید برای آن انتخاب شده است که چگالی بیشتری روی مقادیری از دامنهٔ توزیع پسین که احتمال آنها بیشتر است داراست. در عمل می توان این توزیع را نرمال با میانگین x_i و واریانس توزیع پسین در نظر گرفت.

تنها مشکل در کاربرد این روش در مدلهای پویا این است که شکل توزیع پسین برای انجام مرحله سوم الگوریتم فوق مشخص نیست، اما می توان آن را به صورت زیر برطرف نمود.
می دانیم $\pi(\theta_i|data) \propto l(data|\theta).\pi(\theta)$، بنابراین اگر بتوان $\pi(\theta)$ را با استفاده از نمونه مونت کارلویی که از آن در اختیار است برآورد نمود، آنگاه استفاده از روش فوق برای تحلیل مدلهای پویایی که برآوردهای پارامترهای مدل شکل بسته‌ای ندارند، مقدور می باشد. دقت کنید که ضریب تناسب در رابطهٔ بیز در صورت و مخرج رابطه‌ای که در مرحله ۳ الگوریتم است از بین می رود.
برای برآورد تابع توزیع می توان از فرآیند دیریکله استفاده کرد.

۲.۵ فرآیند دیریکله

برآورد تابع توزیع با استفاده از فرآیند دیریکله به صورت زیر است:

اگر $P \in \mathcal{D}(\theta)$ آنگاه بازای هر t $(F(t) \in Be(\alpha(-\infty, t], \alpha((t, \infty)))$ مخاطرهٔ بیز (مربع خطا) در حالت بدون نمونه وقتی مینیمم می شود که

$$\hat{F}(t) = E(F(t)) = F_0(t) = \alpha((-\infty, t])/\alpha(R)$$

واضح است که $F_0(t)$ حدس اولیهٔ ما برای شکل تابع توزیع را نشان می دهد.

برای وقتی که یک نمونه به اندازهٔ n داشته باشیم، قاعدهٔ بیز عبارت است از:

$$\begin{aligned} \hat{F}_n(t|x_1, \ldots x_n) &= \frac{\alpha((-\infty,t]) + \sum_{i=1}^{n} \delta_{x_i}((-\infty,t])}{\alpha(R)+n} \\ &= p_n F_0(t) + (1 - p_n) F_n(t|x_1, \ldots, x_n) \end{aligned} \qquad (۱۲)$$

که $p_n = \alpha(R)/(\alpha(R) + n)$ و $F_n(t|x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}((-\infty, t])$ تابع توزیع شهودی نمونه است.

یک مسألهٔ بسیار مهم تعیین مقدار $\alpha(R)$ است که اغلب در مسائل مجهول است. واضح است که نمی توان از $F_0(t)$ برآوردی برای آن بدست آورد. در مسالهٔ برآورد تابع

توزیع در فرایند دیریکله، باید $F_0(t)$ و $\alpha(R)$ به طور مجزا انتخاب شوند. (رجوع کنید به فرگوسن [۱۰])

فرض کنید در تجزیه و تحلیل مدلی که برای تجزیه و تحلیل در نظر گرفتیم، در مرحلهٔ i ام هستیم و می خواهیم تابع توزیع نمونهٔ مونت کارلویی را که از توزیع پیشین در این مرحله داریم برآورد نماییم. حدس اولیه و یا پیشین ما برای شکل تابع توزیع مشاهده ای که در این مرحله قرار دارد، یک توزیع نرمال با میانگین و واریانس توزیع پسینی است که در مرحلهٔ قبل بدست آمده است. اما برای انتخاب $\alpha(R)$ به ترتیب زیر عمل می کنیم.

تعریف می کنیم Z_n برابر است با تعداد مشاهده‌های مجزایی که در یک نمونه به اندازهٔ n در دسترس است. آنگاه $E(Z_n) = M \times [log((n+\alpha(R))/\alpha(R)))]$ هر چه n افزایش یابد $E(Z_n)$ نیز افزایش می یابد. در مسائلی که $\alpha(R)$ مجهول است از طریق Z_n استنباطهایی برای این مقادیر مجهول بدست می آورند. با توجه به این موضوع که تعداد مشاهدات یکسان در هر نمونه مونت کارلو A_t، تقریبا ۰ است بنابراین مقدار Z_n برابر با n است و از طریق رابطهٔ بالا و حل آن به روش عددی، برآوردی برای M بدست می آید که از آن برای برآورد $\alpha(R)$ استفاده شده است. حال برآورد تابع توزیع در هر مرحله با توجه به رابطهٔ (۱۲) حاصل می شود.

٦ مثال عددی

داده‌های جدول زیر را در نظر گیرید. که در آن تیمارها دو نوع روش برای برداشتن غدهٔ

تیمار	زمانهای بقا(روز)
روش ۱	۱۷،۴۲،۴۴،۴۸،۶۰،۷۲،۷۴،۹۵،۱۰۳،۱۰۸،۱۲۲، ۱۴۴،۱۶۷،۱۷۰،۱۸۳،۱۸۵،۱۹۳،۱۹۵،۱۹۷، ۲۰۸،۲۳۴،۲۳۵،۲۵۴،۳۰۷،۳۱۵،۴۰۱،۴۴۵، ۴۶۴،۴۸۴،۵۲۸،۵۴۲،۵۶۷،۵۷۷،۵۸۰،۷۹۵، ۸۵۵،۱۱۷۴،۱۲۱۴،۱۲۳۲،۱۳۶۶،۱۴۵۵، ۱۵۸۵،۱۶۲۲،۱۶۲۶،۱۷۳۶،
روش ۲	۱،۶۳،۱۰۵،۱۲۵،۱۸۲،۲۱۶،۲۵۰،۲۶۲،۳۰۱، ۳۰۱،۳۴۲،۳۵۴،۳۵۶،۳۵۸،۳۸۰،۳۸۳،۳۸۳، ۳۸۸،۳۹۴،۴۰۸،۴۶۰،۴۸۹،۴۹۹،۵۲۳،۵۲۴، ۵۳۵،۵۶۲،۵۶۹،۵۷۵،۶۷۶،۷۴۸،۷۷۸،۷۸۶، ۷۹۷،۹۵۵،۹۶۸،۹۷۷،۱۲۴۵،۱۲۷۱،۱۴۲۰، ۱۴۲۰،۱۴۶۰،۱۵۱۶،۱۵۵۱،۱۶۹۰،۱۶۹۴،

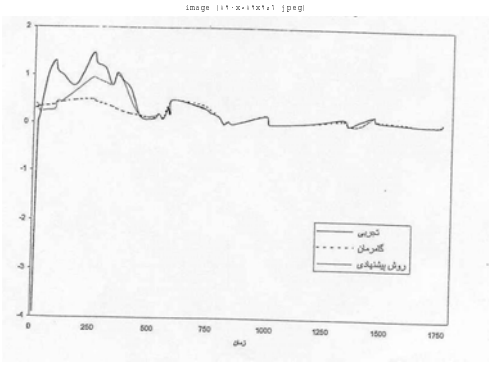
سرطانی می باشند. هدف مقایسه اثرات دو تیمار فوق روی زمان بقای بیماران سرطانی است. این کار را می توان با استفاده از مدل کاکس با یک متغیر کمکی که یک تابع نشانگر $z_۱$ است (۱ برای روش اول و ۰ برای روش دوم) انجام داد. دلایلی وجود دارد که اثر این متغیر کمکی در طول زمان ثابت نیست و بنابراین نمی توان از مدل معمولی کاکس استفاده نمود. کارتر[۱] (۱۹۸۳) فرض کرد که تغییرات این اثر در طول زمان به صورت $\theta_۰ + \theta_۱ t$ است و مقادیر $\theta_۰, \theta_۱$ را برآورد نمود. گامرمان (۱۹۹۱) تقریب سادهٔ اثرات تیمارها را در نظر گرفت و با استفاده از مدلهای پویا آنها را برآورد نمود.(رجوع کنید به گامرمان [۱۲]) ما با استفاده از روشی که ارائه دادیم این اثرات را به صورت خطی در هر بازه برآورد می نماییم.

ابتدا در شروع مطالعه یعنی در زمان صفر، توزیع پیشینی که برای $\theta_{۰۱} = (\theta_{۰۱۱}, \theta_{۰۱۲}), \theta_{۱۱} = (\theta_{۱۱۱}, \theta_{۱۱۲})$ در نظر می گیریم، یک توزیع نرمال با میانگین صفر و واریانس $۱۰۰۰ I_۴$ است. زیرا فرض کرده ایم که $\theta_{۰۱} = (\theta_{۰۱۱}, \theta_{۰۱۲}), \theta_{۱۱} = (\theta_{۱۱۱}, \theta_{۱۱۲})$ از یکدیگر مستقل هستند. البته این فرضی است که برای همه $\theta_i = (\theta_{۰i}, \theta_{۱i}), \quad i = ۱, ..., N$ ها در نظر گرفته‌ایم.

اطلاعاتی که از مشاهدات درون هر بازه بدست می آید، یعنی همان عوامل تجزیه شده تابع درستنمایی در هر بازه را نیز به تک تک مشاهدات تقسیم کرده‌ایم، به این صورت که با توجه به استقلال متغیرهای تصادفی زمانهای بقا در بازه‌های زمانی، در بازه i ام برای آنالیز بیزی از یک مشاهده استفاده می نمائیم ، بنابراین باید مشاهدات درون هر بازه را مرتب نمود و این کار به این صورت انجام می گیرد که ابتدا زمانهای مرگ بیمارانی که در بازهٔ i ام می میرند را به ترتیب صعودی مرتب و سپس برای بیمارانی که در طول بازه i ام زنده هستند، مقدار τ_i را قرار می دهیم. لازم است که نماد گذاری مناسبتری به کار بریم. قرار می دهیم $D_{i-۱,j}$ عبارت است از مجموع اطلاعات موجود تا قبل از زمان τ_i و اطلاعات حاصل از اولین مشاهده مرتب شده تا j امین مشاهده مرتب شده در بازهٔ i ام. همچنین p_{ij} و m_{ij} به ترتیب برابرند با میانگین توزیع پیشین و پسینی که از تابع درستنمایی j امین بیمار در بازه i ام بدست می آید.

همانگونه که قبلاً اشاره گردید از میانگین پسین m_{ij} در هر مرحله برای برآورد توزیع پیشین در مرحله بعد استفاده می نمائیم تا از تمام توابع درستنمایی مشاهدات مرتب شده در بازه i ام استفاده کنیم. آنگاه با توجه به معادلهٔ سیستم وارد مرحله (بازه $i+۱$) می شویم و همینطور ادامه می دهیم تا به آخرین بازه برسیم. برآوردهایی که برای θ_i ها بدست می آیند، میانگین توزیع پسینی هستند که از آخرین مشاهده در بازهٔ i ام بدست می آیند.

[۱] Carter



شکل ۱: نمودار مقدار اثر متغیر کمکی در طول زمان و مقایسه روشهای تقریب

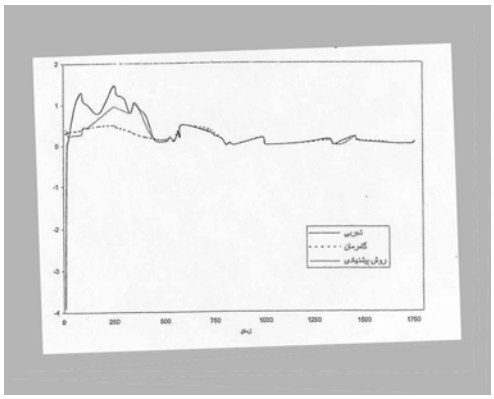
در شکل صفحهٔ بعد مقادیر برآورد پیشنهادی در این رساله و مقادیر برآورد گامرمان و مقادیری را که از قبل و از تجربه در اختیار است،مشاهده می نمایید.

همانطور که در شکل ملاحظه می نمایید مقادیر پیشنهادی نزدیک به مقادیر تجربی هستند و به وضوح می توان دید که تقریب اثر متغیر کمکی بهتر شده است.

مراجع

Antoniak, C. E. "Mixtures of dirichlet process with applications to Bayesian nonparametric problems". (1974) *Ann. Statist,* 2, 1152-1174.

Chang, Myron. N. and Chung Daehyun. "Isotonic window estimation of the baseline hazard function in Cox's regression model under order Restric-



شکل ۲: نمودار مقدار اثر تمغیر کمکی در طول زمان با استفاده از روش پیشنهادی

tion". (1998) *Scand. J. Statist.* 25, 151-161.

Chap, T. L. " Applied Survival analysis". (1997) New York: Wiely.

Chatfield, Christopher."Time series analysis". (1985) Boston: PWS-Kent Publishing.

Collet, D. " Modelling survival data in medical research". (1994) London: Chapman and Hall.

Cox, D. R and Oakes, D. " Analysis of survival data". (1984) London: Chapman and Hall.

Dellaportas, P. and Smith, A. F. M. "Bayesian inference for generalized and proportional hazard Model via Gibbs sampling". (1993) *Appl. Statist.* 42, 443-459.

Escobar, D. and West, M. "Bayesian density estimation and inferense using mixture". (1995) *J. Am. Statist. Ass*, 90, 577-588.

Escobar, D. "Estimating normal means with a dirichlet process prior". (1994) *J. Am. Statist Ass*, 89, 268-277.

Ferguson, T, S. "A Bayesian analysis of some nonparametric problems". (1973) *Ann. Statist,* 1, 209-230.

Ferguson, T, S. "Bayesian density estimation by mixtures of normal distributions". (1994) New York: Academic Press, pp. 287-302.

Gamerman, D. "Dynamic Bayesian models for survival data". (1991) *Appl. Statist.*, 40, 63-79.

Ghorai, J. K. "Nonparametric Bayesian estimation of a survival function under the proportional hazard model". (1989) *Commun, Statist,* 18, 1831-1842.

Gilks, W. R. and Wild. P. "Adaptive rejection sampling for Gibbs sampling". (1992) *Appl. Statist*, 41, 337-348.

Meinhold, R. J and Singpurwalla, N. D. "Understanding the Kalman filter". (1983) *Am. Statistn*, 37, 123-127.

Muller, P. "Posterior integration in dynamic models". (1999) Technical Report, Duke University, Institue of Statistics and Decision Science.

Smith, A. F. M. and Gelefand, A. E. "Bayesian statistics whitout tears: A sampling-resampling perspective". (1992) *Am. Statistn*, 46. 84-88.

Tsai, W. Y. " Estimation of survival curves from dependent censorship models via a generalized self-consistent property with nonparametric Bayesian estimation application". (1986) *Ann. Statist,* 14, 238-249.

West, M. and Harrison, P. J. and Migon, H. S. Dynamic generalized linear models and Bayesian forecasting (with discussion)". (1985) *J. Am. Statist. Ass*, 80, 73-97.

West, M. “Bayesian kernel density estimation”. (1990) Technical Report, 90-A02, Duke University, Institue of Statistics and Decision Science.

علیرضا رضایی، *برآورد تقریب خطی اثرات تغییر پذیر در مدل کاکس با استفاده از مدلهای پویا*، رسالهٔ کارشناسی ارشد، دانشکدهٔ علوم ریاضی—گروه آمار، دانشگاه شهید بهشتی (۱۳۷۹) .

روشهای احتمالاتی در حل مسائل دترمینیستیک

بیژن ظهوری زنگنه

$P۱۱۹۰۰$

دانشکده علوم ریاضی، دانشگاه صنعتی شریف

چکیده: اثبات قضیه‌های احتمال بر اساس تکنیکهای آنالیز ریاضی را در اغلب قضیه‌های نظریه احتمال دیده‌ایم. در این سخنرانی قصد داریم که جریان معکوس این پدیده را یعنی کاربرد روشهای تصادفی را در حمله به مسائل آنالیز کلاسیک بررسی کنیم. یکی از ابتدایی‌ترین مثالهای این روشها اثبات قضیه تقریب وایراشتراس به وسیله احتمالات است. این روشها در حل مسائل نظریه پتانسیل، مسئله دیریشله و مسئله شرایط مرزی مارتین نیز کاربرد دارد. در این سخنرانی عمومی سعی خواهیم کرد با زبان شهودی و غیررسمی به بعضی از این کاربردها بپردازیم.

واژه‌های کلیدی: روشهای احتمالاتی

احتمال به روایت تاریخ

نظریهٔ احتمال در ابتدا و نیز پس از آن برای مدتی طولانی، عبارت بود از صورت آرمانی و تحلیل برخی پدیده‌های زندگی واقعی در خارج از حیطهٔ ریاضیات، اما اندک اندک در نیمه نخست این قرن احتمال ریاضی، بخشی معمولی از ریاضیات شد. تا قرن پانزدهم هیچگونه بررسی علمی در مورد پیشامدهای تصادفی انجام نشد. دانش پژوهان ایتالیایی لوکا با چولی (۱۴۴۵-۱۵۱۴)، نیکولا تارتاگلیا (۱۴۹۹-۱۵۵۷)، چرولاموکاردانو (۱۵۰۰-۱۵۷۱)، از جمله پیشکسوتان دانش ریاضی هستند که احتمالهای مربوط به بسیاری از بازیهای تصادفی را محاسبه کرده‌اند. به‌قول دیود مامفرد در مقاله طلوع عصر روشهای تصادفی «اگر جلوتر بیاییم، می‌بینیم که در عصر رنسان، کاردانو شخصیت بی‌نظیری است. او به خاطر کتاب فن کبیرش (۱۵۴۵) غالباً مبدع i خوانده می‌شود. ظاهراً وی یکی از خبره‌ترین افراد در زمینهٔ عملیات صوری جبر بود به‌طوری که تبعات قواعد منطقی جبر را یک گام فراتر از اسلاف خویش برد. ولی او در عین حال، معتاد به قمار هم بود و در کتاب «بازیهای شانسی» خود نخستین تحلیل را از قوانین شانس ارائه کرد، اما خجالت می‌کشید آن را انتشار دهد و این کتاب تا ۱۶۶۳ به چاپ نرسید، یعنی تقریباً مقارن با زمانی که یاکوب برنولی کار خود را آغاز کرد.» (مامفرد، نشر ریاضی ۱۳). به هر حال پیشرفت واقعی

در فرانسه از سال ۱۶۵۴ آغاز شد، از وقتی بلز پاسکال (۱۶۲۳-۱۶۶۲) و پیردو فرما (۱۶۰۱-۱۶۶۵) دو ریاضیدان نامی نامه‌هایی به یکدیگر رد و بدل کردند که در این در مورد نامه‌ها روشهای کلی محاسبه احتمالها بحث کرده‌اند. در سال ۱۶۵۵ دانشمند معروف آلمانی کریستین هویگنس (۱۶۲۹-۱۶۹۵) به آنها پیوست، و این همکاری بسیار پر ثمر بود. در سال ۱۶۵۷ هویگنس اولین کتاب درباره احتمال را تحت عنوان «درباره محاسبات بازیهای شانسی» نوشت. این کتاب به منزلهٔ تولد واقعی احتمال محسوب می‌شود.

بعد از این تمام غول‌های ریاضی مانند برنولی، لاپلاس، پواسن و گاوس که استاد دقت در رشته‌های دیگر ریاضی بودند، قضیه‌های غلط و یا کم دقتی را در احتمال ثابت کردند تا اینکه در سال ۱۹۰۰ در کنگره بین‌المللی ریاضیدانها در پاریس، دیوید هیلبرت (۱۸۶۳-۱۹۴۳) ۲۳ مساله را که به عقیده او حل آنها در پیشرفت ریاضیات موثر بود پیشنهاد کرد. یکی از این مسایل، بحث اصول موضوعی نظریه احتمال بود. در راستای رسیدن به این هدف کارهایی به وسیله امیل بورل (۱۸۷۱-۱۹۵۶)، و برنشتاین (۱۸۸۰-۱۹۶۸) انجام شد، تا اینکه در سال ۱۹۳۳ اندری کولموگورف (۱۹۰۳-۱۹۸۷) به صورتی موفقیت‌آمیز نظریه احتمال را اصل موضوعی کرد.

نظریه احتمال کلموگرف بر اساس نظریه اندازه لبگ (۱۹۰۲) بنیان گذاشته شد. کلموگرف در نخستین صفحات تکنگاشت مشهور خود درباره نظریه احتمال صراحتاً می‌گوید که متغیرهای تصادفی حقیقی مقدار، همان توابع اندازه‌پذیرند و امیدهای ریاضی انتگرالهای آنها مع‌هذا، اندازه‌پذیری یک تابع حقیقی مقدار را تعریف می‌کند و وقتی که می‌خواهد امید ریاضی یک متغیر تصادفی را تعریف کند صاف و ساده نمی‌گوید که این مقدار برابر است با انتگرال متغیر تصادفی نسبت به اندازه احتمال مفروض، بلکه انتگرال را نیز تعریف می‌کند (دوب، نشر ریاضی ۱۲) این سنت در اغلب کتب نظریه احتمال باقی مانده است (به‌عنوان مثال رجوع شود به Chung 1977).

کلموگورف در تکنگاشت خود مطالب بسیار مهمی را مطرح کرد. او فضای احتمال، ساختن فرآیند تصادفی روی فضای بینهایت بعدی احتمال و امید شرطی نسبت به یک میدان سیگمایی را معرفی کرد. فهم قضیه توسیع کلموگرف روی فضای بینهایت بعدی مدتها برای بسیاری از ریاضیدانها مشکل بود. جوزف دوب از بنیان‌گذاران احتمال در آمریکا می‌گوید:

«نویسنده به یاد می‌آورد که منظور کولموگروف از اندازه روی فضای تابعی را تا زمانی دراز پس از آنکه تکنگاشت وی را خوانده بود، در نمی‌یافته است» (دوب، نشر ریاضی ۱۲)

این رویکرد مورد توجه و توافق اغلب احتمال‌دانها قرار گرفت. تا جایی که خیلی از احتمال‌دانها احتمال را جزیی از آنالیز می‌دانند. دوب می‌گوید «برخی ریاضیدانان بر آنند که هرگاه با خواص تحلیلی احتمال و امید ریاضی سروکار داشته باشیم، موضوع بخشی از آنالیز است، ولی اگر با دنباله‌های نمونه‌ای و توابع نمونه‌ای سروکار داشته باشیم، موضوع عبارت است از احتمال، نه آنالیز. این مولفان در موقعیت جالب توجهی هستند از این رو که در نظر کردن به تابع دو متغیره $(t,\omega) \to x(t,\omega)$ مثلاً در فرآیندهای تصادفی اگر خانواده توابع $x(t,.)$ هنگامی که t تغییر می‌کند مورد مطالعه باشد آن را آنالیز می‌خوانند، ولی اگر خانواده توابع $x(.,\omega)$ هنگامی که ω تغییر می‌کند مورد بررسی باشد آن را احتمال می‌نامند و قطعاً آنالیز به حساب نمی‌آورند. دقیقتر بگوییم، ایشان بحث پیرامون توزیعها و پرسشهای مربوطه را آنالیز می‌دانند، اما بحثهای به زبان توابع نمونه‌ای را آنالیز نمی‌دانند. این دیدگاه در قول ذیل بیان شده است.

پروتر: ایتو (Ito) در سال ۱۹۴۴، با ارائهٔ انتگرالش که در آن فرآیندهای تصادفی انتگرال بودند، توانست بخش چند بعدی را با تکنیکهای احتمالاتی محض مورد بررسی قرار دهد، که نسبت به روشهای آنالیو فلر بهتر است»(دوب، نشر ریاضی ۱۲) .

در هر حال دوب این ایده را که احتمال مستقل از آنالیز باشد قبول ندارد و در آخر این مقاله با آوردن استدلالی درباره توابع رادماچر (Rademacher)، و قضیه پل لوی می‌نویسد:

«دیگر بر عهده خواننده است که داوری کند کدامیک از این نتایج نظریه اندازه‌ای است و کدام یک احتمالاتی، و آیا اصلاً بیرون راندن احتمال ریاضی از قلمرو آنالیز فایده‌ای دارد، و اگر دارد، آیا نظریه اندازه را هم نباید بیرون راند؟» (دوب، نشر ریاضی ۱۲).

در مقابل رهیافت دوب، رهیافت دیگری است که به‌وسیله دیوید مامفرد مطرح می‌گردد که معتقد است که «رهیافت دیگر آن است که مفهوم «متغیر تصادفی» در مرکز توجه قرار بگیرد و همه کارها با انواع و اقسام دستکاری در متغیرهای تصادفی انجام شود.» (مامفرد، نشر ریاضی ۱۳).

دیوید مامفورد در مقاله «طلوع عصر روشهای تصادفی» می‌خواهد ریاضی جدیدی به‌وجود آورد او می‌نویسد «در رهیافت تقلیل‌گرا، متغیر تصادفی برحسب اندازه تعریف می‌شود که خود برحسب نظریه اعداد حقیقی تعریف می‌شود، و این را هم نظریه مجموعه‌ها تعریف می‌کند که خودش بر اساس حساب محمولات تعریف می‌شود. در عوض من می‌خواهم بگویم که باید علی‌الاصول بتوان متغیرهای تصادفی را در مبانی منطق و ریاضیات ادغام کرد و به صورتبندی شفافتر و کاملتری از دیدگاه تصادفی رسید. من خودم هنوز صورتبندی کامل و قطعی از این قضیه ندارم».

در هر حال چه ما مانند «دوب» نظریه احتمال را جزیی از آنالیز بدانیم چه مانند «دیوید مامفرد» قصد داشته باشیم ریاضی جدیدی با تغییر در مبانی آن به‌وجود آوریم و یا مانند پروتر بخشی را آنالیز و بخشی را احتمال بنامیم، این مسئله مسلم است که نظریه احتمال کاملاً آغشته به آنالیز است و بر اساس آن توسعه یافته حال چه بخواهد در درون آنالیز باقی بماند و چه از آن خارج شود. روشهای تصادفی دارای شهود ویژه خود است که در بقیه قسمتهای آنالیز وجود ندارد. این شهود باعث مطرح شدن مسائل زیادی در احتمال گشته و منبع گسترش آن است. در اغلب مسائل برای اثبات قضیه‌های احتمال از آنالیز کمک می‌گیریم. در این سخنرانی قصد داریم برای اثبات مسائل کلاسیک آنالیز از روشهای تصادفی کمک بگیریم و آنها را اثبات کنیم. بنابراین با نگاه یک احتمال دان به اشیاء آنالیز ریاضی می‌پردازیم.

نگاه احتمالاتی به آنالیز کلاسیک

بازه بسته $[۰,۱]$ را در نظر بگیریم. می‌خواهیم به این بازه دترمینیستیگ با تعبیر احتمالی جان تازه‌ای بدمیم.

امیل بورل در سال ۱۹۰۹ هر $x \in [۰,۱]$ را به‌صورت بسط دودوئی

$$x = ۰.x_۱, x_۲, \ldots$$

نوشت که رقم x_j یا صفر است یا یک: این رقمها توابعی از x هستند. «اگر بازه $[۰,۱]$ را با اندازه لبگ در نظر بگیریم که یک اندازه احتمال بر این بازه است، این تابعها به شکل معجزه آسایی متغیرهایی تصادفی می‌شوند که دقیقاً همان توزیعهایی را دارند که در محاسبه احتمالات پرتاب سکه به‌کار می‌روند. یعنی ۲^{-n} برابر است با احتمال منسوب به این رویداد که در یک آزمایش پرتاب سکه نخستین n پرتاب دنباله معینی از شیر و خط به دست بدهد، و ۲^{-n} همچنین طول کل (=اندازه لبگ) تعدادی متناهی بازه است که نقاط متعلق به آنها بسطهایی دودویی با دنباله‌ای مشخص از صفرها و یکها در n جایگاه خاص دارند.» (دوب، نشر ریاضی ۱۲). با این دید بازه $[۰,۱]$ فضای نمونه‌ای آزمایش برنولی است (رجوع شود به Adams, Guiliemin).

از بسط دودویی بالا و آزمایش برنولی یعنی پرتاب مستقل بینهایت بار سکه، می‌توان نشان داد که $\frac{x_۱+x_۲+\cdots+x_n}{n}$ به $\frac{۱}{۲}$ میل می‌کند. «ولی بیان قویتری از قانون اعداد بزرگ حکمی بود که بورل به‌دست آوردید ـطی یک برهان اشتباه و غیرقابل تصحیحـ که این دنباله از میانگینها به ازای تقریباً هر x به $\frac{۱}{۲}$ میل می‌کند (با احتمال ۱). یک سال بعد فیبر

(Faber) برهان درستی برای این حکم ارائه کرد و از آن هنگام برهانهای بسیار ساده‌تری هم به دست آمده‌اند، [Billingsley,Breiman,Chung و زنگنه]. فرشه، حرمت بورل را نگه داشت «برهان بورل بیش از حد کوتاه است. در آن چندین استدلال میانی حذف شده است و نیز احکامی بدون برهان فرض شده‌اند» (دوب، نشر ریاضی ۱۲).

با استفاده از قضیه قانون قوی اعداد بزرگ برای دنباله‌های برنولی می‌توان به اثبات سادهٔ برنشتین از قضیه وایراشتراس برای تقریب توابع پیوسته با چند جمله‌ای‌ها دست یافت.

گیریم $f = f(x)$ یک تابع پیوسته روی بازه $[۰, ۱]$ باشد. ثابت می‌کنیم f حد یکنواخت چندجمله‌ایهای برنیشتین به‌صورت

$$B_n(x) = \sum_{k=۰}^{n} f(k/n) C_n^k x^k (۱ - x)^{n-k}$$

است.

برهان. فرض کنیم $X_۱$ و $X_۲$ و ... X_n و ⋯ دنباله‌ای از متغیرهای تصادفی (i.i.d.) برنولی با

$$P = \{X_i = ۱\} = x,\ P\{X_i = ۰\} = ۱ - x$$

است و $S_n = X_۱ + X_۲ + \cdots + X_n$، آنگاه

$$\begin{aligned} E(f(\tfrac{S_n}{n})) &= \sum_{k=۰}^{n} f(\tfrac{S_n}{n}) P\{S_n = k\} \\ &= \sum_{k=۰}^{n} f(k/n) C_n^k x^k (۱ - x)^{n-k} \\ &= B_n(x). \end{aligned}$$

چون تابع f روی $[۰, ۱]$ پیوسته است، پیوسته یکنواخت است و بنابراین برای هر $t > ۰$، $\delta > ۰$ وجود دارد طوری که

$$|x - y| \le \delta \Rightarrow |f(x) - f(y)| \le t$$

اما تابع f چون پیوسته است بنابراین کراندار است در نتیجه یک M چنان وجود دارد که برای هر x

$$|f(x)| \le M < \infty$$

با به‌کار بردن این نامساوی داریم

$$\begin{aligned}
|f(x)-B_n(x)| &= |\sum_{k=0}^{n} f(x)C_n^k(1-x)^{n-k} \\
&\quad -\sum_{k=0}^{n} f(k/n)C_n^k x^k(1-x)^{n-k}| \\
&= |\sum_{k=0}^{n}(f(x)-f(k/n))C_n^k x^k(1-x)^{n-k}| \\
&\le \sum_{k=0}^{n}|f(x)-f(k/n)|C_n^k x^k(1-x)^{n-k} \\
&= \sum_{\{k:|k/n-x|\le\delta\}}|f(x)-f(k/n)|C_n^k x^k(1-x)^{n-k} \\
&+ \sum_{\{k:|k/n-x|>\delta\}}|f(x)-f(k/n)|C_n^k x^k(1-x)^{n-k} \\
&\le \epsilon\sum_{k=0}^{n} C_n^k x^k(1-x)^{n-k} \\
&+ 2M\sum_{\{k:|k/n-x|>\delta\}} C_n^k x^k(1-x)^{n-k}
\end{aligned}$$

حال چون برای $n\ge 1$، $0\le k\le n$، جرم احتمال در نقطه k احتمال دوجمله‌ای

$$P_n(k)=C_n^k x^k(1-x)^{n-k}$$

است، بنابراین

$$\sum_{\{k:|k/n-x|>\delta\}} P_n(k)=P\{|\frac{S_n}{n}-x|>\delta\}$$

در نتیجه

$$|f(x)-B_n(x)|\le\epsilon+2MP\{|\frac{S_n}{n}-x|>\delta\} \qquad (*)$$

اما چون برای هر متغیر تصادفی ξ داریم

$$P\{|\xi-E(\xi)|\ge\epsilon\}\le\frac{V(\xi)}{\epsilon^2}$$

و در حالت $\xi=\frac{S_n}{n}$، $E(\xi)=x$ و

$$Var(\frac{S_n}{n})=\frac{Var(S_n)}{n^2}=\frac{nx(1-x)}{n^2}=\frac{x(1-x)}{n}$$

بنابراین

$$P\{|\frac{S_n}{n}-x|>\delta\}\le\frac{Var(\frac{S_n}{n})}{\delta^2}=\frac{x(1-x)}{n\delta^2}\le\frac{1}{4n\delta^2}$$

درنتیجه داریم

$$\begin{aligned} |f(x) - B - n(x)| &\leq \epsilon + ۲M P\{|\frac{S_n}{n} - x| > \delta\} \\ &\leq \epsilon + ۲M \frac{۱}{۴n\delta^۲} \\ &= \epsilon + \frac{M}{۲n\delta^۲} \end{aligned}$$

و بنابراین

$$\lim_{n\to\infty} \max_{۰\leq x\leq ۱} |f(x) - B_n(x)| = ۰$$

که نتیجه قضیه وایرشتراس است.

یکی از کاربردهای جالب احتمالات در آنالیز قضیه (Szegö) است برای اثبات آن می‌توان به (Adams, Guiliemin صفحه ۱۶۷) رجوع کرد.

یکی از کاربردهای مهم احتمالات در آنالیز کلاسیک اثبات قضیه رادن‌ـنیکودیم به‌وسیله مارتینگل‌ها است. چون ما معمولاً احتمال و امید شرطی را به‌وسیله قضیه رادن‌ـنیکودیم تعریف می‌کنیم این نظریه مارتینگل‌ها است که بر اساس این قضیه استوار می‌گردد. اما اثبات قضیه رادن‌ـنیکودیم به‌وسیله مارتینگل‌ها چندان هم غیرمنتظره نیست، برای دیدن این قضیه رجوع کنید به ظهوری زنگنه، نظریه فرآیندهای تصادفی .

آخرین مسئله دترمینیستیک که ما می‌خواهیم در این سخنرانی به آن بپردازیم مسئله دیریشله است. این مسئله یکی از مسئله‌های مهم نظریه معادلات دیفرانسیل پاره‌ای بیضوی است و نظریه پتانسیل دیریشله به‌حساب می آید.

مسئله دیریشله

فرض کنیم D یک میدان در $\mathcal{R}^۲$ باشد. فرض کنیم تابع f روی مرز D تعریف شده باشد می‌خواهیم ثابت کنیم یک تابع هارمونیک u روی میدان D چنان موجود است که $u|_{\partial D} = f$.

اثبات وجود این مسئله با روش‌های دترمینیستیک پیچیده است و تنها وجود جواب می‌توان را ثابت کرد. در صورتی که در راه‌حل احتمالاتی که ارائه می‌دهیم نه تنها وجود جواب را ثابت می‌کنیم بلکه فرمولی برای جواب نیز ارائه می‌دهیم. اثبات با روشهای احتمالاتی این قضیه بر اساس آنالیز تصادفی و خواص حرکت برونی است. در بند بعد ما به‌طور شهودی و توصیفی درباره حرکت برونی و خواص آن صحبت می‌کنیم و سپس مسئله دیریشله را ثابت می‌کنیم. برای مطالعه عمیق حرکت برونی و خواص آن رجوع کنید به Revuz & Yor, Karatzas.

حرکت برونی و مسئله دیریشله

حرکت برونی نامی است که به حرکت نامنظم گرده گیاهان که در آب معلق هستند داده شده است. رابرت براون گیاه‌شناس معروف انگلیسی برای اولین بار در ۱۸۲۸ با مشاهده این حرکت، متوجه اهمیت آن در مطالعه ذرات معلق میکروسکوپی شد. پس از آن دامنه کاربرد حرکت برونی از مطالعه ذرات معلق میکروسکوپی بسیار فراتر رفته است و شامل مدل‌سازی قیمتهای سهام، نوفه حرارتی در مدارهای الکتریکی، برخی حالتهای حدی در سیستمهای صف و موجودی و اختلالات تصادفی در انواع دیگر از سیستمهای فیزیکی، زیستی، اقتصادی و مدیریت شده است.

آنچه براون در ابتدا مشاهده نمود این بود که گرده‌های گیاهان درون مایع دارای حرکت‌اند و علاقه‌مند شد تا قانون و علت این حرکت را بیابد، اما از عهده این کار برنیامد و مساله بدون پاسخ باقی ماند. سپس در سال ۱۹۰۶ میلادی، انیشتین موفق به حل مسئله شد و علت حرکت را بمباران دانه‌های گرده توسط ملکولهای مایع معرفی نمود. با این حال اولین مدل ریاضی حرکت برونی در تز دکتری ریاضی بشلیه (Bachelier) در سال ۱۹۰۰ میلادی در دانشگاه پاریس و برای مدل اقتصادی مطرح شد.

بشلیه توزیعهای مهم متعددی استخراج کرده بود که همگی به فرآیند حرکت براونی در $\mathcal{R}$ مربوط بودند، «از جمله توزیع مربوط به تغییر بیشینه در طول یک بازه زمانی. بدین منظور وی توزیعهای متناظر با یک قدم زدن تصادفی گسسته را پیدا می‌کرد و سپس حد را هنگامی که طول قدمها به سمت صفر میل می‌کرد به‌دست می‌آورد. دقیقتر بگویم، آنچه بشلیه استخراج نمود توزیعهایی بودند که برای فرآیند حرکت برونی کارایی داشتند، به فرض آنکه اصلاً چیزی تحت عنوان حرکت برونی وجود داشته باشد، و به فرض اینکه بشود آن را با آن قدم زدنهای تصادفی تقریب زد.» (دوب، نشر ریاضی ۱۲)

پس از آن نوربرت وینر ریاضیدان برجسته و نابغه قرن بیستم در سال ۱۹۱۸ مدل ریاضی این حرکت را به‌طور کامل بررسی کرد. «توجه کنید که شکی در وجود حرکت برونی نیست: حرکت براونی را می‌شود زیر میکروسکوپ نظاره کرد. ولی هنوز برهانی برای وجود یک فرآیند تصادفی، یک حساب ریاضی، با خواص مطلوب در دست نبود. وینر (۱۹۲۳) فرآیند مطلوب حرکت برونی را که امروزه گاه فرآیند وینر نامیده می‌شود ساخت. بدین منظور وی از رهیافت دانیل به نظریه اندازه استفاده کرد تا اندازه‌ای با خواص ذیل برفضای S از توابع پیوسته به دست آورد: اگر $x(t,.)$ متغیری تصادفی باشد که با مقداریک تابع در S در زمان t تعریف شده باشد، فرآیند تصادفی این متغیرهای تصادفی فرآیندی تصادفی است با اعضای S به‌عنوان توابع نمونه‌ای، و با توزیعهای توأمی که

برای فرآیند حرکت برونی داشتیم به‌عنوان توزیعهای توأم متغیر تصادفی» (دوب، نشر ریاضی ۱۲)

ریاضیدانان زحمت برای تولید یک نظریه عمیق ریاضی می‌کشید اما متأسفانه آنچه باقی می‌ماند نتیجه این نظریه به صورت یک مقاله است، همانطور که ۷۵ ریاضیدان در سال ۱۹۶۲ در بیانیه‌ای در مورد تدریس ریاضی اعلام کردند «تفکر ریاضی تنها استدلال استنتاجی نیست، همچنین اثبات صورت صرف هم نمی‌باشد. فرآیندهای ذهنی و فکری که اثبات و چگونگی اثبات را ارائه می‌کنند همانند خود اثبات که نتیجه تفکر ریاضی است بخشی از تفکر ریاضی محسوب می‌شود. استخراج مفاهیم درست از وضعیت‌های محسوس و ملموس، تعمیم از حالات شهود، استدلال استقرایی، استدلال از طریق تمثیل، زمینه‌های شهودی که برای آشکار کردن یک حدسیه به کار می‌روند همگی سبک و طریقه ریاضی گونهٔ تفکر است.» (دوب، نشر ریاضی ۱۲) خوشبختانه نوربرت وینر با نوشتن کتابهای توصیفی مانند «من یک ریاضیدان هستم» پشت پردهٔ تفکر و فرآیند به‌وجود آوردن یک نظریه را تا اندازه‌ای بررسی کرده است، بنابراین داستان حرکت برونی را از زبان نوربرت وینر می‌شنویم. وینر در .M.I.T استخدام شده بود «ساختمان .M.I.T در ساحل رودخانهٔ چارلز ساخته شده بود و طوری قرار داشت که می‌شد مستقیماً و از پنجره‌های آن، از چشم انداز گسترده سرزمین زیبای دور و بر آن لذت برد، به‌خصوص وجود رودخانه، موجب شادی بود. به‌نظر می‌رسید که می‌توان از بام تا شام به تماشای ناز و کرشمه‌های عجیب و غریب آب نشست. ولی آن چه در میان این همه زیبایی مرا به طرف خود می‌کشید، ریاضیات و فیزیک بود. آن قانون‌مندی‌های ریاضی، که همه این توده بی‌نظم و ناآرام آب را هدایت می‌کند، کدام است؟ مگر اهمیت اصلی ریاضیات در این نیست که می‌تواند نظم و ترتیبی را که زیر این هرج و مرج و نابسامانی ظاهر دور و بر ما پنهان شده است، پیدا کند؟ رودخانه چارلز، گاهی ناگهان از موج‌های بلند، با شانه‌های بلندکف، پوشیده می‌شود و گاه چنان چین خوردگی ملایمی دارد که به زحمت می‌توان موج‌های کوتاه آن را دید. طول موج‌های آن، گاه از دو یا سه بند انگشت تجاوز نمی‌کند و گاه به چند متر می‌رسد. چگونه می‌توان بیان ریاضی همه این پدیده‌ها را داد؟ از چه دستگاهی باید استفاده کنیم تا در تنوع بی‌پایان جزئیات این منظره غرق شویم؟ برایم روشن بود که این مسئله، با مسئله میانگین آماری بستگی دارد که با انتگرال لبگ خویشاوند است.» (وینر،انتشارات فاطمی، صفحهٔ ۴۱).

وینر در کتاب خود به آشنایی به آثار ویلارد «ویلارد گیبس، اشاره می‌کند: یکی از بزرگ‌ترین دانشمندان آمریکایی است که در واقع رشته تازه‌ای از دانش را پایه گذاشت، رشته‌ای که در حد فاصل فیزیک و ریاضیات قرار دارد» (وینر،انتشارات فاطمی، صفحه

۴۲). «گیبس آثار بسیار جالبی هم در فیزیک و هم در ریاضیات دارد، ولی کارهای اساسی او در زمینه مکانیک آماری، بیش از هر چیز دیگری، برایم جالب بود، همین کارهای او بود که، تا حد زیادی، مسیر خاص زندگی مرا مشخص کرد.»

دیدگاه سنتی در فیزیک، که از نیوتون بزرگ سرچشمه می‌گیرد، بستگی خلل‌ناپذیری با تصورهای دترمینیستیکی دارد و، بر طبق آن، معرفت دقیق چگونگی جهان و یا هر قسمت بسته‌ای از آن در یک لحظه معین، شامل معرفت دقیق آن در زمان‌های بعدی هم می‌باشد. بنابر تصور اصلی نیوتون، اگر موقعیت و سرعت ذره‌ها را در موج‌های سطح رودخانه چارلز بدانیم، می‌توان حرکت این موج‌ها را در همه سده‌های آینده محاسبه کنیم. متاسفانه با وسیله‌های اندازه‌گیری که در اختیار داریم، و همه آن با دستهای آدمی ساخته شده‌اند، نمی‌توانیم مقادیر مطلقاً دقیق و سرعت همه ذره‌ها را، در لحظه اولین زمان، به‌دست آوریم. به این ترتیب، فیزیک، که باید عملاً به درد پدیده‌های طبیعت بخورد، ناگزیر مواجه با مشکلی می‌شود: چگونه می‌توان با تکیه بر این داده‌های تقریبی مربوط به وضع اولیه، که به کمک وسیله‌های موجود به دست می‌آیند، درباره واقع امر قضاوت کرد؟» (وینر،انتشارات فاطمی، صفحه ۴۳).

با این ایده وینر فضای نمونه‌ای را فضای توابع پیوسته یا در واقع فضایی که هر عضو آن یک موج باشد در نظر گرفت. متغیر تصادفی از فضای توابع پیوسته $\Omega = C[0, T]$ به اعداد حقیقی تعریف شده بود. دراینجا وینر مفهوم مهمی را کشف کرده بود، توابعی که روی فضایی تعریف می‌شوند که دامنه آن نقطه (در فضای چندبعدی) نیست بلکه دامنه آن یک فضای بینهایت بعدی است. خود می‌گوید «به تعمیم مفهوم احتمال، در مواردی مربوط می‌شود که «حالت‌های ممکن» را نمی‌توان به‌صورت نقطه‌های یک صفحه یا حوزه‌ای از فضا در نظر گفت، ولی خصلت منحنی‌های را دارند که معرف اشیای متحرکی هستند.» (وینر،انتشارات فاطمی، صفحه ۴۴).

به این ترتیب، حرکت برونی موقعیتی را در برابر ما قرار می‌دهد که در آن، ذره‌ها به رسم منحنی‌هایی مشغول‌اند، و این منحنی‌ها، به مجموعه‌ای آماری از منحنی‌ها تعلق دارند. «این حرکت، بهترین زمینه برای اندیشه‌های من در مورد به‌کار بردن انتگرال‌گیری لبگ در فضای منحنی‌ها بود. و ضمناً دارای این خصوصیت بود که موضوع آن، از لحاظ فیزیکی، به دنیای واقع مربوط می‌شد و دقیقاً به اندیشه‌های گیبس بستگی داشت. در واقع، در این جا بود که توانستم با به‌کار بردن نظرات خود در تعمیم نظریه انتگرال‌گیری، به موفقیت بزرگی برسم. خود حرکت برونی، موضوعی نبود که در فیزیک، بدون بررسی باقی مانده باشد. ولی در کارهای اساسی و عمیقی که اینشتین و سمولوخوفسکی در این زمینه کرده‌اند، یا به رفتار یک ذره در یک لحظه زمانی ثابت پرداخته‌اند و یا به خصلت‌های

آماری مجموعه بزرگی از ذره‌ها در جریان زمان: ولی خاصیت‌های ریاضی خط سیر ذره‌های جداگانه، هیچ‌گاه مورد مطالعه قرار نگرفته بود.» در مورد موضوع اخیر، تقریباً هیچ چیز روشن نبود، البته اگر اظهار نظر عمیق یه رن فیزیک دان فرانسوی، را که در کتاب خود به‌نام «اتم‌ها» آورده است، به‌حساب نیاوریم. او می‌گوید» خط سیر به‌کلی بی‌نظم ذره‌ها، که در اثر حرکت برونی به وجود می‌آید، آدمی را به یاد منحنی‌های پیوسته ریاضی دانان می‌اندازد که در هیچ نقطه خود مشتق نداشته باشند» (وینر،انتشارات فاطمی، صفحه ۴۸ و ۴۹).

«با کمال تعجب، و در عین حال لذت، دریافتم که با چنین درکی از حرکت براونی، می‌توان نظریه صوری آن را در حد بالایی از کمال و ظرافت تنظیم کرد. در چارچوب این نظریه توانستم ملاحظه یه‌رن را ثابت کنم که، به استثنای چند موردی که احتمال آن را در مجموع برابر صفر است، مسیرهای حرکت براونی، منحنی‌های پیوسته‌اند، که در هیچ کجا مشتق ندارند» (وینر،انتشارات فاطمی، صفحه ۴۹).

بدین ترتیب وینر در سال (۱۹۲۳) از خاصیت گاوسی بودن حرکت برونی استفاده می‌کند و حرکت برونی را به‌صورت یک سری (فوریه) با پایه دنباله‌های توابع شاودر، و ضرایب متغیرهای تصادفی گوسی می‌سازد. حرکت برونی بعدها با اصلاحاتی به‌وسیله پل لوی در سال (۱۹۴۸) و چیشلسکی (۱۹۶۱) به‌صورت ساده تری ساخته می‌شود (برای این روش رجوع کنید به Lamperti 1966). حرکت برونی یک فرآیند تصادفی است که دارای خاصیت مارکفی، گوسی، مارتینگلی، با نموهای ایستا و مستقل است. دارای مسیرهای پیوسته‌ای است که در هیچ نقطه‌ای مشتق ندارد.

حال به حل مسئله دیریشله می‌پردازیم. با توجه به این واقعیت که حرکت براونی دارای خاصیت قوی مارکف نیز می‌باشد، یعنی اگر B_t یک حرکت براونی و T یک زمان توقف باشد آنگاه فرآیند جدید $B_{T+t} - B_T$ یک حرکت براونی جدید است. حرکت براونی یک مارتینگل موضعی است یعنی اگر T یک زمان توقف باشد آنگاه $B_{T\wedge t}$ (که در اینجا $T\wedge t$ می‌نیمم t و T است) یک مارتینگل است.

فرض کنیم D یک مجموعه فشرده نسبی باشد، یعنی $\bar{D}$ فشرده باشد. قرار می‌دهیم

$$\tau_D = \inf\{t : B_t \in D^c\}$$

یعنی زمان برخورد حرکت برونی با مرز مجموعه D یا اولین زمانی که حرکت برونی می‌خواهد از مجموعه D خارج شود. می‌توان ثابت کرد که τ_D یک زمان توقف است، توجه شود که τ_D متناهی است. یعنی حرکت براونی با احتمال 1 از مجموعه D خارج می‌شود. اگر f تابعی باشد که روی D تعریف شده باشد آنگاه $f(B_t)$ فقط برای $t < \tau_D$ تعریف

شده است.

قضیه: فرض کنیم f یک تابع هارمونیک در D باشد (یعنی $\triangle f = ۰$ در D) و فرض کنیم $B_۰ = x \in D$ (یعنی حرکت براونی در لحظه صفر در نقطه x باشد) آنگاه $f(B_t)$ یک مارتینگل موضعی است.

برهان: با استفاده از آنالیز تصادفی و فرمول ایتو Karatzas, Oksendal, Revuz & Yor] و ظهوری زنگنه] داریم

$$f(B_t) = f(x) + \int_۰^t \bigtriangledown f(B_s) dB_s + \frac{۱}{۲} \int_۰^t \triangle f(B_s) ds$$

چون f هارمونیک است، $\triangle f = ۰$ و از آنجا

$$f(B_t) = f(x) + \int_۰^t \bigtriangledown f(B_s) dB_s$$

چون انتگرال تصادفی یک مارتینگل است، نتیجه برقرار است .Q.E.D

قضیه (مسئله دیریشله)

فرض کنیم D یک میدان باشد که در آن برای تمام xها

$$P^x(\tau_D < +\infty) = ۱$$

و گیریم f یک تابع مثبت برل اندازه‌پذیر باشد. تابع $h(x) := E^x(f(B_{\tau_D}))$ را در نظر می‌گیریم، اگر x موجود باشد که $h(x) < +\infty$ آنگاه h روی D هارمونیک است.

تبصره ۱. توجه کنید P^x و E^x، احتمال و امید شرطی به شرط $\{B_۰ = x\}$ (یعنی حرکت برونی در لحظه صفر در نقطه x است) باشد.

تبصره ۲. در اینجا فرمول صریح $h(x) := E^x(f(B_{\tau_D}))$ حل مسئله دیریشله را به ما می‌دهد.

برهان. گیریم $x \in D$ و گیریم V گویی به مرکز x باشد که $\bar{V} \subset D$ قرار می‌دهیم

$$\tau_v = \inf\{t : B_t \in \partial V\}$$

اگر $x = B_۰$، آنگاه $\tau_v < \tau_D$. لذا

$$\begin{aligned} E^x(f(B_{\tau_D})) &= E^x(E(f(B_{\tau_D})|\mathcal{F}_{\tau_v})) \\ &= E^x(E(f(B_{\tau_D})|B_{\tau_v})). \\ &= h(B_{\tau_v}) \end{aligned}$$

بنابراین $h(x) = E^x(h(B_{\tau_v}))$. اما حرکت براونی نسبت به دوران ناوردا است، لذا B_{τ_v} روی ∂V یکنواخت است. حال

$$h(x) = \frac{۱}{|\partial V|} \int_{\partial V} h(y)dy$$

درنتیجه برای هر گوی به مرکز x که در D واقع شود، اگر h متناهی یا نامتناهی باشد

$$h(x) = \frac{۱}{|V|} \int_{V} h(y)dy$$

پس $h(x)$ متناهی است اگر و فقط اگر h در یک همسایگی x متناهی باشد، که نشان می‌دهد مجموعهٔ $\{x : h(x) < +\infty\}$ باز است. از همبند بودن D نتیجه می‌شود که اگر x یی در D موجود باشد که $h(x) < +\infty$، آنگاه برای هر x، $h(x) < +\infty$، لذا h هارمونیک است.

مراجع

Adams, M., Guiliemin, V., Measure Theory and Probability.

Bass, R.F. Probabilistic Techniques in Analysis, Springer-Verlag 1995.

Billingsley, P., Probability and Measure, New York: Wiley.

Breiman, Probability, Siam 1992.

Chung, K.L., A Course in Probability Theory, Acadamic Press, 1977.

Kac, M., Statistical Independence in Probability, Analysis and Number Theory. Mathematical Association of America (Carus Mathematical Monograph, no 12) 1959.

Karatzas, Shreve E., Brownian Motion and Stochastic Cal culus Springer-Verlag New York 1988.

Lamperti, J., probability W.A. Benjamin Inc 1966.

OKsendal, B., Stochastic Differential Equations Springer-Verlag, Berlin.

Revuz, D., Yor, M., Continuous Martingales and Brownlan Motion Springer Berlin 1992.

در باب برنامه‌ریزی درسی دبیرستان، ترجمه جواد حاجی‌بابایی، رشد آموزش ریاضی شماره ۴۶، ۱۳۷۴.

دوب، جوزف (ترجمه عطاءالله تقاء) سیر پیدایش دقت در احتمال ریاضی (۱۹۰۰-۱۹۵۰)، نشر ریاضی، سال ۱۲، شماره ۱ و ۲.

ظهوری زنگنه، بیژن، نظریه فرآیندهای تصادفی، جزوات درسی، دانشگاه صنعتی شریف.

ظهوری زنگنه، بیژن، آنالیز تصادفی، جزوات درسی، دانشگاه صنعتی شریف.

مامفرد، دیوید، ترجمه شاپور اعتماد، طلوع عصر روشهای تصادفی، نشر ریاضی سال ۱۳، شماره ۱.

وینر، نوربرت، ترجمهٔ پرویز شهریاری، من یک ریاضیدانم، انتشارات فاطمی، خرداد ۱۳۶۸.

توسعه روش سری زمانی فازی و ارائه یک مورد کاوی

رسول نورالسناء، عباس سقائی

$P۱۱۰۱۸$

دانشکده صنایع، دانشگاه علم و صنعت

چکیده: پس از ارائه مجموعه های فازی توسط دکتر زاده،کاربردهای متفاوتی برای آن تعریف گردیده است. پیش بینی داده هایی که ذاتاً قطعی نبوده باعث گردید تا اصول سری زمانی فازی برای پیش بینی فرایندهای پویایی که مشاهداتی زبانی دارند، توسط محققین‌مطرح گردد. روش ارائه شده مورد تحلیل آماردانان قرارگرفت و ایراداتی بر آن واردشد. دراین مقاله سعی شده است تا باتوجه به ابهاماتی که روش مذکور داشته است، الگویی براساس تلفیق مدلهای سری زمانی وابسته به زمان و مستقل از زمان برای مشاهدات زبانی مطرح گردد که بتواند ابهامات مربوطه را مرتفع نماید. دراین مدل از شبکه های عصبی مصنوعی نیزاستفاده گردیده است.بعلاوه جایگاه کاربردی مناسبی برای این تکنیک مشخص گردیده که درآن حیطه یک موردکاوی نیز انجام پذیرفته است.

واژه‌های کلیدی: سری زمانی، مجموعه‌های فازی، داده‌های زبانی، شبکه‌های عصبی.

۱ مقدمه

تئوری فازی درسال ۱۹٦۵ برای اولین بارتوسط زاده(۱۹٦۵) طی مقاله‌ایی به‌نام «مجموعه های فازی» به عنوان ابزاری نوین برای مدلسازی عدم قطعیت معرفی شد. ایده مجموعه های فازی پاسخ ذهن خلاق زاده به ناتوانی روشهای موجود درتحلیل رفتارسیستمهای پیچیده و خصوصاً سیستمهای جاندار و طبیعی بود. درسال ۱۹٦۲ چیزی رابدین مضمون برای سیستم های بیولوژیک نوشت « مااساساً به نوع جدیدی ریاضیات نیازمندیم، ریاضیات مقادیرمبهم یا فازی که توسط توزیع احتمالات قابل توصیف نیستند» پس ازآن وی ایده اش را درمقالهٔ «مجموعه های فازی تجسم بخشید.باپیدایش تئوری فازی، بحث و جدلها پیرامون آن نیز آغاز گردید.بعضی‌ها آنرا تایید نموده و کار روی این زمینهٔ جدید را شروع کردند و برخی دیگر آن را برخلاف اصول علمی می دانستند.بسیاری از مفاهیم بنیادین تئوری فازی به وسیلهٔ زاده دراواخر دههٔ ٦۰ و اوایل دههٔ ۷۰ میلادی مطرح گردید.نظریهٔ مجموعه های فازی که شاید بتوان گفت به انگیزهٔ ایجاد ابزاری برای مدلسازی عدم قطعیت درسیستم های پیچیده معرفی شده به تبیین نوعی دیگر از عدم قطعیت پرداخت

که شاید بهترین لغت برای توصیف آن «ابهام» باشد.

یکی از بخشهای آمار پیش بینی است.وجود هرگونه عدم قطعیت موجب شد تاتکنیکهای زیادی برای پیش بینی تدوین گردد. این علم به عنوان یکی از شاخه های اصلی در آمار کاربردی بطور وسیعی در سایر علوم مورد استفاده قرار گیرد.در این بین تجزیه و تحلیل سری های زمانی نقش بسیار مهمی را ایفا می کند. یک سری زمانی را مجموعه ایی از مشاهدات می دانیم که هر یک متعلق به زمانی مشخص بوده و اغلب فاصلهٔ بین زمان مشاهدات متوالی برابر می باشند.که این مقادیر با $Y_1, Y_2, ...$ (از متغییر تصادفی Y) در زمانهای $t_1, t_2, ...$ نشان داده می شوند. لذا Y تابعی از t بوده و آنرا به صورت $Y=F(t)$ مشخص می کنند. آنچه دراین مقاله مورد توجه قرار گرفته « سری زمانی فازی » است. بطوری که مشاهده می شود این عبارت در ارائه مدلها و روشهای مختلف مورد استفاده قرار گرفته است. در این مقاله سعی داریم تا تحقیقات موجود را که با عنوان « سری زمانی فازی » توام بوده‌اند را مشخص نموده و بر حسب موضوع مربوطه تقسیم بندی نماییم.و به تشریح هریک بپردازیم.در بخش دوم مقاله سری زمانی فازی با داده های قطعی را بررسی می کنیم.و سری زمانی فازی با داده های زبانی [1] را در بخش سوم مورد بررسی قرار می دهیم در این بخش روشی را توسعه داده و الگوریتمی برای آن پیشنهاد نموده‌ایم.در بخش چهارم مثالی را برای تشریح بیشتر روش پیشنهادی مطرح نموده و در خاتمه نتیجه گیری نموده‌ایم.

۲ سری زمانی فازی با داده های قطعی

روشهایی که دراین بخش مورد توجه قرار می دهیم، روشهای است که داده های سری زمانی، مقادیری قطعی هستند که به توسط مدلسازی فازی و سری زمانی مورد تحلیل قراراگرفته و پیش بینی مقادیر را ممکن نموده است. باید توجه کرد که مفاهیم سری زمانی فازی پس از توسعهٔ رگرسیون فازی مطرح شده است. دراین بین تاناکا(۱۹۹۲،۱۹۸۷،۱۹۸۲)رگرسیون فازی را مورد بررسی قرارداد و روشی را توسعه داد که محیط فازی را جایگزین خطای مدل نمود.این مدل پیش بینی را به توسط فاصله‌ایی نشان می داد.یکی از ضعفهای این روش زمانی است که نقاط دورافتاده در مشاهدات وجود داشته باشند، وموجب وسیع شدن بیش از حد فاصلهٔ ارائه شده گردد. یکی از اولین روشهای کاربردی پیش بینی فازی تحت عنوان رگرسیون فازی توسط واتادا(۱۹۹۲)مطرح شد.دراین روش فاصله‌ایی ،امکان رخداد را نشان می‌دهد. این مدل براساس مدل رگرسیون امکان تدوین شده و یکی از ضعفهای آن این است که وزن توابع هدف بطور ذهنی تعیین می گردد. چن(۱۹۹۶)،

[1] Linguistic Data

روش سری زمانی فازی را براساس مطالب سونگ و چیسوم (که در بخش بعد به تفضیل مطرح شده است) بنا نهاد. داده های مورد تحلیل قطعی بوده و از رابطهٔ فازی برای تشریح رابطهٔ بین داده ها استفاده گردید. این روش قادر بود پیش بینی را براساس یک مقدار (نه بصورت یک فاصله) ارائه کند. یکی از جدیدترین روشهای سری زمانی فازی را تسنگ و همکارانش (۲۰۰۱) ارائه نمودند. سریهای زمانی ARIMA که توسط باکس و جنکینز(۱۹۷۶)ارایه شده‌اند به عنوان تکنیکهای کاربردی در علوم اجتماعی، اقتصادی، مهندسی و... از جایگاه ویژه‌ای برخوردار است. تحقیقات تسنگ و همکارانش بر سری زمانی فازی که داده های قطعی را در مدلهای ARIMA مورد بررسی قرار می‌دهند معطوف گردیده است.علاوه براین در این روش دو مقدار «بهترین امکان» و « بدترین امکان »برای پیش بینی ارائه می‌گردد.و درنتیجه فاصله‌ای برای امکان رخداد مشخص می شود. ایشان این روش را FARIMA [۲] نامیدند.

یکی از دلایل اصلی که‌باعث شده است تا مدلهای سری زمانی فازی با داده های قطعی مورد توجه قرار گیرند، آن است که مدلهای سری زمانی (مانند ARIMA) جهت بررسی و پیش بینی به ۵۰ یا ۱۰۰ مشاهده برای تعیین مناسب مدل و پیش بینی مقادیر آن نیاز دارند. ولی درسری زمانی فازی، به تعداد کمتری مشاهده برای دقتی یکسان نیاز داریم. خلاصه روشهای مورد بحث در جدول شمارهٔ۱ آمده‌است.

جدول ۱: مقایسهٔ روشهای سری زمانی با داده های قطعی

Watada	*Chen*	*FARIMA*	*ARIMA*	*Compare*
ندارد	ندارد	دارد	دارد	تطابق‌بامدل باکس‌وجنکینز
تابع فازی	رابطه فازی	تابع فازی	تابعی‌براساس مشاهدات قبل	ارتباط‌ورودی وخروجی مدل
فاصلهٔ امکان	برآورد نقطه‌ای	فاصلهٔ امکان	فاصلهٔ اطمینان	چگونگی پیش بینی
کمتراز ARIMA	کمتراز ARIMA	کمتراز ARIMA	حداقل ۵۰نمونه برآوردخوب۱۰۰	تعدادنمونه‌های موردنیاز

[۲] Fuzzy ARIMA

۳ سری زمانی فازی با داده‌های زبانی

درزندگی روزمره اغلب کلماتی وجود دارند که برای توصیف متغیرها استفاده می شوند. به عنوان مثال هنگامی که می گوییم« امروز هوا سرد است.» ما از واژهٔ سرد استفاده می‌کنیم. واضح است که متغیر « دمای هوای امروز » می تواند مقادیری مانند ۱۰، ۰، ۷— یا ۲۰— را اختیار کند.هنگامی که متغیر، واژه‌ها را به عنوان مقدار می گیرد، در آنصورت چهارچوب مشخصی برای فرموله کردن آن در ریاضیات کلاسیک نداریم.برای ایجاد چنین چهارچوبی، مفهوم متغیرهای زبانی تدوین گردیده‌است.

۱.۳ نظریهٔ‌موضوع

یک متغیر زبانی بوسیلهٔ چهارپارامتر (X,T,U,M) مشخص می‌گردد.

X: نام متغیر زبانی / دمای هوای امروز

U: دامنه‌ای که متغیر درآن مقدار می‌گیرد / U=[-20,50]

T: مجموعه نامهایی متغیر زبانی است. / خیلی سرد،سرد،متعادل،گرم،خیلی گرم T=

M: یک قاعده است که هر مقدار زبانی در T را به یک مجموعهٔ فازی در U مرتبط می‌سازد و غالباً به وسیلهٔ توابع تعلق نشان داده‌می شود.

پیش بینی مقادیر زبانی به کمک مشاهدات گذشته می تواند کاربرد وسیعی داشته باشد. این موضوع توسط محققینی مورد توجه قرار گرفته، که از مهمترین آنها می توان به تحقیقات سونگ و چیسوم (۱۹۹۳) اشاره نمود. ایشان برای اولین بار سری زمانی فازی با داده های زبانی را مطرح نمودند و شرایط استفاده از این روش را؛ پویا بودن فرایند، فازی بودن مشاهدات، قابل تعریف بودن دامنهٔ مقادیر فازی در R^1 و غیر قابل حل بودن توسط مدلهای سنتی سری زمانی در این روش دو نوع مدل مطرح شده است که عبارتند از مدل پایدار در زمان و مدل ناپایدار در زمان برشمردند.و تعاریف و قضایایی نیز برای هریک ارائه شده که برخی از آن در ذیل آورده شده است (برای اطلاعات بیشتر، رجوع شود به سونگ و چیسوم(۱،۲—۱۹۹۳و۱۹۹۴))

تعریف ۱: فرض کنید Y(t) ، (t=...,0,1,2,...) زیر مجموعه‌ای از R^1 است، به عنوان دامنهٔ مجموعه‌های فازی $f_i(t)$ (i=1,2,...) تعریف گردد و F(t) دسته‌ای از(i=1,2,...) $f_i(t)$, باشد، لذا F(t) را سری زمانی برحسب Y(t) , (t=...,0,1,2,...) می‌نامند.

تعریف ۲: فرض کنید F(t) فقط توسط F(t-1) ایجاد می‌گردد.و به صورت $F(t-$

$F(t) \rightarrow$ (۱ ویا $F(t)= F(t-1) \circ R(t,t-1)$ نشان می‌دهیم.بطوری که $R(t,t-1)$ نشاندهندهٔ رابطهٔ فازی بین $F(t)$ و $F(t-1)$ می باشد و 'o' عملگر max-min است.

تعریف ۳: فرض کنید $F(t)$ فقط توسط $F(t-1)$ یاتوسط $F(t-1)$ یا $F(t-2)$ یا $(m >$ ...$F(t-m)$...) ایجاد می‌گردد. این رابطه فازی را می توان بصورت تساوی زیر نشان داد.

$$F(t)= (F(t-1) \cup F(t-2) \cup \ldots \cup F(t-m)) \circ R(t,t-1)$$

تعریف ۴: فرض کنید $R(t,t-1)$ یک مدل مرتبهٔ اول از $F(t)$ باشد. اگر به ازای هرt، $R(t,t-1)$ مستقل از t باشد، یعنی برای هر t تساوی $R(t,t-1)=R(t-1,t-2)$ برقرار باشد، آنگاه $F(t)$ یک سری زمانی مستقل از زمان و درغیر اینصورت یک سری زمانی وابسته به زمان نامیده می‌شود.

قضیهٔ ۱: $F(t)$ یک سری زمانی فازی است. اگر به ازای هرt تساوی $F(t)=F(t-1)$ برقرار باشد و همچنین $F(t)$ دارای تعداد عناصر محدود باشد آنگاه $F(t)$ یک سری زمانی مستقل از زمان نامیده می‌شود.

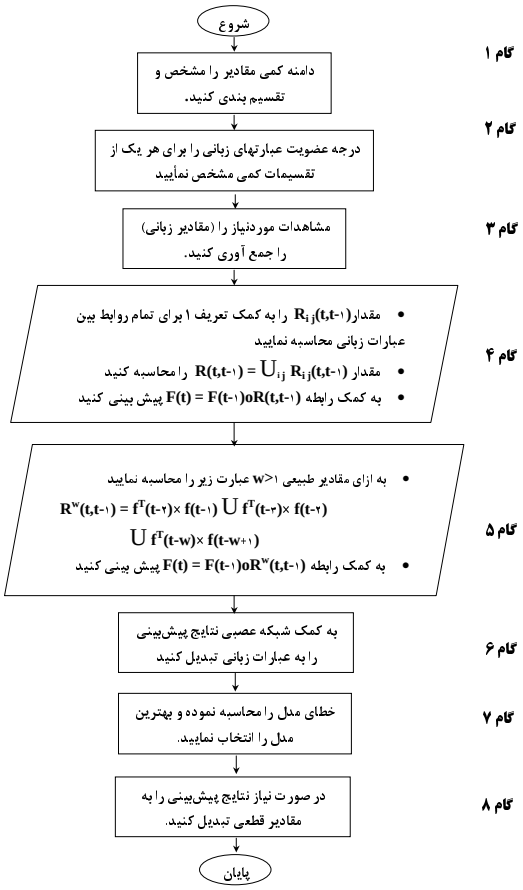
قضیهٔ ۲: اگر $F(t)$ یک سری زمانی فازی باشد، $F(t)=F(t-1)$ و برای هرt و $F(t)$ عناصر محدود $f_i(t)$, $(i=1,2,\ldots)$ وجود داشته باشد،خواهیم داشت:

$$R(t,t-۱) = \ldots f_{i۱}(t-۱) * f_{i۰}(t) U f_{i۲}(t-۲) * f_{i۱}(t-۱) U \ldots U f_{im}(t-m) * f_{im-۱}(t-m+۱) \ldots$$

بطوری که $m > ۰$ بوده و تمام زوجهای مجموعه‌های فازی متفاوت باشند.

این سری مقالات مورد بحث و توجه خاصی قرار گرفت که می توان به مقالهٔ سولیوان و وودال(۱۹۹۴) اشاره کرد که ایراداتی رانیزمنعکس نموده بودند. مباحثی که مطرح شده باعث شد تا اصلاحاتی را در روش سونگ و چیسوم پیشنهاد نماییم و الگوریتمی جدید را که درشکل ۱ آمده است توسعه دهیم. کلیات ایراداتی که بر آن گرفته شده بود و در روش پیشنهادی اصلاح گردیده است به شرح ذیل می‌باشد.

* اصرار محققین برای استفاده از این مدل جهت پیش بینی داده های قطعی کارایی این مدل را در مقایسه با روشهای آماری متعدد قرارداده بود.که در روش پیشنهادی ساختار مدل برای داده های فازی طراحی گردیده است.دو جنبه کاربردی برای آن پیشنهاد شده است که عبارتند از: الف)هنگامی که داده ها ذاتاً مقادیری از متغیر های زبانی هستند یا بدلیل سهولت بیشتر یا هزینه کمتر اندازه گیری مقادیر زبانی جایگزین مقادیر قطعی می‌گردند. ب) در بسیاری از مواقع بدلیل مشخص نبودن مقادیر دقیق پیشین یک متغیر، نمی توان مدلی برای پیش بینی ارائه نمود.در صورتی که شاید بتوان مقادیر قبل را بطور



شکل ۱: الگوریتم پیشنهادی

نادقیق بیان کرد.و به کمک مدل پیشنهادی مقادیر آتی را پیش بینی نموده و به تدریج مقادیر دقیق را جمع آوری و تکمیل نمود.

* تعریف توابع عضویت در مقالات یکسان نبوده‌اند. که‌این مورد نیز مرتفع گردیده‌است.

* برای انتخاب و تشخیص مدلهای ناپایدار و پایدار در زمان روش مناسبی پیشنهاد نگردیده بود که در روش پیشنهادی معیار انتخاب خطای کمترمدل تعریف گردیده است.

* شبکه عصبی مصنوعی استفاده شده بطور نامناسب بکار بسته شده است. که‌این مورد نیز مرتفع گردیده‌است.

۴ مثال

درکارخانه‌ای سوابق تولید و فروش درسنوات قبل ثبت شده ولی سوابق تقاضای محصول ثبت نگردیده است. و مدیریت اکنون به این نتیجه رسیده که برای برنامه‌ریزی تولید و کنترل موجودی نیازمند پیش بینی تقاضای سال آتی می‌باشد. مطابق مدلهای سنتی پیش بینی می بایست مشاهدات آتی تا چند دهه جمع‌آوری شده و پس از کفایت مشاهدات، مدل مناسب برازش گردد.ولی مطابق مدل سریِ زمانیِ فازیِ پیشنهاد شده در شکل ۱ ، گامهای اشاره شده در ذیل برداشته شد.

گام۱: ابتدا دامنهٔ تقاضای سالانه بصورت U=[13000,20000] تعریف گردید.و به نواحی زیر تقسیم شد.

$$U_۱=[13000,14000],$$
$$U_۲=[14000,15000],$$
$$U_۳=[15000,16000],$$
$$U_۴=[16000,17000],$$
$$U_۵=[17000,18000],$$
$$U_۶=[18000,19000],$$
$$U_۷=[19000,20000]$$

گام۲: عبارتهای های زبانی A با هفت مجموعهٔ فازی به عناوین؛

$A_۱=$(بسیاربسیار کم), $A_۲=$(بسیارکم), $A_۳=$(کم), $A_۴=$(متوسط),

$A_۵=$(زیاد), $A_۶=$(بسیار زیاد), $A_۷=$(بسیار بسیار زیاد)

در فضای U با درجهٔ عضویت های ذیل تعریف گردیدند.

$$A_۱ = \{U_۱/۱,\ U_۲/۰.۵,\ U_۳/۰,\ U_۴/۰,\ U_۵/۰,\ U_۶/۰,\ U_۷/۰\}$$
$$A_۲ = \{U_۱/۰.۵,\ U_۲/۱,\ U_۳/۰.۵,\ U_۴/۰,\ U_۵/۰,\ U_۶/۰,\ U_۷/۰\}$$
$$A_۳ = \{U_۱/۰,\ U_۲/۰.۵,\ U_۳/۱,\ U_۴/۰.۵,\ U_۵/۰,\ U_۶/۰,\ U_۷/۰\}$$
$$A_۴ = \{U_۱/۰,\ U_۲/۰,\ U_۳/۰.۵,\ U_۴/۱,\ U_۵/۰.۵,\ U_۶/۰,\ U_۷/۰\}$$
$$A_۵ = \{U_۱/۰,\ U_۲/۰,\ U_۳/۰,\ U_۴/۰.۵,\ U_۵/۱,\ U_۶/۰.۵,\ U_۷/۰\}$$
$$A_۶ = \{U_۱/۰,\ U_۲/۰,\ U_۳/۰,\ U_۴/۰,\ U_۵/۰.۵,\ U_۶/۱,\ U_۷/۰.۵\}$$

$$A_{۷} = \{U_{۱}/۰,\ U_{۲}/۰,\ U_{۳}/۰,\ U_{۴}/۰,\ U_{۵}/۰,\ U_{۶}/۰.۵,\ U_{۷}/۱\}$$

گام۳: در گام سوم بر اساس نظرات خبرگان کارخانه، میزان تقاضای سنوات قبل بطورذهنی و به صورت عبارتهای زبانی تعیین گردید که در جدول ۲ آورده شده است. روابط احصا شده از جدول ۲ عبارتند از:

$$A۱ \to A۱$$
$$A۱ \to A۲$$
$$A۲ \to A۳$$
$$A۳ \to A۳$$
$$A۳ \to A۴$$
$$A۴ \to A۳$$
$$A۴ \to A۶$$
$$A۶ \to A۱$$
$$A۷ \to A۷$$
$$A۷ \to A۶$$

سال	مقادیر زبانی	نتایج پیش بینی (درجات عضویت)							نتایج پیش بینی مقادیر زبانی
		U۱	U۲	U۳	U٤	U٥	U٦	U۷	
۱۳٦۰	A۱	..	..	..	..	..	..	..	..
۱۳٦۱	A۱	..	..	..	..	..	..	..	..
۱۳٦۲	A۱	..	..	..	..	..	..	..	..
۱۳٦۳	A۲	..	..	..	..	..	..	..	..
۱۳٦٤	A۲	..	..	..	..	..	..	..	..
۱۳٦٥	A۳	۰,٥	۰,٥	۰,٥	۰,٥				A۳
۱۳٦٦	A۳	۰,٥	۰,٥	۱	۰,٥				A۳
۱۳٦۷	A۳	۰,٥	۰,٥	۱	۰,٥				A۳
۱۳٦۸	A٤		۰,٥	۱	۰,٥				A٤
۱۳٦۹	A٤		۰,٥	۰,٥	۰,٥	۰,٥			A٤
۱۳۷۰	A٤		۰,٥	۰,٥	۱	۰,٥			A٤
۱۳۷۱	A۳		۰,٥	۰,٥	۱	۰,٥			A٤
۱۳۷۲	A۳		۰,٥	۰,٥	۱	۰,٥			A٤
۱۳۷۳	A۳		۰,٥	۱	۰,٥	۰,٥			A۳
۱۳۷٤	A۳		۰,٥	۱	۰,٥	۰,٥			A۳
۱۳۷٥	A۳		۰,٥	۱	۰,٥				A٤
۱۳۷٦	A٤		۰,٥	۱	۰,٥				A٤
۱۳۷۷	A٤		۰,٥	۰,٥	۰,٥	۰,٥			A٤
۱۳۷۸	A٦					۰,٥	۰,٥	۰,٥	A٦
۱۳۷۹	A۷					۰,٥	۱	۰,٥	A٦
۱۳۸۰	A۷					۰,٥	۰,٥	۰,٥	A٦
۱۳۸۱	A٦					۰,٥	۰,٥	۱	A٦
۱۳۸۲						۰,٥	۰,٥	۱	A٦

جدول ۲. نتایج پیش‌بینی مدل انتخاب شده

گام۴: در این مرحله از الگوریتم، برای هر رابطه فوق ماتریس $R_{ij}(t,t-۱)$ محاسبه و در نهایت $R(t,t-۱)$ بدست می آید. و در نهایت مطابق رابطه ارائه شده در شکل ۱، پیش بینی مقادیر به صورت درجات عضویت انجام پذیرفت.

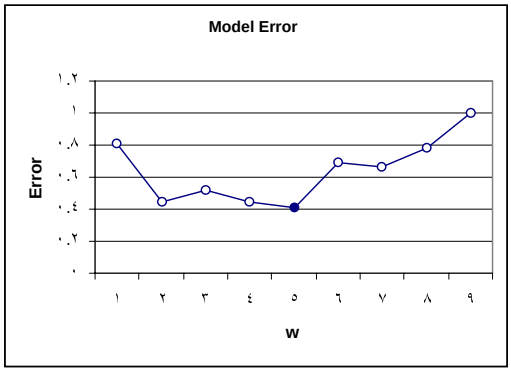
گام۵: مطابق رابطه ارائه شده در شکل ۱،به ازای مقادیر $w>۱$ رابطه $R^w(t,t-۱)$ محاسبه گردید و مقادیر پیش بینی به صورت درجات عضویت انجام پذیرفت.

گام٦: نتایج حاصل از هر دو روش گام ۴ و گام ۵ که به صورت مقادیر پیش بینی بر حسب درجات عضویت می باشند به کمک شبکه عصبی RBF[۱] به مقادیر زبانی تبدیل گردید. ورودی این شبکه عصبی از برداری با ۷ درایه که درجات عضویت تعریف شده درگام ۲ می باشند، تشکیل گردیده است و خروجی آنرا رده های متغیرهای زبانی تعریف شده درگام ۲ تشکیل می دهند.این شبکه به کمک نرم افزار MATLAB (5.3.1) آموزش دیده و جهت تبدیل درجات عضویت مقادیر پیش بینی شده به مقادیر زبانی مورد استفاده قرار گرفته است.

گام۷: در این گام خطای مدل پیش بینی محاسبه می گردد.بدین ترتیب که ابتدا مقدار

[۱] Radial Basis Function Neural Network

ران



شکل ۲: خطای مدل

رده پیش بینی بدست آمده از شبکه عصبی، به صورت عدد صحیح گرد می شود. سپس قدر مطلق اختلاف بین این مقادیر و مقادیر رده واقعی محاسبه می گردد. در نهایت با تقسیم مجموع کلیه مقادیر قدر مطلق بر تعداد پیش بینی انجام شده، میزان خطای مدل بدست می آید. بر اساس همین روش، مقدار خطا برای مدل گام ۴ و مدلهای گام ۵(به ازای w های مختلف) محاسبه گردیده است. نتایج محاسبات به صورت یک نمودار در شکل ۲ نشان داده شده است.همانطور که در شکل مشاهده می شود کمترین مقدار خطا متعلق به مدل گام ۵(به ازای ۵ =w) می باشد.نتایج عددی محاسبات مربوط به مدل مذکور شامل مقادیرزبانی واقعی، مقادیرزبانی پیش بینی شده و مقادیردرجات عضویت در جدول ۲ موجود می باشد.

گام۸: در صورت نیاز می توان مقادیر پیش بینی را به صورت مقادیر عددی قطعی نیز تبدیل نمود. برای این منظور می توان میانه بخشی از دامنه(که در گام ۱ معرفی گردیده است)، و دارای بالاترین درجه عضویت است را به عنوان مقدار پیش بینی قطعی در نظر گرفت.(به عنوان مثال مقدار ۱۹۵۰۰ برای سال ۱۳۸۲) و یا در روشی دیگر، میانگین وزنی درجات عضویت و تقسیمات دامنه را محاسبه نماییم.(مقدار ۱۸۷۵۰ برای سال ۱۳۸۲)

اکنون مدیریت کارخانه مقدار تقاضای سال ۱۳۸۲ را به ۳ شکل؛فازی ،عبارت زبانی و مقدار قطعی پیش بینی نموده است.

۵ نتیجه گیری

دراین مقاله توانایی مدلسازی فازی را درسریهای زمانی مورد بررسی قراردادیم و مشاهده نمودیم که در داده های قطعی می توان تعداد مشاهدات مورد نیاز را برای تخمین مدل وانجام پیش بینی در مقایسه با روشهای معمول آماری کاهش داد.به علاوه توانستیم با بررسی روشهای سری زمانی فازی که برای مشاهدات زبانی تعبیه گردیده است روشی را توسعه دهیم که با ترکیب مدلسازی فازی و شبکه های عصبی، قادر است با دریافت مشاهدات زبانی مقادیر پیش بینی را به صورت فازی،عبارت زبانی و مقادیر قطعی ارائه نماید.

مراجع

Cressie, N. (1993), *Statistics for Spatial Data*, Revised edition, John Wiley, New York.

David, M. (1977), *Geostatistical Ore Reserve Estimation*, Eevier Scintific Publishing Co., 364 P.

Box,G.P. Jenkins,G.M.(1976), *Forecasting and Control*, Holden-day Inc.,San Francisco, CA.

Chen,S.M. (1996), *Forecasting enrollments based on fuzzy time series*, Fuzzy Sets and System, vol.81(3), 311-319 P.

Tanaka,H. (1987), *Fuzzy data analysis by possibility linear models*, Fuzzy Sets and Systems, vol.24(3), 363-375 P.

Tanaka,H. Ishibuchi,H. (1992), *Possibility regression anaysis based on linear programming*, Fuzzy Regression Analysis, Omnitech Press, Heidelberg, 47-60 P.

Tanaka,H. Uejima,S. Asai,K. (1982), *Linear regression anaysis with fuzzy model*, IEEE Trans. System Man Cybernet.,vol.12(6), 903-907 P.

Watada,J. (1992), *Fuzzy time series analysis and forecasting of sales volume*, Kacprzyk,J. Fedrizzi,M. (Eds.), Fuzzy Regression Analysis, Omnitech Press, Heidelberg, 211-227 P.

Song,Q. Chissom,B.S. (1993), *Fuzzy time series and its models*, Fuzzy Sets and Systems, vol.54(3), 269-277 P.

Song,Q. Chissom,B.S. (1993), *Forecasting enrollments with fuzzy time series-Part I*, Fuzzy Sets and Systems, vol.54(1), 1-9 P.

Song,Q. Chissom,B.S. (1994), *Forecasting enrollments with fuzzy time series-Part II*, Fuzzy Sets and Systems, vol.62(1), 1-8 P.

Tseng,Fang-Mei. Tzeng,Gew-Hshiung. Yu,Hsiao-Cheng. Yuan,J.C. (2001), *Fuzzy ARIMA model for forecasting the foreign exchange market*, Fuzzy Sets and Systems, vol.118, 9-19 P.

Sullivan,J. Woodall,W,H. (1994), *A comparison of fuzzy forecasting and Markov modeling*, Fuzzy Sets and Systems, vol.64, 279-293 P.

Zadeh,L.A. (1965), *Fuzzy Sets*, Inform and Control, vol.8, 338-353 P.

آزمایشهای آمیزه‌ای و کاربرد آن در فرمولاسیون مواد غذایی

مجید صیوتی، فتاح میکائیلی

P۱۳۲۲۶

گروه آمار، دانشگاه علامه طباطبایی

چکیده: بسیاری از محصولات مورد مصرف امروزی از مخلوط کردن دو یا بیشتر از دو مولفه بوجود می آیند، برای مثال تولید آلیاژهای فولاد و چدن، بتونهای سیمانی، رنگهای نقاشی و یا تولید معجون‌های مختلف و ... از این گونه‌اند. همچنین در اکثر موارد خواص محصولات فقط به نسبت مواد تشکیل دهنده محصول بستگی دارد.

طرح آزمایشهای آمیزه‌ای، شاخه‌ای از طرح آزمایشهاست که در آن پاسخ، فقط تابعی از نسبت مولفه‌های تشکیل دهنده آمیزه است. تفسیر آزمایشهای آمیزه‌ای به ایده روش‌شناسی رویه پاسخ وابسته است که یک روش دنباله‌ای است که با استفاده از ابزارهای طرح آزمایش و رگرسیون روابط بین اثرات تیماری را توضیح می‌دهد.

یک گروه کارشناسی متشکل از متخصصین صنایع غذائی و آمار اقدام به فرمولاسیون نوعی نوشیدنی به نام ایزوله پروتئین سویا کرده‌اند. در تهیه این نوشیدنی از یک طرح D-اپتیمم (که به کمک نرم‌افزار SAS تهیه شده است) استفاده شده و هدف بهینه کردن طعم نوشیدنی به ازای ترکیبات نسبت‌های مختلف تشکیل دهنده نوشیدنی بوده است. در این نوشیدنی که برای اولین بار و به صورت آزمایشگاهی تولید شده، در نهایت یک فرمول بهینه برای آن بدست آمده است.

واژه‌های کلیدی: آزمایش آمیزه‌ای، سادک، چندجمله‌ای کانونی، روش شناسی رویه پاسخ، طرح‌های بهینه، فرمولاسیون.

۱ مقدمه

بسیاری از محصولات مورد استفاده امروزی از ترکیب دو یا چند جزء از مواد مختلف بدست می آید، مثلاً فولاد و چدن آلیاژهای هستند که از ترکیب نسبتهای خاصی (متفاوت) از کربن و آهن بوجود می آیند و یا معجونهای مختلف از نسبت‌های مختلفی از عصاره‌های میوه‌های مختلف تشکیل می‌شود.

در مواقعی خواص محصولات یا آمیزه‌ها به نسبتِ (بر حسب حجم یا وزن و یا حجم

مولکولی و غیره) اجزاء تشکیل دهنده آن بستگی دارد، در مثالهای بالا خواص فولاد و چدن (از لحاظ شکنندگی، استحکام و غیره) با هم متفاوت است و یا طعم معجون بستگی به نسبت عصاره‌های تشکیل دهندهٔ آن دارد. بررسی این گونه آزمایشها در آمار به نوع خاصی از طراحی آزمایشها برمی‌گردد که به **آزمایشهای آمیزه‌ای**[۱] معروفند.

۲ مساله اصلی

با مقدمه کوتاهی که در مورد آزمایشهای آمیزه‌ای بیان شد به تعریف کلی و دقیق مساله می‌پردازیم.

تعریف: آزمایش آمیزه‌ای، آزمایشی است که در آن پاسخ فقط تابعی از نسبت اجزاء تشکیل دهنده آمیزه است و به مقدار کل آمیزه بستگی ندارد.

در واقع در اینگونه از آزمایشها عاملها، نسبت‌های اجزاء تشکیل دهنده می‌باشند. این نسبتها نامنفی هستند و مجموع آنها برابر یک است. فرض کنید که در آمیزهٔ q مولفه‌ای، x_i نسبت مولفهٔ iام مخلوط باشد، داریم که:

$$\sum_{i=۱}^{q} x_i = ۱ \quad , x_i \leq ۰ \quad , i = ۱, ۲, \ldots, q \qquad (۱)$$

قید اخیر محدودیت اساسی مربوط به اجزاء متشکله آزمایش آمیزه‌ای است، به این معنی که هر مولفه بوسیله دیگر مولفه‌ها تعیین می‌شود و درجه آزادی یک سیستم آمیزه‌ای با q مولفه برابر با $q - ۱$ است.

۳ فضای پارامتری

با توجه به قیود (۱) فضای پارامتری شامل مجموعه نقاط واقع بر روی یا داخل **سادک**[۱] است.

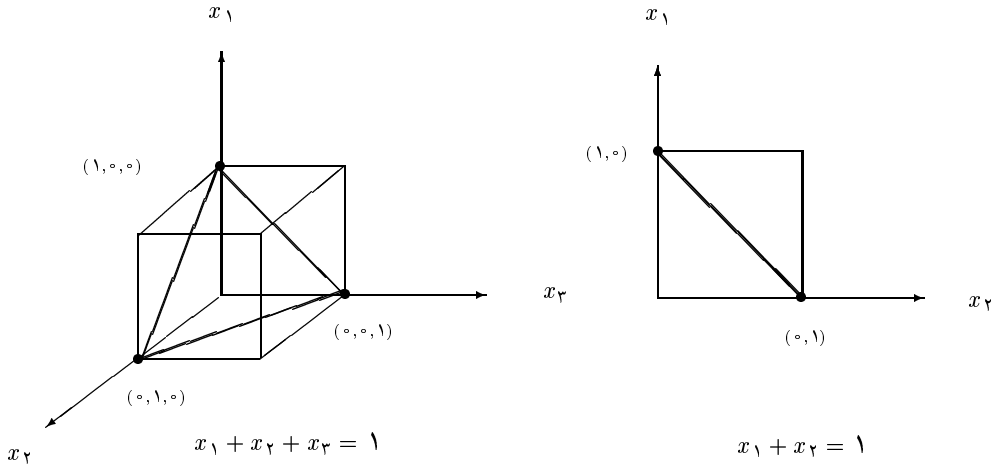
تعریف: یک سادک، یک چند وجهی محدب و منتظم n بعدی است که دقیقاً $n + ۱$ رأس دارد. در حالت صفر بعدی یک سادک یک نقطه، یک بعدی یک خط، دو بعدی مثلث و در حالت سه بعدی یک چهار وجهی و الی آخر می‌باشد. در شکل (۱) فضای پارامتری برای $q = ۲$ و $q = ۳$ نشان داده شده است.

---

[۱] Mixtures Experiments

[۱] Simplex

در دستگاه مختصات سادک برای مثال با $q = ۳$ مولفه، ناحیه آزمایشی یک مثلث است که رئوس این مثلث بر ترکیب خالص یعنی ترکیب **تک مولفه‌ای**[۲] منطبق است، یعنی نقطه‌ای که با $x_j = x_k = ۰$، $x_i = ۱$ برای $i, j, k = ۱, ۲, ۳$ و $i \neq j \neq k$ مشخص می‌شود. یالهای (ضعلها) مثلث ترکیبات **دو مولفه‌ای**[۳] (جزء سوم برابر با صفر) را نشان داده و نقاط داخل مثلث ترکیبات **سه مولفه‌ای**[۴] که تمام مولفه‌ها مخالف با صفر هستند را نشان می‌دهد.



شکل ۱: (الف) فضای عوامل برای $q = ۲$ سادک یک بعدی و (ب) فضای عوامل برای $q = ۳$ سادک دو بعدی

۴ روش‌شناسی رویه پاسخ

در مسایل آمیزه‌ای، داده‌های حاصل از آزمایش در یک مقیاس کمی مثل عملکرد یا خصیصه فیزیکی نهایی تعریف می‌شود. هدفِ برنامهٔ آزمایشی، ارائه الگوی پاسخ آمیزه در قالب یک معادلهٔ ریاضی برای پیش‌بینی پاسخ برای ترکیبی از اجزاء می‌باشد و یا به عبارت دیگر باید با آن معادلهٔ ریاضی بتوانیم برای اثری که هر جزء به صورت تکی یا به صورت ترکیبی روی آمیزه دارد، یک اندازه بدست آوریم. برای تحلیل مسایل آمیزه‌ای

[۲] Simple Component
[۳] Bineary Components
[۴] Ternary Components

از مجموعه‌ای از فنون ریاضی و آماری سودمند به نام **روش‌شناسی رویه پاسخ**[۱] استفاده می‌کنیم. در این روش ما قادر خواهیم بود رابطهٔ بین عوامل و پاسخ را به صورت منحنی یا رویه بیان کنیم. به طور کلی این روش بر چهار گام استوار است:

۱) داده‌ها با استفاده از یک آزمایش از پیش طراحی شده جمع‌آوری می‌گردد.
۲) یک مدل ریاضی (معمولاً چند جمله‌ای) با استفاده از تکنیکهای برازش منحنی (روشهای رگرسیونی) به داده‌ها برازش می‌شود.
۳) به دنبال ناحیه‌ای که مقادیر پاسخ بهینه در آن باشد، جستجو می‌گردد (با استفاده از منحنی‌های تراز پاسخ)
۴) نقاط اضافی برای صحّت مدل برازش شده جمع‌آوری می‌گردد.

۵ چند جمله‌ایهای کانونی

قیود (۱) نه تنها فضای پارامتری را تغییر می‌دهد بلکه چند جمله‌ایهای که برای برآورد پاسخ در این گونه آزمایشها به کار می‌رود نیز متفاوت خواهد بود.

فرض کنید چند جمله درجه اول با q مولفه بصورت زیر باشد.

$$\eta = \beta_0 + \sum_{i=1}^{q} \beta_i x_i$$

حال با ضرب‌کردن $\sum_{i=1}^{q} x_i$ در β_0 داریم:

$$\begin{aligned}\eta &= \beta_0 \sum_{i=1}^{q} x_i + \sum_{i=1}^{q} \beta_i x_i \\ &= \sum_{i=1}^{q} \beta_i^* x_i\end{aligned}$$

که در آن برای هر i داریم:

$$\beta_i^* = \beta_0 + \beta_i \qquad , i = 1, 2, \ldots, q$$

[۱] Response Surface Mathodology

این چند جمله‌ایها به **چند جمله‌ایهای کانونی**[۱] معروفند که چند جمله‌ای درجهٔ دوم و درجهٔ سوم به صورت زیر با اعمالی مشابه بدست می‌آید.

$$\eta = \sum_{i=1}^{q} \beta_i x_i + \sum_{1 \le i < j \le q} \beta_{ij} x_i x_j$$

$$\eta = \sum_{i=1}^{q} \beta_i x_i + \sum_{1 \le i < j \le q} \beta_{ij} x_i x_j + \sum_{1 \le i < j \le q} \beta_{ij} x_i x_j (x_i - x_j) + \sum_{1 \le i < j < k \le q} \beta_{ijk} x_i x_j x_k$$

۶ فرمولاسیون نوشیدنی ایزوله پروتئین سویا

۱.۶ بیان مساله

فرآورده‌های سویا یک منبع مهم و کم هزینه پروتئین، مواد معدنی و ویتامین می‌باشد که نقش مهمی در سلامتی دارند و جهت جلوگیری از بیماریهای قلبی، چشم، سرطان، دیابت، پوکی استخوان و چاقی توصیه می‌شود. **ایزوله پروتئین سویا**[۱] (SPI)، فرآورده‌ای از آرد لوبیای سویای چربی زدایی شده است که شامل بیش از ۹۰ درصد پروتئین می‌باشد.

اهمیت پروتئین سویا در رژیم غذایی، بعلت ظرفیت بالای اسیدهای آمینه ضروری، میزان بالای آن در دانه سویا و ارزان بودن آن است، پروتئین سویا تمام اسیدهای آمینه ضروری بدن را فراهم می‌کند. ایزوله پروتئین سویا به دلیل خصوصیات عملکردی و ارزش تغذیه‌ای به منظور تولید فرآورده‌های جانشین لبنیات، فرآورده‌های گوشتی، سوپ و پودر نوشیدنی‌های فوری با طعمهای متفاوت به کار می‌رود. از طرف دیگر در رژیم غذایی ورزشکاران، جهت تامین پروتئین اضافه توصیه می‌شود، چرا که ورزشکاران به مقدار پروتئین بیشتری نسبت به سایر افراد نیازمندند. همچنین لازم است با دادن یک رژیم غذایی مناسب افزایش وزن ورزشکار را به افزایش تودهٔ عضلانی تبدیل کرد. از آنجایی که نیاز ورزشکاران به مقدار پروتئین بیشتر از سایر افراد است در نتیجه مقدار دریافت آب بدن آنها نیز افزایش می‌یابد تا دفع مواد حاصل از فعل و انفعالات ترکیبات ازتی به راحتی انجام گیرد، که به کار بردن یک نوشیدنی پروتئینی به حل این مشکل کمک می‌کند.

---

[۱] Canonical Polynomials

[۱] Soyabeon Protein Isolate

۲.٦ تعیین پاسخ

طعم فرآورده‌های سویا مشکل عمده و عامل محدود کنندهٔ گسترش کاربرد پروتئین سویا در غذای انسانی است. با بهبود خواص عملکردی و طعم ایزوله پروتئین سویا، می‌توان از آن جهت تولید یک نوشیدنی پروتئینی استفاده کرد. با تهیه فرمولاسیون مناسب از ایزوله پروتئین سویا در غالب یک نوشیدنی و با تاکید بر طعم و پذیرش کلی آن از طرف مصرف کننده می‌توان رغبت بیشتری در میان مصرف کنندگان ایجاد نمود.

۳.٦ اهداف

به طور کلی هدف‌های این تحقیق بصورت زیر بوده است:

۱) برازش رویه پاسخ مناسب.
۲) تهیه بهترین فرمولاسیون برای این محصول.

۴.٦ تعیین مولفه‌ها

جهت تهیه پودر نوشیدنی، SPI، شیرین کننده (شکر)، ایزوله پروتئین سویا، اسانس و مالتود کسترین به صورت خشک مخلوط می‌شود و سپس در آسیاب کاملاً همگن می‌گردند. در نهایت پودر حاصله به نسبت ۱ به ۵ در شیر حل می‌شود. به دلیل محدودیتهای آزمایشگاهی و همچنین قیدهایی که بر روی مولفه‌های تشکیل دهندهٔ مشابه این محصول وجود دارد چهار مولفهٔ تشکیل دهندهٔ ایزوله پروتئین سویا به صورتی که در جدول ۱ مشاهده می‌شود محدود شده‌اند.

جدول ۱: محدودیتها بر روی مولفه‌های واقعی (بر حسب گرم)

مولفه‌ها	کد	حد پایین	حد بالا	حد پایین (درصد)	حد بالا (درصد)
شکر	$C_۱$	۵	۱۱	۲۵	۵۵
ایزوله پروتئین سویا	$C_۲$	٦	۱۱	۳۰	۵۵
اسانس	$C_۳$	۱٫۰	۲	۵٫۰	۱۰
مالتودکسترین	$C_۴$	۱	۲	۵	۱۰

مولفه‌های تشکیل دهنده در مقیاس گرم اندازه‌گیری شده‌اند و مولفه‌ها در هر آزمایش ۲۰ گرم بوده است که این مقدار تقریباً در ۱۰۰ سی‌سی شیر حل شده است.

۵.٦ ارزیابی پاسخ

ارزیابی پاسخ یک روش علمی است که در آن برای اندازه‌گیری تعبیر و تفسیر خصوصیات حسی غذا از حواس انسان استفاده می‌شود، اهمیت این روشها به قابلیت ویژه آنها برمی‌گردد. به عنوان مثال طعم غذا، بوی غذا و اینکه غذا چگونه به نظر می‌رسد یا احساس می‌شود توسط هیچ روش فیزیکی و شیمیایی قابل تشخیص و اندازه‌گیری نیست.

برای همین منظور گروهی متشکل از کارشناسان صنایع غذایی و آمار جهت ارزیابی خصوصیات **ارگانولیپتکی**[۱] شامل طعم، رنگ، مزه و به طور کلی پذیرش کلی تست پانل توسط ٦ داور انجام گرفته و میانگین نمرات آنها به عنوان پاسخ برای آن ترکیب در نظر گرفته شده است. تست پانل در قالب تست **هدنیک**[۲] پنج نقطه‌ای که ۲− بدترین نمره و ۲+ بهترین نمره بوده است. البته ملاحظاتی از قبیل یکنواختی نمونه، دمای نمونه، ظروف نوشیدنی، کدگذاری، ترتیب ارائه نمونه، زمان آزمایش، محیط آزمایش و ... در نظر گرفته شده است.

٦.٦ انتخاب طرح

همانگونه که در بخش ۴.٦ گفته شد مولفه‌های به کار رفته برای تهیه ایزوله پروتئین سویا در کرانهای بالا و پایین محدود هستند بنابراین طرحهای آمیزه‌ای استاندارد برای این آزمایش مستقیماً نمی‌تواند مورد استفاده قرار گیرد. با استفاده از **مولفه‌نماها**[۳] که به صورت تبدیل زیر انجام می‌شود آزمایش آمیزه‌ای به کار گرفته می‌شود.

$$X_i = \frac{a_i - L_i}{۲۰ - \sum L_i}$$

که در آن a_i وزن واقعی مولفه iام آمیزه است.

اسنی[۴] [۱۹۷۵] با استفاده از اصولی موفق شد طرحی را برای مدل درجهٔ دوم ارائه کند که این طرح شامل رئوس فرین، مراکز صفحات مقید، نقطه مرکزی کلی و مراکز بزرگترین یالها است. نقاط نامزد برای انتخاب طرح شامل ۳۱ نقطهٔ رئوس، نقاط میانی یالها و وجه‌ها را تشکیل می‌دهند. به علت محدودیت‌های زمانی و امکانات از ۳۱ نقطه نامزد، ١٦ نقطه به عنوان طرح D-اپتیمم با استفاده از از روال OPTEX[۵] در SAS انتخاب شده است. این نقاط با فرض درست بودن مدل درجهٔ دوم شفه بوده است.

---

[۱] Organoleptic
[۲] Hedonic
[۳] Pseudocomponents
[۴] Snee
[۵] Optimal Experimental Designs

طرح بهینه شامل ۹ نقطهٔ راس فرین، ۴ نقطه میانی یالها، ۳ نقطه میانی وجه‌ها است. ۴ نقطه در رئوس فرین هر کدام هر کدام یک بار تکرار شده تا آزمون نکویی برازش میسّر گردد.

۷.٦ تحلیل داده‌ها

تحلیل داده‌ها یکی از مهمترین مراحل طرح آزمایشی است، برای تحلیل داده‌ها و برازش رویه پاسخ به چهار مولفه‌نما از روال RSREG[۱] و GLM[۲] در نرم افزار SAS استفاده شده است.

جدول ۲: آزمون نکویی برازش و تحلیل واریانس

Regression	df	Sum of Squre	F-Value	P-Value
Model	۹	۱۳٫۸٦	۳٫۳۸	۰٫۰۳۵
Error	۱۰	۴٫۵۵		
Lack of fit	٦	۳٫۸۲	۳٫۴۹	۰٫۱۲
Pure error	۴	۰٫۱۸		
Total	۱۹	۱۸٫۴۱		

در پیوست برنامه مربوط به آن با جزئیات بیشتری آورده شده است. آماره‌های مهم و مرتبط از خروجی مورد نظر استخراج شده و در ادامه بحث خواهد شد. همانگونه که در جدول ۲ دیده می‌شود آزمون LOF مناسب بودن مدل درجهٔ دوم را نشان می‌دهد.

مقدار R^2 برابر با ۰٫۷۵۰۴ و همچنین مقدار Root MSE برابر با ۰٫٦۸ است که نشان دهنده خوبی مدل درجه دوم است. همچنین با استفاده از روال GLM برای برازش مدل و برآورد ضرایب استفاده شده است.

جدول ۳: برآورد پارامترهای رویه پاسخ

Variable	Coefficient	S.E	t-Value	P-Value
X_1	−۵٫۲۵	۱٫۸۳	−۲٫۸٦	۰٫۰۱۲
X_2	−۱٫۳٦	۱٫۷۰	−۰٫۸۰	۰٫۴۳
X_3	−۴۰٦٫۰۴	۱۰۴٫۲٦	−۳٫۸۹	۰٫۰۰۱
X_4	۳۲٫۵۰	۹٫۰۷	۳٫۵۸	۰٫۰۰۳
X_1X_3	۵۲۰٫۸۰	۱۲۷٫۸٦	۴٫۰۷	۰٫۰۰۱
X_2X_3	۴٦۸٫۹۳	۱۲۸٫٦۴	۳٫٦۴	۰٫۰۰۲

[۱] Response Surface Regression
[۲] General Linear Models

همچنین آزمونهای نرمالیتی و همسانی واریانس باقیمانده‌ها نیز انجام گرفته که نتایج نرمال بودن باقیمانده‌ها و همسانی واریانس برای تمام سطوح X را نشان می‌دهد.

نسبتهای بهینه (نسبتهای از مولفه‌ها که به ازای آنها پاسخ بهینه، طعم مطلوب، بدست آمده است) به صورت زیر(بر حسب درصد) است:

$$P_۱ = ۴۱ \qquad P_۲ = ۴۷$$
$$P_۳ = ۳ \qquad P_۴ = ۹$$

مراجع

فرهوش، ر. (۱۳۷۴). *بررسی امکان تولید ایزوله پروتئین سویا*، پایان‌نامه کارشناسی ارشد علوم و صنایع غذایی، دانشکده کشاورزی، دانشگاه تربیت مدرس.

جهانیان، ل. (۱۳۸۱). *تهیه نوشیدنی ایزوله پروتئین سویا*، پایان‌نامه کارشناسی ارشد علوم و صنایع غذایی، دانشکده کشاورزی، دانشگاه تربیت مدرس.

Cornell, J. A. (1990), *Expriments with Mixtures: Designs, Models and the Analysis of Mixture Data*. 2nd edition. John Wiley & Sons, New York, 1990.

Myers, R. H. and Montgomery, D. C. (1995), *Response Surface Methodology: proccess and product optimization using designed experiments*. John Wiley & Sons, New York, 1995.

Ruguo, Hu. (1999), *Food Product Design: A Computer-Aided Statistical Approach*, Technomic Publishing Company, Inc.

Scheffé, H. (1958), Experiments with Mixtures. *J. R. Stat. Soc, B, 20, No.2, 344-360.*

Snee, R. D. (1979), Experimenting with Mixtures. *CHEMTECH, 9, 702-710.*

Snee, R.D. (1975) Experimental Design for Quadratic Models in Constrianed Mixture Spaces. *Techonometrics, 17, 149-159*

پیوست

روال RSREG روش حداقل مربعات را برای برآورد پارامترهای درجهٔ دوم به کار می‌گیرد. مدلهای رویهٔ پاسخ مخصوصاً هنگامی که هدف برآورد پاسخ بهینه است به کار گرفته می‌شوند. به طور کلی روال RSREG می‌تواند برای اهداف زیر مورد استفاده شود:

۱) آزمون برای عدم برازش
۲) آزمون اثرات عاملهای تکی
۳) تحلیل ساختار کانونیک رویه پاسخ برآورد شده
۴) تحلیل ستیغی برای رسیدن به پاسخ بهینه
۵) پیش‌بینی مقادیر جدید پاسخ

فرم کلی استفاده از این روال در SAS به صورت زیر است:

```
PROG RSREG <options>;
  MODEL response=independent <options>;
  RIDGE <options>;
  WEIGHT <options>;
  ID variables;
  BY variables;
```

استفاده از PROC RSREG و عبارت MODEL اجباری است. در ادامه خروجی کامپیوتری با استفاده از ADX و سایر روالهای موجود در SAS با جزئیات بیشتری آورده شده است.

تخمین بهینه وزنها در نمودارهای EWMA و DEWMA با استفاده از MCDM

رضا برادران کاظم‌زاده [۱]، عزیزا... معماریانی [۱]، مهدی کرباسیان [۲]

P۱۱۰۲۴

[۱] بخش مهندسی صنایع، دانشگاه تربیت مدرس
[۲] دانشگاه مالک اشتر

چکیده: ضرایب پیش‌بینی کننده و تصحیح کننده در نمودارهای کنترل میانگین متحرک موزون نمایی EWMA و نمودارهای کنترل میانگین متحرک موزون نمایی جفتی DEWMA نقش اساسی در تنظیم عملیات کنترل کیفیت آماری در فرایند ساخت نیمه‌هادی، الکتروسرام و قطعات اپتیکی دارند. تخمین بهینه این ضرایب که معمولاً بر اساس حداقل کردن مربع خطا .M.S.E انجام می‌شودباعث تشخیص سریع هر گونه تغییری در میانگین و براکندگی مشخصه کیفی فرایند می شود.

در این مقاله از روش تصمیم‌گیری با معیارهای چندگانه M.C.D.M و تکنیک برنامه‌ریزی آرمانی G.P برای کاهش اریب و انحراف معیار آماره آن نمودارها به صورت جداگانه و بر اساس نظر و آرمانهای تصمیم‌گیرنده .D.M و اولویتهایی که او قائل است استفاده می‌شود. این مساله با توجه به اهمیت نظرات مدیریت و کارشناسان در طراحی، شروع و ادامه نمودارهای کنترلی، حائز اهمیت است.

۱ مقدمه

نمودارهای کنترل برای اولین‌بار توسط دکتر **والتر شیوارت** (۱۹۹۳) معرفی شده‌اند . پیج در سال ۱۹۹۵ برای اولین‌بار پیشنهاد کرد که از حدود کنترلی به نام حدود هشدار، جدا از حدود کنترل شیوارت استفاده شود. (پاگ ۱۹۹۵) بعد از پیج حدود و معیارهای دیگری برای حساس سازی بیشتر نمودارهای کنترل توسط افرادی از قبیل **چمپ** و **دال** در سال ۱۹۹۷ ابداع گردید. پیج در سال ۱۹۵۴ استفاده از نمودارهایی با عنوان جمع تجمعی [۱] CUSUM را مطرح نمود (پاگ ۱۹۵۴). دراین نمودار از اطلاعات موجود در نمونه‌های قبلی نیز استفاده می‌شود. نمودار میانگین متحرک موزون نمایی [۲] EWMA با خواصی شبیه به نمودارهای CUSUM برای اولین‌بار توسط **روبرتس** در سال ۱۹۵۹ توسعه یافت. **لوکاس و سکیوس** در سال ۱۹۹۰ نتیجه‌گیری کرده‌اند که عملکرد نمودارهای CUSUM

[۱] *Cumulative Sum Control Charts*
[۲] *Exponentially Weighted Moving Average*

و EWMA بسیار شبیه به هم هست (لوکاس ۱۹۹۰). در مقالات متعددی نیز به کارایی قویتر نمودارها EWMA در مقایسه با نمودارهای CUSUM اشاره شده است (وان براکل و گان) به هر حال زمانی که پی‌بردن به وجود تغییرات کوچک مورد نظر باشد، نمودارهای EWMA و CUSUM بهتر از نمودارهای شیوارت عمل می‌کنند. البته **لوکاس** در سال ۱۹۷۳ نشان داده است که نمودارهای شیوارت نسبت به تغییرات بزرگ در میانگین، حساستر از نمودارهای EWMA می‌باشند.

نمودارهای EWMA غالباً از مشاهدات انفرادی استفاده می‌کنند. البته آماره این نمودار به علت اینکه میانگین موزونی از مشاهدات انفرادی باشد. همچنین این آماره کاربرد زیادی در مدل‌سازی سریهای زمانی و پیش‌بینی دارد (جنکینز ۱۹۷۶) و (پِری ۱۹۷۳).

بالتر و **استفانی** در فرایندهای متحول و آشوبناک نمودارهایی پیشنهاد کرده‌اند که از یک ضریب تصحیح کننده – پیش‌بینی کننده استفاده می‌کند. این نمودارها با نام DEWMA[1] شناخته می‌شوند.

۲ نمودارهای EWMA و DEWMA

آماره میانگین متحرک موزون نمایی در نمودار DEWMA به صورت زیر تعریف شده است:

$$Z_t = \lambda X_t + (1-\lambda) Z_{t-1} \qquad (1-2)$$

در این رابطه λ دارای مقدار ثابتی بین $0 < \lambda \leq 1$ است. مقدار Z_t برابر μ_0 می‌باشد. در بعضی مواقع نیز مقدار Z_0 از میانگین اطلاعات اولیه $(\bar{X})$ تخمین زده می‌شود. رابطه (۱–۲) به صورت زیر بازنویسی می‌شود:

$$\begin{aligned} Z_t &= \lambda X_t + (1-\lambda)[\lambda X_{t-1} + (1-\lambda) Z_{t-2}] \qquad (2-2) \\ &= \lambda X_t + \lambda(1-\lambda) X_{t-1} + (1-\lambda)^2 Z_{t-2} \\ Z_t &= \lambda \sum_{j=0}^{t-1} (1-\lambda)^j X_{t-1} X_t + (1-\lambda)^t Z_0 \end{aligned}$$

[1] *Double* EWMA

وزنهای $\lambda(۱-\lambda)^j$ به صورت هندسی با افزایش عمر کاهش می‌یابند. همچنین جمع وزنها به سمت یک میل می‌کند:

$$\lambda\sum_{j=۰}^{t-۱}(۱-\lambda)^j=\lambda[\frac{۱-(۱-\lambda)^t}{۱-(۱-\lambda)}]=۱-(۱-\lambda)^t$$

اگر مشاهدات X_i متغیرهای تصادفی مستقل با واریانس σ باشد آنگاه:

$$\begin{aligned}\sigma_{z_t}&=\sigma^۲\lambda^۲[\frac{۱-(۱-\lambda)^{۲t}}{۱-(۱-\lambda)^۲}]\\&=\sigma^۲(\frac{\lambda}{۲-\lambda})[۱-(۱-\lambda)^{۲t}]\end{aligned}\qquad(۳-۲)$$

با افزایش σ_{z_t} به مقدار حدی خود یعنی:

$$\sigma^۲_{Z_t}=\sigma^۲(\frac{\lambda}{۲-\lambda})\qquad(۴-۲)$$

میل خواهد کرد.

از طرف دیگر اگر میانگین هر x_t را μ بنامیم، در آن صورت امیدریاضی Z_t به صورت زیر خواهد بود:

$$E(Z_t)=\mu-(۱-\lambda)^t(\mu_۰-\mu)\qquad(۵-۲)$$

در نمودارهای DEWMA آماره W_t یک تصحیح کننده است، به صورت زیر محاسبه می‌شود:

$$W_t=\lambda_۲(x_t-z_{t-۱})+(۱-\lambda_۲)W_{t-۱}\qquad(۶-۲)$$

که در آن $W_۰=۰\ ,\ W_۰<\lambda_۲$، ۱.

رابطه فوق را می‌توان به صورت زیر بازنویسی کرد:

$$W_t=\lambda_۲\sum_{j=۰}^{t-۱}(۱-\lambda_۲)^jF_{t-j}$$

که در آن $F_{t-j}=x_t-Z_{t-۱}$ می‌باشد و میزان آماره DEWMA برابر مقدار $R_t=w_t+z_t$ خواهد بود. اگر مشاهدات x_t متغیرهای تصادفی مستقل با واریانس $\sigma^۲$ باشند و با توجه به رابطه (۲—۴) و استقلال $Z_{t-۱}$ و X_t، آنگاه R_t دارای واریانس زیر خواهد بود:

$$si^۲_{R_t}=\sigma^۲+(\frac{\lambda_۱}{۲-\lambda_۱})(\frac{\lambda_۱}{۲-\lambda_۱})\qquad(۷-۲)$$

از طرف دیگر اگر میانگین X_t برابر μ باشد و با توجه به رابطه (۲–۵) خواهیم داشت:

$$E(R_t) = \mu - (۱ - \lambda_۱)^t(\mu_۰ - \mu) + [۱ - (۱ - \lambda_۱)^t][\mu - (۱ - \lambda_۱)^{t-۱}(\mu_۰ - \mu)]$$

۳ طراحی نمودارهای کنترل EWMA و DEWMA

مهمترین پارامترهایی که در طراحی این نوع نمودارها مورد استفاده قرار می‌گیرند مقادیر λ، $\lambda_۱$ و $\lambda_۲$ مهمترین مقادیر می‌باشند. این امکان وجود دارد که بتوان این پارامترها را به گونه‌ای انتخاب نمود تا هرگونه انحراف از میانگین و یا افزایش پراکندگی با سرعت تشخیص داده شود. **دل کاستلو** در سال ۱۹۹۹ یک مدل غیر خطی برای بدست‌آوردن مقادیر بهینه این ضرایب بر اساس کمینه کردن میزان میانگین مربع خطا [۱] MSE بدست آورده است.

۴ استفاده از تکنیکهای MCDM برای تخمین بهینه λ، $\lambda_۱$ و $\lambda_۲$

توجه محققین در دهه اخیر معطوف به مدلهای چندمعیاره [۲] (MCDM) برای تصمیم‌گیری پیچیده گردیده است. شاید بتوان گفت برنامه‌ریزی آرمانی GP[۳] یکی از قدیمیترین مدلهای موجود در تصمیم‌گیری چند معیاره است که با کاربردهای وسیع به کارگیری شده است.

چارلز و کوپر اولین مقاله را درباره G.P در سال ۱۹۵۵ منتشر کردند به طوری که آنها می‌نیمم کردن مجموع قدرمطلق انحرافات از مقاصد مشخصی را مورد بررسی قرار داده‌اند. تلاش در G.P. بر آن است که منطق مدلهای ریاضی بهینه تواماً با تمایل تصمیم‌گیرنده [۴] در تامین مقاصد مشخص از اهداف مختلف مورد توجه قرار بگیرند.

در این قسمت میزان اریب و واریانس آماره نمودارهای EWMA و DEWMA به صورتی آرمانی بهینه می‌گردد، آرمان در اریب صفر می‌باشد (بی اریبی) و آرمان در مورد واریانس بستگی به نظر تصمیم‌گیرنده، دارد.

با این اوصاف مدل G.P. در نمودار EWMA به صورت زیر است:

$$lexMin\{n_۱ + p_۱ \ , \ n_۲\} \qquad (۱ - ۴)$$

[۱] *Mean Square Error*
[۲] Multi Squre Error
[۳] Goal-Programing
[۴] Decision Maker

$$(۱-\lambda)^t(\mu_۰-\mu)+n_۱+p_۱=۰ \qquad (۱)$$

$$\sigma^۲(\frac{\lambda}{۲-\lambda})+n_۲=S \qquad (۲)$$

$$۰<\lambda\leq ۱ \qquad (۳)$$

$$n_۱,n_۲,p\geq ۰ \qquad (۴)$$

باتوجه به اهمیت اریب، از مدل لیکسوگراف با اولویت اریب نسبت به واریانس استفاده شده است.

S میزان واریانس آماره است که توسط D.M تعیین می‌شود.

مدل G.P. در حالت DEWMA به صورت زیر می‌باشد:

$$lextMin\{n_۱+p_۱\ ,\ n_۲\} \qquad (۲-۴)$$

$$(۱-\lambda_۱)^t(\mu_۰-\mu)-[۱-(۱-\lambda_۱)^t][\mu-(۱-\lambda_۱)^{t-۱}(\mu_۰-\mu)]+n_۱+p_۱=۰ \quad ۱$$

$$(\frac{\lambda_۲}{۲-\lambda_۲})+n_۲=S$$

$$۰<\lambda\leq ۱$$

$$n_۱,n_۲,p_۱\geq ۰$$

۵ نتیجه‌گیری و پیشنهادات

تعیین وزنهای بهینه λ، $\lambda_۱$ و $\lambda_۲$ در نمودارهای EWMA و DEWMA، در کاهش اریب و واریانس و در نتیجه تشخیص سریع تغییر در میانگین و افزایش پراکندگی و کاهش خطای نوع دوم (β) کاملاً موثر است. همچنین مقادیر بهینه λ، $\lambda_۱$ و $\lambda_۲$ باعث این پایداری این نمودارها می‌شود.

می‌توان گفت وزنهای بهینه λ، $\lambda_۱$ و $\lambda_۲$ را در حالتی که نمودارهای فوق برای آماره‌هایی مانند تعداد عیب و یا تعداد قطعات معیوب به کار می‌روند محاسبه کرد. همچنین استفاده از برنامه‌ریزی آرمانی موزون و همچنین محاسبه وزنهای بهینه در این نوع برنامی‌ریزی برای فعالیتهای آینده پیشنهاد می‌شود.

مراجع

Shewart,W.A(1993).*"The Economic control of Qualitiy Manufactured Products"*. Van Nostrand, New York

Pages, E.S. (1995). *"Control Charts With Warning Lines"* Biometrica, 41, 243-259.

Champ, C.W., and Woodall, W.H., (1987). *"Exact Result For Shewart Control Charts With Supplementary Run Rules"* Technometrics, 29, 393-399.

Pages, E.S. (1954).*"Continuous Inspection Schemes"* Biometrica, 41, 100-114.

Roberts, S.W. *"Control Chart Tests Base On Geometric Moving Averages"* Technometrics,(1959) 1, pp 239.250

Lucas,J,M. and Saccucci, M.S.(1990). *"Exponentially Moving Averages Control Chart: Properties and Enhancements"* Technometrics, 31, 1-12.

Van Brackle, L.N., III; Rynols, M.R., Jr.*EWMA amd CUSUM Control Charts In THe Procence Of Corrolation, Comminications In Statistics -Simultain"* Vol:26, lss: 3, pp.979, 1008

Gan, F.F.*" Desigen Of One-and Two-saided exponential EWMA Charts"* Jornal of Quality Technology.

Lucas, J.M.(1973)*"A modified V-Mask Scheme."* Technometrics, 15,833-847

Box, G.E.P., and G.M.jenkins(1976) *"Time Series Analysis, Forecasting ,and Control,"*Holden-Day, Sanfransisco.

Perry, R.I. (1973). *"Skip-Lot sampling Planes."* Journal of Quality Technology, Vol. 5.

Bulter, S. W. and Stfain, J. a. (1994). *Supervisory Run-To-Run Control of a Polysilicon Gate etch using in Situ Ellipsometry."* IEEE Transaction on Semiconductor Manufacturing. 7,193-201.

Castillo, E. D.(1999) *"Long Run and Transient Analysis of a Double Ewma Feedback Controller."*, IIE Transaction, 31, 1137-1169.

مقایسه مدل رگرسیون پواسن و مدل رگرسیون دوجمله‌ای منفی در تعیین عوامل مؤثر بر حاملگی ناخواسته

انوشیروان کاظم‌نژاد، سیدمهدی سادات‌هاشمی، شبنم کریمی

P۱۳۲۰۸

گروه آمار زیستی، دانشگاه تربیت مدرس

چکیده: موضوع تحقیق حاضر« مقایسه مدل رگرسیون پواسن و مدل رگرسیون دوجمله‌ای منفی درتعیین عوامل مؤثر بر حاملگی ناخواسته» می‌باشد و متغیر وابسته تعداد حاملگی‌های ناخواسته درنظر گرفته شده است.

در آنالیز رگرسیون در حالتی که متغیر وابسته گسسته و نامنفی باشد، مدل رگرسیون پواسن مورد استفاده قرار می‌گیرد. مدل رگرسیون پواسن به عنوان نمونه‌ای از مدل‌های تعمیم یافته خطی است که مدل‌های رگرسیونی را به خانواده نمایی توزیع‌هایی بسط می‌دهد که هر دو توزیع نرمال و پواسن را شامل می‌شوند. شرط اصلی استفاده از مدل رگرسیون پواسن معادل بودن میانگین و واریانس متغیر وابسته می‌باشد. وقتی که میانگین و واریانس داده‌ها به طور تقریبی برابر نباشند، مدل پواسن برآوردهای ناصحیحی از واریانس جملات و استنباط هایی گمراه کننده درباره رگرسیون ایجاد می‌کند. برای حل این مشکل می‌توان از مدل رگرسیونی دوجمله‌ای منفی استفاده کرد.

به منظور تعیین مهمترین عوامل خطرساز در حاملگی ناخواسته و تعیین برآورد پارامترها از مدل‌های رگرسیونی پواسن و دوجمله‌ای منفی استفاده شده و مقایسه‌ای بین آنها صورت گرفته است. از بین کلیهٔ متغیرهای مورد بررسی، متغیرهای سطح تحصیلات مادر، تعداد فرزندان زنده دختر، تعداد فرزندان زنده پسر و استفاده از روش پیشگیری مؤثر بوده و متغیرهای سن، سطح تحصیلات پدر و آگاهی قبلی از روشهای تنظیم خانواده از نظر آماری از مدل حذف شدند و اثرات متقابل نیز وارد مدل نشدند.

واژه‌های کلیدی: مدل‌های تعمیم یافته خطی، رگرسیون پواسن، رگرسیون دوجمله‌ای منفی، حاملگی ناخواسته.

۱ مقدمه

۱.۱ مدل‌های تعمیم یافته خطی

بخش وسیعی از آمار زیست سنجی و اقتصاد، به تحلیل داده‌های پاسخ گسسته مربوط می‌شود. در مواردی که متغیر پاسخ گسسته باشد(متغیرهای مستقل از هر نوع که باشند) مدل‌های تعمیم یافته خطی را به کار می‌بریم.

مدل‌های تعمیم یافته خطی دارای دو ویژگی کلی هستند. اول اینکه برای متغیر وابسته y_i با مقدار مورد انتظار μ_i، توزیع احتمال y_i به شرط μ_i، عضوی از خانواده نمایی است. دوم اینکه یک تابع پیوند وجود دارد(تابع تبدیل $g(.)$) که مقدار مورد انتظار y_i را خطی می‌کند [1].

مدل رگرسیون پواسن و مدل رگرسیون دوجمله‌ای منفی نمونه‌هایی از مدل‌های تعمیم یافته خطی‌اند که برای آنالیز داده‌های شمارشی مورد استفاده قرار می‌گیرند [2].

۲.۱ مدل رگرسیون پواسن

ساده‌ترین مدل برای داده‌های شمارشی، مدل رگرسیون پواسن می‌باشد. مدل رگرسیون پواسن به عنوان نمونه‌ای از مدل‌های تعمیم یافته خطی است که مدل‌های رگرسیونی را به خانواده نمایی توزیع‌هایی بسط می‌دهد که هر دو توزیع نرمال و پواسن را شامل می‌شوند.

روش رگرسیون پواسن مابین آنالیز مبتنی بر داده‌های پیوسته و آنالیز مبتنی بر داده‌های دو حالتی قرار گرفته است. در رگرسیون پواسن متغیر وابسته شمارشی بوده و دارای توزیع پواسن می‌باشد که این موضوع اساس استنباط را تشکیل می‌دهد. معادله رگرسیون پواسن، یک متغیر شمارشی یا یک نرخ را به یک سری متغیرهای مستقل مربوط می‌سازد و ساختاری را برای تحلیل آماری ارائه می‌دهد. در مدل رگرسیون پواسن، لگاریتم نرخ بر اساس مجموع وزنی متغیرهای مستقل (متغیرهای خطر) مدل‌بندی می‌شود. در واقع این مدل، میانگین متغیر وابسته را بر حسب متغیرهای مستقل، مدل‌بندی می‌کند[3].

محدوده اصلی مدل رگرسیون پواسن بر این شرط استوار است که واریانس متغیر وابسته با میانگین آن برابر باشد، اما در بسیاری از پدیده‌ها داده‌ها بسیار پراکنده‌اند. به عبارت دیگر ممکن است واریانس به طور معنی داری بزرگتر از میانگین باشد. وقتی که میانگین و واریانس داده‌ها به طور تقریبی مساوی نباشند، واریانس‌های ضرایب برآورد شده در مدل پواسن اریب خواهند بود. در این حالت دیگر برازش رگرسیون پواسن بر روی داده‌ها مناسب نمی‌باشد، ولی می‌توان این محدودیت را با استفاده از توزیع دوجمله‌ای منفی از میان برداشت[4].

در تحقیقی که در سال ۱۳۷۸ توسط آقای مهران آل ابریشم کارزاده با عنوان« کاربرد مدل رگرسیون پواسن در تعیین رابطهٔ کیست تخمدان ناشی از کاشت نورپلانت با ریسک فاکتورهای مربوطه» انجام شد، ایشان به تفصیل در ارتباط با رگرسیون پواسن و کاربرد آن به بحث و بررسی پرداخته‌اند[۵].

۳.۱ رگرسیون دوجمله‌ای منفی

توزیع دوجمله‌ای منفی برای داده‌های گسسته و نامنفی مناسب بوده و به برابری میانگین و واریانس حساس نیست. پس به لحاظ مشابه بودن فرم فضای نمونه‌ای، می‌توان از توزیع دوجمله‌ای منفی به جای توزیع پواسن استفاده کرد و برآوردهای دقیق‌تری را به دست آورد.

تقریباً در همهٔ تحقیقات علوم اجتماعی انتظار داریم که تغییرات تبیین نشده‌ای بین موارد وجود داشته باشد، که این تغییرات تبیین نشده، اختلافهای مرتبط با پیش بینی کننده‌های منظور نشده را منعکس می‌کند.

رگرسیون دوجمله‌ای منفی را می‌توان شکل دیگری از رگرسیون پواسن دانست که شامل مؤلفه‌ای تصادفی است که عدم ثبات در نرخ‌های حقیقی رویدادهای واقع شده برای اشخاص را منعکس می‌کند. در واقع مسئله اختلافهای بین موردی تبیین نشده، در مدل رگرسیون دوجمله‌ای منفی حل شده است [2].

اولین کاربردهای توزیع دوجمله‌ای منفی توسط استیودنت (۱۹۰۷) در مقاله‌ای که در آن از توزیع پواسن نیز بحث می‌کند آمده است. استیودنت از این توزیع به عنوان صورت دیگری از توزیع پواسن درشمارش‌های توصیف شده روی طرح‌هایی از یک هماسیتومتر استفاده کرده است. همچنین یول (۱۹۱۰) با استفاده از این توزیع، تعداد مرگ و میرها را در بروز یک بیماری معین در نظر گرفته است. گرین وود و یول (۱۹۲۰) این توزیع را به عنوان یک نتیجه از فرضهای ساده خاصی در مدل پدیده‌های تصادفی به کار گرفتند[٦].

۴.۱ حاملگی ناخواسته

چنانچه زوجین یا یکی از آنها در حوالی به وقوع پیوستن بارداری تصمیم به بچه‌دار شدن نداشته باشند و یا یکی از روشهای پیشگیری از بارداری را استفاده کرده باشند، بارداری به وقوع پیوسته به نام بارداری برنامه ریزی نشده تلقی می‌گردد. حال اگر بعد از اثبات وقوع این بارداری، یکی از زوجین یا هر دو آنها تمایلی به ادامه یافتن بارداری نداشته و خواهان خاتمهٔ آن باشند، بارداری ناخواسته تلقی می‌گردد که در برخی موارد این قبیل بارداری‌ها، حاملگی برنامه ریزی نشده — ناخواسته نیز نامیده می‌شود[7].

در زمینهٔ بررسی علل وقوع حاملگی ناخواسته تحقیقات مختلفی در ایران صورت گرفته که در اینجا به ذکر نمونه‌هایی از آن می‌پردازیم.

در تحقیقی جهت بررسی مشخصات زنان با حاملگی ناخواسته در بین مراجعین به مراکز بهداشتی — درمانی جنوب شهر تهران، خانم امامی افشار در طی ۴۵ روز در زمستان سال ۱۳۶۸ از چهار مرکز بهداشتی — درمانی که به صورت تصادفی از بین یازده مرکز در منطقه جنوب شهر تهران انتخاب شدند، ۳۵۰ پرونده به روش نمونه‌گیری تصادفی انتخاب و از طریق پرسشنامه‌ای شامل ۳۵ سؤال از مادرانی که حداقل تجربه یک بار حاملگی و زایمان را داشتند، اطلاعات مورد نیاز را جمع آوری نمود. متغیرهای مورد بررسی عبارت بودند از: سن مادر، تعداد فرزندان زنده، میزان تحصیلات مادر و همسرش. جهت تجزیه و تحلیل این اطلاعات از تست کای اسکوئر استفاده گردید و مشخص شد که در این داده‌ها بین دو متغیر میزان تحصیلات و حاملگی ناخواسته ارتباطی معکوس وجود دارد و همچنین رابطهٔ معنی داری بین تعداد فرزندان زنده و حاملگی ناخواسته وجود داشت یعنی با افزایش تعداد فرزندان، درصد حاملگی ناخواسته در این مادران افزایش می‌یافت [۸].

در تحقیق دیگری که در سال ۱۳۷۵ توسط خانم فریفته منصوریان صورت گرفت ۶۱۲۵ خانم تحت مطالعه قرار گرفتند. هدف از این مطالعه بررسی عوامل مؤثر بر حاملگی ناخواسته و اقدام به سقط بوده است. متغیرهای مورد بررسی عبارت بودند از: شاخص‌های جمعیتی پدر و مادر، تعداد فرزندان زنده، فاصلهٔ حاملگی از زایمان یا سقط قبلی، مراجعه برای معاینات و وزن نوزاد. از نظر آماری متغیرهای سن پدر و مادر، تحصیلات پدر و مادر، شغل پدر و مادر، قومیّت پدر و مادر، تعداد فرزندان زنده و فاصلهٔ حاملگی از زایمان یا سقط قبلی به عنوان عوامل مؤثر بر حاملگی ناخواسته شناخته شدند [۹].

۲ مواد و روشها

۱.۲ مواد

داده‌هایی که در این مقاله مورد تجزیه و تحلیل قرار می گیرد، از پرسشنامه‌های مربوط به طرح بررسی تاثیر مشاوره تنظیم خانواده پس از زایمان در پیشگیری از حاملگی‌های ناخواسته در دو سال اول پس از زایمان، که توسط مرکز ملی تحقیقات بهداشت و باروری صورت گرفته، جمع آوری شده است.

در این تحقیق جامعهٔ آماری، کلیهٔ خانم‌هایی هستند که در فاصله سالهای ۱۳۷۷ ـ ۱۳۷۵ برای زایمان به یکی از مراکز دهگانه شهر تهران (بیمارستان‌های مهدیه، میرزا

کوچک خان، شریعتی، لقمان، امام حسین، فیروزگر، امام خمینی، رهنمون، آرش و اکبر آبادی) مراجعه کرده‌اند.

با توجه به ماهیت مطالعه، از بین خانم‌هایی که برای زایمان به یکی از مراکز دهگانه فوق در طی سالهای ۱۳۷۷ ـ ۱۳۷۵ مراجعه کرده بودند، نمونه تصادفی ساده به حجم ۴۱۷۷ نفر استخراج گردید.

تعداد حاملگی‌های ناخواسته متغیر وابسته (متغیر پاسخ) و سایر عواملی را که در این قسمت معرفی می‌کنیم متغیرهای مستقل مدل (پیش‌بینی کننده‌ها) می‌باشند.

۱) AGE: این متغیر سن افراد مورد مطالعه را مشخص می‌کند.

۲) GRADE: این متغیر سطح تحصیلات مادر را مشخص می‌کند و در پنج رده(۱— بی‌سواد، ۲— ابتدایی، ۳— متوسطه، ۴— دیپلم و ۵— بالاتر از دیپلم) طبقه‌بندی شده است.

۳) FGRADE: این متغیر سطح تحصیلات پدر را مشخص می‌کند و در پنج رده(۱— بی‌سواد، ۲— ابتدایی، ۳— متوسطه، ۴— دیپلم و ۵— بالاتر از دیپلم) طبقه‌بندی شده است.

۴) DAUT: این متغیر نشان دهندهٔ تعداد فرزندان زندهٔ دختر می‌باشد.

۵) SON: این متغیر نشان دهندهٔ تعداد فرزندان زندهٔ پسر می‌باشد.

٦) CONTRA: این متغیر نشان دهندهٔ استفاده از روش پیشگیری از حاملگی قبل از وقوع حاملگی است و در دو رده(۱— پیشگیری داشته و ۲— پیشگیری نداشته) طبقه بندی شده است.

۷) KNOW: این متغیر نشان دهندهٔ آگاهی قبلی مادر در ارتباط با روشهای تنظیم خانواده است و در دو رده(۱— آگاهی قبلی داشته و ۲— آگاهی قبلی نداشته) طبقه‌بندی شده است.

۲.۲ روشها

در مدل رگرسیون پواسن، لگاریتم نرخ براساس مجموع وزنی متغیرهای مستقل مدل‌بندی می‌شود. تحت فرض اینکه y_i (متغیر وابسته) دارای توزیع پواسن با میانگین μ_i باشد، داریم:

$$Pr[Y_i = y_i] = p(y_i) = (\mu_i^{y_i} e^{-\mu_i})/y_i \ , i = ۱, ۲, ..., n$$

به طوری که مقادیر برآورد شده از رابطهٔ زیر محاسبه می‌شود: $Ln(\mu_i) = \beta X_i$

و شکل نمایی آن به صورت زیر است: $\mu_i = E[Y_i] = e^{\beta X_i}$

که در آن μ_i تعداد وقایع مورد انتظار، X_i بردار متغیرهای پیش بینی کننده و β بردار ضرایب رگرسیون می‌باشد.

مدل $\mu_i = exp(\beta X_i)$ مدل تابع نرخ نمایی است که همواره نامنفی می‌باشد. این نوع تابع نرخ در پردازش انواع مختلف داده‌های شمارشی بسیار انعطاف پذیر است[4]. رگرسیون پواسن معرفی شده به وسیلهٔ رابطهٔ فوق توسط روش‌های حداکثر درستنمایی استاندارد قابل محاسبه است[۵].

توزیع پواسن یک توزیع داوطلب اولیه برای متغیرهای تصادفی گسسته می‌باشد، ولی اگر میانگین و واریانس داده‌ها به طور تقریبی برابر نباشد، رگرسیون پواسن مدل نامناسبی خواهد بود. یک روش ساده برای حل این مشکل استفاده از توزیع دوجمله‌ای منفی می‌باشد.

مدل دوجمله‌ای منفی از مدل پواسن با اضافه کردن جمله خطای ϵ که به طور مستقل توزیع شده به دست می‌آید: $Ln(\mu_i) = \beta X_i + \epsilon$

جمله خطای ϵ دارای توزیع گاما با میانگین یک و واریانس α می‌باشد[2]. این مدل ارتباط بین میانگین و واریانس را به صورت زیر نتیجه می‌دهد[4]:

$$Var(y_i) = E(y_i)[۱ + \alpha E(y_i)]$$

α در رابطهٔ فوق همان واریانس جمله خطاست و یک پارامتر سنجش پراکندگی است که به وسیلهٔ روشهای حداکثر درستنمایی استاندارد قابل برآورد است[10].

اگر $\alpha = ۰$ باشد مدل دوجمله‌ای منفی به مدل پواسن تبدیل می‌شود$(= E(y_i) Var(y_i))$. ولی اگر α اختلاف معنی داری از صفر داشته باشد، مدل دوجمله‌ای منفی انتخابی شایسته و مدل پواسن نامناسب می‌باشد.

در مدل دوجمله‌ای منفی با استفاده از ضریب تعدیلی α محدودیت واریانس مدل پواسن برطرف می‌شود، به دلیل اینکه واریانس ϵ یعنی α پارامتری اضافه شده به مدل واریانس می‌باشد، رگرسیون دوجمله‌ای منفی در مدل بندی ارتباط بین مقدار مورد انتظار و واریانس y_i دارای انعطاف پذیری بیشتری نسبت به مدل به شدت محدود پواسن می‌باشد[4].

در قسمت نتایج به کمک مدل رگرسیون پواسن و مدل رگرسیون دوجمله‌ای منفی اثر هر یک از متغیرها را بر متغیر وابسته (تعداد حاملگی‌های ناخواسته) مورد بحث و بررسی قرار داده‌ایم.

۳ نتایج

با توجه به ماهیت متغیر وابسته (تعداد حاملگی‌های ناخواسته) به منظور مدل بندی عوامل مؤثر بر آن از مدل‌های رگرسیونی پواسن و دوجمله‌ای منفی استفاده می‌کنیم. روش مورد استفاده رگرسیون گام به گام پیشرو با $P_E = ۰/۱۵$ و $P_R = ۰/۲$ می‌باشد(P_E مقدار P-Value برای گنجاندن متغیر در مدل و P_R مقدار P-Value برای حذف متغیر از مدل می‌باشد).

ضریب پراکندگی α اختلاف معنی داری از صفر ندارد($\alpha = ۱/۰۵e-۰۷$). بنابراین در این داده‌ها مدل‌های رگرسیونی پواسن و دوجمله‌ای منفی برازشی مشابه دارند.

متغیرهای سطح تحصیلات مادر، تعداد فرزندان زندهٔ دختر، تعداد فرزندان زندهٔ پسر و استفاده از روش پیشگیری وارد مدل شدند و هیچ یک از اثرات متقابل وارد مدل نشدند.

۱.۳ برآورد پارامترها

ضرایب برآورد شده در مدل رگرسیون پواسن در جدول (۱) و ضرایب برآورد شده در مدل رگرسیون دوجمله‌ای منفی در جدول (۲) نشان داده شده است.

متغیر	ضرایب برآورد شده	خطای استاندارد	آماره والد	P-Value	فاصلهٔ اطمینان ۹۵%
GRADE(2)	−۰/۱۹۶۵	۰/۱۲۶۸	−۱/۵۴۹	۰/۱۲۱	(−۰/۴۴۵۲, −۰/۰۵۲۱)
GRADE(3)	−۰/۱۸۷۶	۰/۱۲۵۵	−۱/۴۹۵	۰/۱۳۵	(−۰/۴۳۳۶, −۰/۰۵۸۴)
GRADE(4)	−۰/۲۶۹۵	۰/۱۳۲۷	−۲/۰۳۱	۰/۰۴۲	(−۰/۵۲۹۵, −۰/۰۰۹۵)
GRADE(5)	−۰/۴۳۵۶	۰/۲۰۶۷	−۲/۱۰۷	۰/۰۳۵	(−۰/۸۴۰۷, −۰/۰۳۰۴)
DAUT	۰/۴۵۲۱	۰/۰۳۴۱	۱۳/۲۵۴	۰/۰۰۰	(۰/۳۸۵۳, ۰/۵۱۹۰)
SON	۰/۳۰۱۹	۰/۰۲۰۱	۱۵/۰۱۶	۰/۰۰۰	(۰/۲۶۲۵, ۰/۳۴۱۳)
CONTRA(2)	−۰/۶۵۴۲	۰/۰۷۱۷	−۹/۱۲۹	۰/۰۰۰	(−۰/۷۹۴۷, −۰/۵۱۳۸)
ثابت	−۱/۴۵۳۶	۰/۱۳۲۵	−۱۰/۹۷۳	۰/۰۰۰	(−۱/۷۱۳۳, −۱/۱۹۴۰)

جدول ۱ : برآوردهای حداکثر درستنمایی برای مدل پذیرفته شده رگرسیون پواسن

با توجه به ستون مقدار آمارهٔ« والد» در جدول (۱) به نتایج زیر می‌رسیم:

به نظر می‌رسد که مهمترین و مؤثرترین متغیر تعداد فرزندان زندهٔ پسر بوده است و پس از آن به ترتیب متغیرهای تعداد فرزندان زندهٔ دختر، و استفاده از روش پیشگیری تأثیر نسبتاً زیادی برمتغیر وابسته، یعنی تعداد حاملگی های ناخواسته داشته‌اند.

همچنین باتوجه به ستون ضدایب برآورد شده به این نتیجه می‌رسیم که بین متغیرهای تعداد فرزندان زندهٔ پسر و تعداد فرزندان زندهٔ دختر و متغیر وابسته یک ارتباط مستقیم برقرار بوده است(برآورد پارامتر برای این دو متغیر مستقل، مثبت است) یعنی با افزایش

متغیر	ضرایب برآورد شده	خطای استاندارد	آماره والد	P-value	فاصلهٔ اطمینان ۹۵%
GRADE(2)	−۰/۱۹۶۵	۰/۱۲۷۷	−۱/۵۳۹	۰/۱۲۴	(−۰/۴۴۶۸, −۰/۰۵۳۷)
GRADE(3)	−۰/۱۸۷۶	۰/۱۲۶۵	−۱/۴۸۳	۰/۱۳۸	(−۰/۴۳۵۶, −۰/۰۶۰۳)
GRADE(4)	−۰/۲۶۹۵	۰/۱۳۳۲	−۲/۰۲۳	۰/۰۴۳	(−۰/۵۳۰۵, −۰/۰۰۸۴)
GRADE(5)	−۰/۴۳۵۶	۰/۲۱۲۹	−۲/۰۴۶	۰/۰۴۱	(−۰/۸۵۲۸, −۰/۰۱۸۴)
DAUT	۰/۴۵۲۱	۰/۰۳۴۱	۱۳/۲۴۵	۰/۰۰۰	(۰/۳۸۵۲, ۰/۵۱۹۰)
SON	۰/۳۰۱۹	۰/۰۲۰۱	۱۵/۰۳۸	۰/۰۰۰	(۰/۲۶۲۵, ۰/۳۴۱۲)
CONTRA(2)	−۰/۶۵۴۲	۰/۰۷۰۶	−۹/۲۶۱	۰/۰۰۰	(−۰/۷۹۲۷, −۰/۵۱۵۸)
ثابت	−۱/۴۵۲۶	۰/۱۳۳۷	−۱۰/۸۷۶	۰/۰۰۰	(−۱/۷۱۵۶, −۱/۱۹۱۷)

جدول ۲: برآوردهای حداکثر درستنمایی برای مدل پذیرفته شده رگرسیون دوجمله‌ای منفی

تعداد فرزندان زندهٔ پسر(یا تعداد فرزندان زندهٔ دختر) خطر وقوع حاملگی ناخواسته بیشتر می‌شود.

برآورد پارامتر منفی برای متغیر سطح تحصیلات مادر و استفاده از روش پیشگیری، نشان دهندهٔ این مطلب است که بین این متغیرها و متغیر وابسته یک رابطهٔ منفی وجود داشته است، یعنی با افزایش میزان سطح تحصیلات مادر، تعداد حاملگی های ناخواسته کاهش یافته است و چون متغیر استفاده از روش پیشگیری از حاملگی متغیری دوحالتی است(۱— پیشگیری داشته و ۲— پیشگیری نداشته)، وجود رابطهٔ منفی به این معنی است که خطر وقوع حاملگی ناخواسته برای افرادی که پیشگیری داشته‌اند بیشتر از کسانی بوده که پیشگیری نداشته‌اند و این نشان دهندهٔ وجود شکافی در آگاهی از کنترل بارداری، حتی در میان استفاده کنندگان از روشهای پیشگیری از بارداری است.

۲.۳ نسبت خطر

در این بخش می‌خواهیم نسبت خطرهای مربوط به متغیرهای مستقل طبقه‌ای را مورد بررسی قرار دهیم. لازم به ذکر است که در هر دو روش رگرسیونی پواسن و دوجمله‌ای منفی اولین سطح هر متغیر طبقه‌ای به عنوان مرجع درنظر گرفته شده و خطرات نسبی بقیه سطوح نسبت به این سطح سنجیده می‌شود. به شرط ثابت نگهداشتن سایر متغیرها نتایج زیر به دست آمده است:

در ابتدا متغیر مستقل سطح تحصیلات مادر را در نظر می‌گیریم. این متغیر دارای پنج رده می‌باشد. نسبت خطر را به صورت زیر درنظر می‌گیریم:

$$\pi = \frac{\text{نسبت حاملگی ناخواسته در یک گروه}}{\text{نسبت حاملگی ناخواسته‌در گروه مرجع}}$$

پس داریم:

$\pi =$ نسبت‌حاملگی‌ناخواسته‌درگروه‌بی‌سواد /نسبت‌حاملگی‌ناخواسته‌درگروه‌باسوادابتدایی

که برابر است با ۰/۸۲۱۶ و فاصله اطمینان ۹۵% برای این نسبت خطر برابر است با (۰/۶۴۰۷, ۱/۰۵۳۵) که نشان دهنده معنی داری این نسبت خطر در سطح خطای ۵% است و خواهیم داشت: $1/\pi = 1/2171$

بنابراین خطر وقوع حاملگی ناخواسته در مادران بی‌سواد نسبت به‌مادرانی که تحصیلاتی در حد ابتدایی دارند، ۱/۲ برابر بیشتر است.

به همین ترتیب نتایج زیر به دست آمده است:

خطر وقوع حاملگی ناخواسته در مادران بی‌سواد نسبت به‌مادرانی که تحصیلاتی در حد متوسطه دارند، ۱/۲ برابر بیشتر است.

خطر وقوع حاملگی ناخواسته در مادران بی‌سواد نسبت به‌مادرانی که تحصیلاتی در حد دیپلم دارند، ۱/۳ برابر بیشتر است.

خطر وقوع حاملگی ناخواسته در مادران بی‌سواد نسبت به‌مادرانی که تحصیلاتی بالاتر از دیپلم دارند، ۱/۵ برابر بیشتر است.

حال متغیر استفاده از روش پیشگیری از حاملگی را درنظر می‌گیریم. این متغیر دارای دو رده (۱— پیشگیری داشته و ۲— پیشگیری نداشته) می‌باشد.

$$\pi = \frac{\text{نسبت حاملگی ناخواسته در گروهی که پیشگیری داشته}}{\text{نسبت حاملگی ناخواسته در گروهی که پیشگیری نداشته}}$$

که برابر است با ۰/۵۱۹۸ و فاصله اطمینان ۹۵% برای این نسبت خطر برابر است با (۰/۴۵۱۷, ۰/۵۹۸۲) و داریم: $1/\pi = 1/9238$

بنابراین خطر وقوع حاملگی ناخواسته در مادرانی که از روشهای پیشگیری از بارداری استفاده می‌کنند نسبت به مادرانی که از روشهای پیشگیری از بارداری استفاده نمی‌کنند، تقریباً دو برابر بیشتر است.

۳.۳ بحث

استیودنت (۱۹۰۷) جزءِ اولین افرادی بود که از مدل دوجمله ای منفی به عنوان صورت دیگری از توزیع پواسن در داده‌های شمارشی استفاده کرد. مک کدریک (۱۹۱۴) و کیونویل (۱۹۴۹) توزیع های آمیخته در رابطه با توزیع دوجمله‌ای منفی را شرح دادند[٦].

گاردنر، مالوی و شاو (۱۹۹۵) مدل های پواسن و دوجمله‌ای منفی را، بر روی داده مربوط به وقوع خشونت در اشخاص مبتلا به ناخوشی‌های روانی، مورد مقایسه قرار

دادند. این داده‌ها از مطالعه‌ای مربوط به وقوع خشونت از جامعه‌ای شامل ۷۹۷ نفر که دچار ناخوشی‌های روانی بودند، برگرفته شده است. در این بررسی بیماران مراجعه کننده به بخش اورژانس یک بیمارستان روانی، مدت شش ماه پیگیری شده‌اند.

در مثال مذکور، داده‌ها تغییرات قابل توجهی را در وقوع خشونت بین بیماران نشان می‌دهد(وقوع ۷/۳ $M = ۳, SD = ۷/۳$).

رگرسیون مربوط به داده‌های خشونت شامل سه متغیر پیش بینی کننده برای خشونت می‌باشد:

۱) سن بیماران ($SD = ۱۱/۱$ و سال $X = ۲۸/٦$)

۲) اندازه‌گیری تکنسین‌های بخش اورژانس درباره اطمینان از اینکه بیمار دچار خشونت شده است.

۳) اندازه‌گیری سابقهٔ خشونت در بیماران قبل از اینکه به بخش اورژانس مراجعه کنند.

متغیرهای مذکور به ترتیب AGE ، CONCERN و HISTORY نام گذاری شده‌اند.

در داده‌های مربوط به خشونت، واضح است که نمی‌توان رگرسیون پواسن با ثابت μ را برازش کرد چون $Var(y) = ۵۲/٦$ بسیار بزرگتر از $y = ۳$ می‌باشد و درنتیجه ضرایب برآورد شده به وسیلهٔ مدل رگرسیون پواسن (که در جدول (۳) نشان داده شده است) بسیار کوچکتر از مقدار حقیقی آنها است و آزمونهای t مربوط به این ضرایب متوّرم شده است.

ضرایب برآورد شده به وسیلهٔ برازش مدل رگرسیون دوجمله‌ای منفی در جدول (۴) نشان داده شده است وهمانطور که ملاحظه می‌شود نتایج حاصله بسیار بهتر از نتایج به دست آمده به وسیلهٔ مدل پواسن است.

بنابراین در مدل بندی داده‌های شمارشی ابتدا باید میانگین و واریانس متغیر وابسته را محاسبه کرد. در صورتی که میانگین و واریانس به طور تقریبی برابر باشند، برای آنالیز داده‌ها از مدل رگرسیون پواسن استفاده می‌کنیم و در غیر اینصورت مدل رگرسیون دوجمله‌ای منفی را مورد استفاده قرار می‌دهیم.

به عبارت دیگر برای استفاده از مدل رگرسیون دوجمله‌ای منفی، ابتدا باید مقدار α را برآورد کرد. در صورتی که α اختلاف معنی داری از صفر داشت، رگرسیون دوجمله‌ای منفی برای آنالیز داده‌ها مدل مناسبی خواهد بود و در غیر اینصورت بهتر است که از مدل پواسن برای مدل بندی داده‌ها استفاده کنیم. در ایران مطالعات مختلفی در زمینهٔ حاملگی

ناخواسته صورت گرفته اما در هیچیک از آنها از رگرسیون پواسن و یا رگرسیون دوجمله‌ای منفی برای آنالیز داده‌ها استفاده نشده است.

متغیر	ضرایب برآورد شده	خطای استاندارد	t	$Pr(>\|t\|)$
AGE	−۰/۰۴۵	۰/۰۰۲۳	−۱۹/۶۹	۰/۰۰۰۱
CONCERN	۰/۰۸۳	۰/۰۰۷۵	۱۱/۲۰	۰/۰۰۰۱
HISTORY	۰/۴۲۰	۰/۰۳۸۰	۱۱/۲۶	۰/۰۰۰۱
ثابت	−۳/۴۱۰	۰/۰۶۹۰	−۴۹/۲۹	۰/۰۰۰۱

جدول ۳: آنالیز رگرسیون پواسن برای نرخ‌های خشونت

متغیر	ضرایب برآورد شده	خطای استاندارد	t	$Pr(>\|t\|)$
AGE	−۰/۰۴۵۹	۰/۰۰۷۹۹	−۵/۷۴	۰/۰۰۰۱
CONCERN	۰/۰۹۶۲	۰/۰۳۰۹۰	۳/۱۲	۰/۰۰۲۸
HISTORY	۰/۵۳۶۰	۰/۱۵۵۰۰	۳/۴۵	۰/۰۰۰۸
ثابت	−۳/۵۵۰۰	۰/۲۶۲۰۰	−۱۳/۵۵	۰/۰۰۰۱

جدول ۴: آنالیز رگرسیون دوجمله‌ای منفی برای نرخ‌های خشونت

مراجع

Mcculagh, P. and Nelder J.A. (1989), *Generalized Linear Models (2nd ed.)*, Chapman & Hall , London.

Piegrosh, W.W. (1990), *Maximum Likelyhood Estimation for the Negative Binomial Dispersion Parameter* , Biometrics., 49:863 P.

Denton, B. , Scott, S. and Chase, B. (1994), *Unintend and Unwanted Pregnancy in Halifax: the Rate and Associated Factors* , Con.J.Public Health, 234 p.

Gardner , W. , Mulvey E.P. and Shaw E.C. (1995), *Regression Analysis of Counts and Rates: Poisson, Overdispersed Poisson and Negative Binomial Models.*, Psycological Bulletin., 118(3):392 P .

Selvin , S. (1995), *Practical Biostatistical Methods*, Duxbury, USA.

Poch, M. and Mannering, F. (1996), *Negative Binomial Analysis of Intersection Accident Frequencies*, Journal of Transportation Engineering., March/April 105 P.

آل ابریشم کارزاده مهران. کاربرد مدل رگرسیون پواسن در تعیین رابطهٔ کیست تخمدان ناشی از کاشت نورپلانت با ریسک فاکتورهای مربوط. پایان نامهٔ کارشناسی ارشد آمار حیاتی، تهران: دانشکده علوم پزشکی دانشگاه تربیت مدرس، ۱۳۷۸.

اندرسن ارلینگ بی. الگوهای آماری گسسته و کاربرد آن در علوم اجتماعی. ترجمهٔ دکتر علی مشکانی و دکتر ابوالقاسم بزرگ نیا. چاپ اول، مشهد: دانشگاه فردوسی، ۱۳۷۲.

امامی افشار نزهت. حاملگی ناخواسته و ارتباط چند متغیر دموگرافی— اجتماعی و روشهای پیشگیری از بارداری در بین مراجعین به مراکز بهداشتی— درمانی جنوب تهران. محیط شناسی، زمستان ۱۳۷۰ .

منصوریان فریفته. کاربرد مدل‌های عمومی لگ خطی در تعیین عوامل مؤثر بر حاملگی ناخواسته و پیامدهای آن. پایان نامهٔ کارشناسی ارشد آمار حیاتی، تهران: دانشکده علوم پزشکی دانشگاه تربیت مدرس، ۱۳۷۵.

تعیین پارامتر پنجره در برآورد تابع چگالی توسط B—اسپلاین

دکتر محسن محمدزاده، رضا صالحی

P۱۳۰۱۹

گروه آمار، دانشگاه تربیت مدرس

چکیده: اغلب در مسائل کاربردی تابع چگالی احتمال نامعلوم است و لازم است بر اساس مشاهدات حاصل از یک نمونه تصادفی برآورد شود. در آمار روشهای مختلفی برای برآورد تابع چگالی بکار برده می‌شود که در این مقاله شیوهٔ برآورد B—اسپلاین مورد مطالعه قرار گرفته و روشی برای تعیین پارامتر پنجره، که میزان همواری برآورد را کنترل می‌کند، ارائه خواهد شد.

واژه‌های کلیدی: B—اسپلاین، پارامتر پنجره، برآورد چگالی احتمال.

۱ مقدمه

تابع چگالی احتمال یک مفهوم اساسی در آمار است که توسط آن رفتار متغیر تصادفی، احتمال پیشامدهای متناظر با آن و بسیاری از خواص متغیر تصادفی قابل توصیف می‌باشد. اما در اغلب مطالعات عملی این تابع نامعلوم است، در عوض یک مجموعه از n مشاهدهٔ $x_1, \ldots, x_n$ داریم، که فرض می‌شود مقادیر n متغیر تصادفی مستقل و هم توزیع و دارای تابع چگالی مجهول f هستند و معمولاً بر اساس این مشاهدات به یکی از دو روش پارامتری یا ناپارامتری برآورد می‌شود. در این مقاله ابتدا روش ناپارامتری برآورد یک تابع دلخواه توسط اسپلاین‌ها مرور شده سپس روش B—اسپلاین‌ها برای برآورد توابع چگالی احتمال مورد بررسی قرار می‌گیرد. چون یکی از مسائل مهم در روش B—اسپلاین‌ها تعیین پارامتر پنجره است، که میزان همواری تابع برآورد شده را کنترل می‌کند، روشی برای تعیین مقدار بهینه این پارامتر ارائه خواهد شد.

در تجزیه و تحلیل رگرسیون برای تعیین ارتباط بین متغیر پاسخ و متغیرهای مستقل و برازش منحنی به داده‌ها فرض‌های محدودکننده‌ای بمنظور حصول یک پاسخ منحصر بفرد در نظر گرفته می‌شود. در بسیاری از مواقع نوع ارتباط بین متغیر ها مشخص نیست و بجای مفروض داشتن یک الگوی خاص پارامتری، از روشی استفاده می شود که داده ها ماهیت روند خود را بهتر نشان دهند. اسپلاین هموار روشی است برای برآورد منحنی

که در آن تنها فرض همواری منحنی در نظر گرفته می شود. با در نظر گرفتن مدل:

$$Y_i = g(t_i) + \epsilon_i \quad , \qquad i = ۱, ..., n \qquad (۱)$$

که در آن g نامعلوم است، در حالت کلی نمی توان آنرا بطور منحصر بفرد برآورد کرد، مگر آنکه محدودیتی روی مدل اعمال شود. به عنوان مثال در روشهای رگرسیون خطی فرض می شود g تابعی خطی از پارامترهاست و به یکی از روشهای برآورد، پارامترهای آن بطور منحصر بفرد برآورد می‌شوند. گذاشتن فرض خطی بودن تابع g ممکن است برای برخی از مجموعهٔ داده ها مناسب باشد لیکن دارای کاستی‌هایی است که برای رهایی از آنها می‌توان g را تابعی دلخواه در نظر گرفت و با فرض هموار بودن، آن را برآورد نمود. میزان ناهمواری تابع g در بازهٔ $[a, b]$ را می توان توسط عبارت $\int_a^b [g''(t)]^۲ dt$ اندازه گیری نمود. لذا می‌توان در برآورد منحنی هموار g، علاوه بر عامل مجموع مربعات باقیمانده‌ها، میزان ناهمواری را نیز در نظر گرفته و ملاک مجموع مربعات جریمه‌ای

$$S(g, \alpha) = \sum_{i=۱}^{n} (Y_i - g(t_i))^۲ + \alpha \int_a^b [g''(t)]^۲ dt \qquad (۲)$$

را که در آن $\alpha > ۰$ پارامتر همواری نامیده می شود و تقابل بین نکویی برازش منحنی به داده‌ها و میزان ناهمواری تابع g را کنترل می‌کند، برای برآورد g بکار برد. تابعی مانند $\hat{g}$ که دارای مشتق مرتبهٔ دوم پیوسته و کمینه کنندهٔ مجموع مربعات جریمه‌ای (۲) باشد، اسپلاین همواری[۱] نامیده می شود. این شیوهٔ برآورد منحنی، رهیافت جریمهٔ ناهمواری نامیده می شود که در اوبانک (۱۹۸۸)، هاردل (۱۹۹۰) و راسنبلات (۱۹۹۱) مطالب مبسوطی در مورد آن آمده است. تعین مقدار مناسب پارامتر همواری α یک مسئلهٔ مهم است که برای مطالعه آن می توان به هاردل، مارون و هال (۱۹۸۸) و گرین و سیلورمن (۱۹۹۴) مراجعه نمود. یکی از روشهای پیدا کردن پارامتر هموارسازی، روش اعتبار متقابل و اعتبار متقابل تعمیم یافته می باشد که در استون (۱۹۷۴)، کریون و واهبا (۱۹۷۹)، گرین و سیلورمن (۱۹۹۴) آمده است. محمدزاده (۱۹۹۸) الگوریتمی مناسب برای تعین مقدار پارامتر هموارسازی بر اساس مشاهدات ارائه نموده است. برای مطالعه‌بیشتر در زمینه اسپلاین‌ها می توان به اوبانک (۱۹۸۴) و شاماکر (۱۹۹۳) مراجعه نمود. وقتی مدل (۸) به روش حداقل مربعات باقیمانده‌ها برازش داده شود، حالات زیر رخ می دهد. اگر نقاط بوسیلهٔ خط راست به یکدیگر متصل شده باشند، مجموع مربعات باقیمانده‌ها صفر خواهند شد. اگر شرط همواری g را اعمال کنیم و منحنی دارای مشتق مرتبهٔ دوم پیوسته باشد، منحنی از تمامی نقاط (t_i, Y_i) می گذرد. در این حالت نیز مجموع مربعات

[۱] Smoothing Spline

باقیمانده‌ها صفراست، اما منحنی مواج و دارای نوسانات زیادی است. مشتق مرتبهٔ دوم اندازهٔ ناهمواری منحنی را در نقاطی که منحنی تغییر جهت می‌دهد، مشخص می‌کند که ممکن است منفی یا مثبت باشد. مجموع توان دوم این مشتق‌ها روی تمام نقاط میزان ناهمواری منحنی را مشخص می‌کند. برای مقدار داده شدهٔ α، کمینه کردن $S(g,\alpha)$ بهترین راه توافق بین میزان همواری و نیکویی برازش مدل است. در روش اسپلاین همواری مقدار α نقش مهمی در برآورد منحنی ایفا می‌کند. اگر α بزرگ باشد، مولفهٔ اصلی در $S(g,\alpha)$ جملهٔ جریمهٔ ناهمواری است و بنابراین کمینه کنندهٔ g خیلی کم انحنا خواهد بود. در حد وقتی α به بی نهایت میل می‌کند، سهم عبارت $\int_a^b [g''(t)]^2 dt$ به سمت صفر میل می‌کند و منحنی $\hat{g}$ همان برازش رگرسیون خطی خواهد بود. از طرف دیگر اگر α نسبتاکوچک باشد، مجموع مربعات باقیمانده‌ها سهم اصلی را در $S(g,\alpha)$ دارد و براورد منحنی $\hat{g}$ تا حد زیادی روند داده‌ها را دنبال می کند، که ممکن است نوسانات زیادی داشته باشد. در وضعیت حدی اگر α به صفر نزدیک شود، برآورد حاصل به اسپلاین درونیاب[۲] میل می‌کند. برآورد حداقل مربعات جریمه‌ای، $\hat{g}$، کمینه کنندهٔ تابع $S(g,\alpha)$ روی کلاس تمام توابع دوبار مشتق پذیر g تعریف می‌شود. برآورد منحنی $\hat{g}$ که کمینه کنندهٔ تابع $S(g,\alpha)$ است، اسپلاین نامیده می‌شود. تابع $S(t)$ را اسپلاین از درجه r با گره‌های[۳] $t_1, ..., t_n$ گویند اگر $a = t_0 < t_1 < ... < t_n < t_{n+1} = b$ و

۱) برای هر $k = 0, ..., n$، $S(t)$ در فاصله (t_k, t_{k+1}) یک چند جمله‌ای درجه نابیشتر از r باشد.

۲) $S(t)$، $S'(t)$،...،$S^{(r-1)}(t)$ توابع پیوسته‌ای روی $[a,b]$ باشند.

هر چند جمله‌ای یک تابع اسپلاین بدون گره می‌باشد و بنابر تعریف بالا r امین مشتق یک اسپلاین از درجه r یک تابع ثابت تکه‌ای با شکستگی در نقاط $t_1, ..., t_n$ است. برعکس r امین تابع اولیه از یک تابع ثابت تکه‌ای یک اسپلاین از درجه r است. تابع g بر فاصله $[a,b]$ اسپلاین درجه سه نامیده می شود اگر برای اعداد حقیقی $a < t_1 < t_2 < ... < t_n < b$ در دو شرط زیر صدق کند:

۱—روی هر یک از فاصله‌های $(a,t_1), (t_1,t_2), ..., (t_n,b)$، g یک چند جمله‌ای درجه سه باشد.

۲—قطعات چند جمله‌ای در نقاط t_i به نحوی برازش یابند که مشتق‌های اول و دوم آن در هر t_i پیوسته باشد که در نتیجه روی تمام بازهٔ $[a,b]$ پیوسته است.

---

[۲] Interpolating Spline

[۳] Knots

۲ B—اسپلاین

B—اسپلاین‌ها رهیافت متنوعی برای بررسی منحنی‌ها در گرافهای کامپیوتری فراهم می‌کنند و مرکب از منحنی‌های چند جمله‌ای متصل در گره‌ها هستند و از درجه یکسان هستند، که درجه B—اسپلاین نامیده می‌شود. فرض کنید $U_1, \ldots, U_r$ یک نمونه تصادفی از توزیع یکنواخت بر فاصلهٔ $[-1/2, 1/2]$ باشد. چگالی مجموع این متغیرهای تصادفی بصورت

$$B^{(r)}(x) = \frac{\partial}{\partial x} P_r(U_1 + U_2 + \ldots + U_r < x), \qquad x \in R$$

است. هر تابع $B^{(r)}$ عضوی از کلاس $C^{(r-2)}$ یعنی توابع دارای مشتق پیوسته مرتبهٔ $r-2$ با تکیه‌گاه $[-r/2, r/2]$ است. برای هر عدد صحیح i، تابع $B^{(r)}(.)$ که به بازهٔ $(i + r/2, i+1+r/2)$ محدود شده است یک چند جمله‌ای بدیهی در خارج بازهٔ $[-r/2, r/2]$، یا از درجه r در داخل آن است. تابع $B^{(r)}(.)$، B—اسپلاین مرتبهٔ r نامیده می‌شود. بعضی B—اسپلاین‌های اساسی بصورت زیر می‌باشند:

برای $r = 1$

$$B^{(1)}(x) = \begin{cases} 1 & -\frac{1}{2} \le x < \frac{1}{2} \\ 0 & o.w. \end{cases}$$

برای $r = 2$، $B^{(2)} = B^{(1)} * B^{(1)}$ بنابراین

$$B^{(2)}(x) = \begin{cases} 1 + x & -1 \le x < 0 \\ 1 - x & 0 \le x \le 1 \\ 0 & o.w. \end{cases}$$

برای $r = 3$، $B^{(3)} = B^{(1)} * B^{(1)} * B^{(1)}$ بنابراین

$$B^{(3)}(x) = \begin{cases} \frac{1}{2}(x + \frac{3}{2})^2 & -\frac{3}{2} \le x < -\frac{1}{2} \\ -x^2 + \frac{3}{4} & -\frac{1}{2} \le x \le \frac{1}{2} \\ \frac{1}{2}(x - \frac{3}{2})^2 & \frac{1}{2} < x \le \frac{3}{2} \\ 0 & o.w. \end{cases}$$

و الی آخر. واضح است که $B^{(1)}$ یک تابع پله‌ای واحد و ثابت است. بنابراین $B^{(2)}$ خطی تکه‌ای، $B^{(3)}$ درجه دو تکه‌ای می‌باشد.

حال بوسیلهٔ انتقال و تغییر مقیاس $B^{(r)}$، می‌توان B—اسپلاین را برای مجموعه‌ای از گره‌ها روی R که بطوریکسان تقسیم شده‌اند، بدست آورد:

$$B^{(r)}_{s,h}(x) = B^{(r)}(\frac{x}{h} - s), \qquad x \in R, s \in Z, h > ۰ \qquad (۳)$$

برای هر عدد صحیح مثبت دلخواه r، هستهٔ اسپلاین بصورت

$$Q^{(r)}(x,y) = \sum_{s \in Z} B^{(r)}(x-s)B^{(r)}(y-s), \qquad (y,x) \in R^{۲} \qquad (۴)$$

است، که برای $h > ۰$ و $(x,y) \in R^{۲}$ هستهٔ مقیاس ساز[۱] بصورت

$$Q^{(r)}_h(x,y) = \frac{۱}{h} Q^{(r)}(\frac{x}{h}, \frac{y}{h}) \qquad (۵)$$

تعریف می‌شود. کرزیکوفسکی (۲۰۰۱) نشان داد هستهٔ اسپلاین (۴) دارای خواص زیر است:

$$\begin{aligned}
&Q^{(r)}(x,y) = Q^{(r)}(y,x) = ۰ && |y-x| > r \\
&\int Q^{(r)}(x,y)dx = ۱ && y \in R \\
&\int xQ^{(r)}(x,y)dx = y && y \in R\ , r > ۱ \\
&\int x^{۲}Q^{(r)}(x,y)dx = y^{۲} + \frac{r}{۶} && y \in R\ , r > ۲
\end{aligned}$$

۳ برآورد تابع چگالی

فرض کنید $X_۱, \ldots, X_n$ یک نمونه تصادفی مستقل با مقادیر حقیقی و چگالی مشترک $f_۰$ باشد. برآوردگر هسته‌ای $\hat{f}_h(x)$ برای برآورد تابع چگالی $f(x)$ در نقطه x بصورت زیر تعریف می‌شود:

$$\hat{f}_h(x) = \frac{۱}{nh} \sum_{i=۱}^{n} K(\frac{x_i - x}{h})$$

[۱] Scaling kernel

که در آن $K(.)$ تابع هسته‌ای نامیده می‌شود و می‌تواند هسته یکنواخت، مثلثی، درجه دو، گوسی، نمایی و غیره باشد و h پارامتر پهنای بازه است و میزان هموارى برآوردگر را کنترل می‌کند.

برآوردگر سری برای تابع $f(.)$ بصورت زیر تعریف می‌شود:

$$\hat{f}(x) = \frac{۱}{n}\sum_{k=۱}^{n} K_m(x - x_k)$$

که در آن

$$K_m(y) = \frac{۱}{۲\pi}\sum_{j=-m}^{m} e^{ijy}$$

و $i = -\sqrt{-۱}$ و m تعداد دوره‌های تناوب ۲π است یعنی برای $m = ۱$ فاصله $[-\pi, \pi)$ و برای $m = ۲$ فاصله $[-۲\pi, ۲\pi)$ بازه تغییرات داده‌ها برای برآورد تابع چگالی می‌باشد.

برآوردگر اسپلاین تابع چگالی بصورت زیر تعریف می‌شود:

$$\hat{f}_{n,h}(x) = \frac{۱}{n}\sum_{j=۱}^{n} Q_h^{(r)}(X_j, x), \qquad h \in R_+ \ , \ x \in R \ , \ r \geq ۱ \qquad (۶)$$

این برآورد نه از نوع برآوردهای هسته‌ای و نه از نوع سری می‌باشد، امّا بعضی خواص این دو برآوردگر را داراست. تشابه آن با برآوردگرهای سری بصورت

$$\hat{f}_{n,h}(x) = \sum_{s \in Z} a_{s,n,h} B_{s,h}^{(r)}(x), \qquad x \in R \ h \in R_+$$

است، که در آن

$$a_{s,n,h} = \frac{۱}{n}\sum_{j=۱}^{n} \frac{۱}{h} B_{s,h}^{(r)}(X_j)$$

به عبارت دیگر $\hat{f}_{n.h}(.)$ یک ترکیب خطی از B–اسپلاین است. برآوردگر (۶) بوسیلهٔ سیسیلسکی (۱۹۸۷) داده شده و برای بررسی عمیق‌تر این برآوردگر و همچنین برآورد چگالی اسپلاین چند متغیره می‌توان به‌سیسیلسکی (۱۹۹۰) مراجعه نمود.

۴ تعیین پارامتر پنجره

در این بخش راه حلی برای تعیین پارامتر پنجرهٔ $h \in R_+$ در رابطهٔ (٦) ارائه می‌شود برای این منظور میانگین نمونه‌ای $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$، یعنی برآوردگر نااریب با واریانس کمینه برای میانگین $m = \int x f_0(x)dx$ و $\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2 = \sigma_n^2$ برآوردگر نااریب با واریانس کمینه برای واریانس $\sigma^2 = \int (x-m)^2 f_0(x)dx$ را در نظر می‌گیریم. با توجه به خواص هسته اسپلاین برای تمام $\omega \in \Omega$ داریم :

$$\int \hat{f}_{n,h}(x)dx = \frac{1}{n}\sum_{j=1}^{n}\int \frac{1}{h}Q^{(r)}(\frac{X_j}{h}, \frac{x}{h})dx = 1$$

در نتیجه برآوردگر چگالی (٦) یک چگالی احتمال است. میانگین نمونه‌ای $\bar{X}_n$ و واریانس نمونه‌ای σ_n^2 بترتیب میانگین و واریانس چگالی $\hat{f}_{n,h}(.)$ هستند، پس داریم:

$$\begin{cases} \int x\hat{f}_{n,h}(x)dx = \bar{X}_n \\ \int (x-\bar{X}_n)^2\hat{f}_{n,h}(x)dx = \sigma_n^2 \end{cases}$$

این سیستم پایه‌ای برای تخمین پارامتر پنجرهٔ $h \in R_+$ خواهد بود. قضیه زیر روشی برای بدست آوردن پارامتر پنجره‌ای ارائه می‌دهد.

قضیه : اگر $\{Q_h^{(r)} : h \in R_+\}$ خانوادهٔ هسته‌های رابطهٔ (۵) و $X_n, \ldots, X_1$ یک نمونه تصادفی باشد، آنگاه برای برآوردگر $\hat{f}_{n,h}$ در رابطهٔ (٦) داریم:

$$\int x\hat{f}_{n,h}(x)dx = \bar{X}_n$$

و درست یک $h_n = h_n(X_1, \ldots, X_n) > 0$ وجود دارد بطوریکه

$$\int (x-\bar{X}_n)^2\hat{f}_{n,h}(x)dx = \sigma_n^2$$

و

$$h_n = h_n(X_1, \ldots, X_n) = \frac{\sigma_n}{\sqrt{kn}}, \qquad k = \frac{r}{6} \qquad (۷)$$

اثبات : بنا به خواص هسته اسپلاین داریم:

$$\begin{aligned} \int x\hat{f}_{n,h}(x)dx &= \frac{1}{n}\sum_{j-1}^{n}\int \frac{x}{h}Q^{(r)}(\frac{X_j}{h}, \frac{x}{h})dx \\ &= \frac{1}{n}\sum_{j=1}^{n}\int uhQ^{(r)}(\frac{X_j}{h}, u)du \end{aligned}$$

$$= h\frac{1}{n}\sum_{j=1}^{n}\frac{X_j}{h}$$
$$= \bar{X}_n$$

و

$$\int (x-\bar{X}_n)^2 \hat{f}_{n,h}(x)dx = \frac{1}{n}\sum_{j=1}^{n}\int x^2 \frac{1}{h}Q^{(r)}(\frac{X_j}{h},\frac{x}{h})dx - (\bar{X}_n)^2$$
$$= \frac{1}{n}\sum_{j=1}^{n} h^2 \int u^2 Q^{(r)}(\frac{X_j}{h},u)du - (\bar{X}_n)^2$$
$$= \frac{1}{n}\sum_{j=1}^{n} h^2((\frac{X_j}{h})^2 + \frac{r}{6}) - (\bar{X}_n)^2$$
$$= \frac{rh^2}{6} + S_n^2$$

که در آن $S_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j-\bar{X}_n)^2$ است. چون $\frac{r}{6}h^2 + S_n^2 = \sigma_n^2$ بنابراین

$$\frac{r}{6}h^2 + \frac{n-1}{n}\sigma_n^2 = \sigma_n^2$$

در نتیجه

$$h^2 = \frac{6\sigma_n^2}{rn} = \frac{\sigma_n^2}{n\frac{r}{6}}$$

اگر قرار دهیم $k = \frac{r}{6}$ آنگاه :

$$h = h_n = h_n(X_1,\ldots,X_n) = \frac{\sigma_n}{\sqrt{nk}}$$

و حکم ثابت است.

بنابراین برآوردگر چگالی براساس نمونهٔ تصادفی $X_1,\ldots,X_n$ به صورت

$$f_n(x) = \hat{f}_{n,h}(x) \qquad x \in R \qquad (8)$$

است، که در آن $h = h_n$ بوسیلهٔ رابطهٔ (۷) داده شده است. کرزیکوفسکی (۲۰۰۱) نشان داد که این برآوردگر سازگاری قوی در نرم L_1 دارد یعنی اگر $X_1,\ldots,X_n$ یک نمونه تصادفی با تابع چگالی f باشد، بطوریکه $E(X_1^2) < \infty$، در اینصورت برای تمام $x \in R$، با احتمال یک داریم

$$||f_n - f||_1 \longrightarrow 0$$

۵ بحث و نتیجه گیری

برآورد ناپارامتری چگالی با استفاده از B—اسپلاین‌ها تعمیمی از برآورد چگالی هیستوگرام‌ها می‌باشد. بطور مثال برای $r = ۱$، B—اسپلاین درجه صفر، همان هیستوگرام‌ها می‌باشند. از آنجا که برآورد چگالی بر مبنای اسپلاین‌هاست، درجه‌ای از هموارى بسادگی کنترل می‌شود. معمولاً در عمل $r = ۱, ۲, ۳, ۴$ در نظر می‌گیرند که برای $r = ۲$ یک چند جمله‌ای خطی، برای $r = ۳$ یک چند جمله‌ای درجه دو و برای $r = ۴$ یک چند جمله‌ای درجه سه می‌باشد. با توجه به نمونهٔ تصادفی داده شده، برآورد تابع چگالی احتمال با استفاده از B—اسپلاین‌های خطی، درجه دو یا درجه سه بسته به اینکه هر کدام خطای $L_۱$ کمتری داشته باشند صورت می‌گیرد.

برآورد چگالی احتمال برای یک نمونهٔ تصادفی مستقل $X_۱, \ldots, X_n$ بصورت ترکیب خطی از B—اسپلاین‌ها می باشد و زمانی که تعداد نمونه افزایش پیدا کند یعنی n بقدر کافی بزرگ باشد بر چگالی احتمال واقعی جامعه منطبق می‌شود.

مراجع

Ciesielski Z. (1987), Local Spline Approximation and Nonparametric Density Estimation, International Conference on Constructive Theory of Function, 25-31 May Varna.

Ciesielski Z. (1990), Asymptotic Nonparametric Spline Density Estimation, Probab. Math. Statist. 12, pp. 1-24.

Ciesielski Z. (1990), Asymptotic Nonparametric Spline Density Estimation in Several Variables. International Series of Numerical Mathematics (Birkhauser Verlag Basel) Vol. 94, pp. 25-53.

Creven, P. and Wahba, G. (1979), Smoothing Noisy Data with Spline Function, Number Math, 31, 377-390.

Eubank, R. L. (1984), Approximate Models and Spline, COMMUN. STATIS. Theory. Meth., 13(4), 433-484.

Eubank, R. L. (1988), Splines Smoothing and Nonparametric Regression, New York, Marcel Dekker.

Green, P. J. and Silverman, B. W. (1994), Nonparametric Regression and Generalized Linear Model, London, Chapman and Hall.

Hardle, W. (1990), Applied Nonparametric Regression, Cambridge: Cambridge University Press.

Hardle, W., Marron, J. S. and Hall (1988), How Far Are Automatically chosen regression Smoothing Parameters from their Optimum, Jornal of the American Statistical Association, Vol. 83, No. 401, Theory and methods.

Krzykowski G. (2001), On Automatic Choice of the Window Parameter in the Nonparametric Density Estimation, University of Gdansk. Poland.

Mohammadzadeh, M. (1998), An Algorithm to Find the Smoothing Parameter in Smoothing Splines, Proceeding of the 4th Iranian Statistical Conference, Shahid Beheshti University , Tehran.

Rossenblatt, M. (1991), Stochastic Curve Estimation, NSFCBMS Regional Conference Serries in Probability and Statistics, Vol 3, Hayward, California: Institute of Mathematical Statistics.

Schumaker, L. L. (1993), Splines Functions: Basic Theory. Florida, Krieger.

Stone, M. (1974), Cross Validatory Choice and Assessment of Statistical Prediction, Sco B.

استفاده از کریگینگ عام در همه‌گیری‌شناسی جغرافیایی بیماریها

محسن محمدزاده، انوشیروان کاظم‌نژاد، سقراط فقیه‌زاده، یداله واقعی

دانشگاه تربیت مدرس

چکیده: یکی از ابزارهای مهم برای تجزیه و تحلیل بیماریها در مبحث همه‌گیری‌شناسی جغرافیایی تهیه نقشه بیماری می‌باشد. اما نوسانات یا اغتشاشات ناشی از عواملی مانند خطاهای اندازه‌گیری و تغییرات سریع میزان بیماری در مناطق مختلف، تعبیر و تفسیر نقشه را مشکل می‌سازد. لذا برای حذف نوسانات شدید لازم است حتی‌الامکان نقشه‌های هموار بدست آورده شوند. نظر به اینکه میزان بیماری در مناطق مختلف اغلب دارای همبستگی فضایی هستند، در این مقاله روش کریگینگ عام برای تهیه نقشه آماری هموار بیماری ها معرفی شده و به کمک آن نقشه میزان بروز بیماری سل ریوی برای سالهای ۱۳۷۷ و ۱۳۷۸ شهرستانهای کشور تهیه و روند جغرافیایی بیماری در این دو سال مقایسه خواهند شد.

کلید واژه: نقشه آماری، کریگینگ عام، ناایستایی، سل ریوی

۱ مقدمه

معمولاً در همه‌گیری‌شناسی جغرافیایی از نقشه‌های بروز یا شیوع بیماریها برای نمایش توزیع جغرافیایی و تجزیه و تحلیل بیماری در بهداشت ملی ـ منطقه‌ای استفاده می‌شود. یکی از قدیمی‌ترین موارد استفاده از نقشه در همه‌گیری‌شناسی، مطالعه جان اسنو (۱۸۵۴) درباره همه‌گیری وبا در شهر لندن می‌باشد، که با استفاده از ترسیم موقعیت مرگ افراد و موقعیت جغرافیایی پمپ‌های تأمین آب، تأثیر منابع تأمین آب در انتشار وبا را مورد بررسی قرار داد. نقشه‌های بیماری برای دستیابی به تغییرات جغرافیایی کمبودهای بهداشتی و در نتیجه اختصاص منابع مالی و انسانی و تنظیم سیاستهای بهداشتی مورد استفاده قرار می‌گیرند. همچنین برای مطالعه ارتباط بین بروز بیماری با متغیرهای توضیحی مانند عوامل اقتصادی، اجتماعی و فرهنگی می‌توانند مفید باشند. این ارتباط توسط لاوسن و کرسی (۲۰۰۰) و لاوسن (۲۰۰۱) از طریق رگرسیون فضایی و تحت عنوان تجزیه و تحلیل اکولوژیکی مورد مطالعه قرار گرفته است. نقشه‌های معمولی میزان بروز یک

بیماری، معمولاً براساس اطلاعات موجود در بعضی یا همهٔ مناطق تهیه می‌شوند. عدم وجود میزان بیماری در بعضی مناطق یا وجود تغییرات جغرافیایی خطای اندازه‌گیری ممکن است اینگونه نقشه‌ها را دچار ناهمواری و نوسانات زیادی کند که در اینصورت تعبیر و تفسیر آنها بسیار دشوار خواهد بود. لذا برای رفع این مشکل لازم است، نقشه آماری به گونه‌ای هموار شود تا بتوان از آن برای تخمین میزان بیماری در مناطق فاقد اطلاع استفاده نمود.

در سه دههٔ گذشته روشهای مختلفی به منظور تخمین مقدار یک متغیر مانند میزان بیماری در موقعیتی مشخص با استفاده از مشاهدات واقع در موقعیتهای مجاور توسعه یافته است، که مهمترین آنها عبارتند از: روش عکس فاصله موزون[۱] (ریپلی، ۱۹۸۱ و واتسون و فیلیپ، ۱۹۸۵) که توسط پوکالا (۱۹۸۹) و گلاتر (۱۹۸۹) برای تهیه نقشه آماری بیماری مورد استفاده قرار گرفته است، روش سطوح روند چندجمله‌ای[۲] (واتسون، ۱۹۷۲ و ریپلی، ۱۹۸۱)، روش اسپلاین‌ها[۳] (دی‌بور، ۱۹۷۸ و لیند و همکاران، ۱۹۹۵) و روشهای مبتنی بر بیز (کلایتون و کالدر، ۱۹۸۷، کلایتون و برناردینلی، ۱۹۹۲، برناردینلی و همکاران، ۱۹۹۵ و ماگلین و همکاران، ۲۰۰۰). روش کریگینگ از جمله فنون آماری است که برای تخمین فضایی و تهیه نقشه آماری بیماری به کار برده می‌شود و علیرغم روشهای مذکور، ساختار همبستگی فضایی میزانهای بیمای را در تجزیه و تحلیل آنها مورد استفاده قرار می‌دهد. بعلاوه با استفاده از این روش می‌توان در هر موقعیت دلخواه انحراف معیار تخمین میزان بیماری را نیز تعیین کرد و در صورت نیاز نقشه آن را تهیه نمود (کارات و والرون، ۱۹۹۲ و کرسی، ۱۹۹۳).

چون میزانهای بروز بیماری در مناطق مختلف اغلب از یک ساختار همبستگی پیروی می‌کنند، که به موقعیت آنها بستگی دارد، نوعی داده فضایی[۴] هستند و این ساختار همبستگی باید در تجزیه و تحلیل آنها لحاظ گردد. در این مقاله نحوهٔ تعیین ساختار همبستگی داده‌ها و استفاده از روش کریگینگ برای تخمین فضایی میزانهای بیماری بیان شده و نقشه هموار بیماری سل ریوی سالهای ۱۳۷۷ و ۱۳۷۸ با استفاده از داده‌های سل ریوی اسمیر مثبت شهرستانهای کشور که توسط اداره کل مبارزه با بیماریها جمع‌آوری شده است، تهیه می شود.

---

[۱] Inverse Distance-Weighted
[۲] Trend Polynomial Surfaces
[۳] Splines
[۴] Spatial Data

۲ مفاهیم و روشهای آماری

داده‌هایی که نوعاً بر حسب موقعیت (مکان) قرارگرفتن آنها در فضای مورد مطالعه همبسته باشند و این همبستگی تابعی از فاصله موقعیت آنها باشد، داده‌های فضایی نامیده می‌شوند. میزان بروز یا شیوع یک بیماری در شهرستانهای مختلف و میزان گازهای آلاینده هوا در ایستگاههای مختلف سنجش آلودگی هوا مصداق‌هایی از مجموعه داده‌های فضایی، در حیطه‌های علوم پزشکی و محیط زیست هستند.

دو مشخصه داده‌های فضایی را از داده‌های معمولی متمایز می‌کند، اولاً هر داده فضایی با موقعیت آن در فضای مورد مطالعه همراه است و ثانیاً این داده‌ها به هم وابسته و این وابستگی مرتبط با موقعیت قرارگرفتن آنها است. داده‌های فضایی با نماد $z(\boldsymbol{t}_1)$،...،$z(\boldsymbol{t}_n)$ نشان داده می‌شود، که در آن n تعداد داده‌ها و $\boldsymbol{t}_i = (x_i, y_i)$ طول و عرض متناظر با موقعیت جغرافیایی مشاهده i ام (در فضای دوبعدی) است. ساختار همبستگی داده‌های فضایی به وسیله تابع تغییرنگار[۱] مشخص می‌شود. این تابع همبستگی داده‌های فضایی را برحسب فاصله نشان می‌دهد و معمولاً مقدار آن با زیاد شدن فاصله افزایش می‌یابد، که نشان‌دهندهٔ کاهش همبستگی بین داده‌ها می‌باشد. ماترون (۱۹۶۳) برآورد تجربی تغییرنگار را به صورت

$$۲\hat{\gamma}(\boldsymbol{h}) = \frac{۱}{|N(\boldsymbol{h})|} \sum_{i=۱}^{|N(h)|} (z(\boldsymbol{t}_i) - z(\boldsymbol{t}_i + \boldsymbol{h}))^۲ \qquad (۱)$$

معرفی نمود، که در آن $N(\boldsymbol{h}) = \{(i,j) : \boldsymbol{t}_i - \boldsymbol{t}_j = \boldsymbol{h}\}$ مجموعه تمام زوج داده‌های در فاصله $\boldsymbol{h}$ از یکدیگر و $|N(\boldsymbol{h})|$ تعداد زوجهای متمایز این مجموعه می‌باشد.

در حالت کلی ساختار همبستگی داده‌ها به فاصله و جهت موقعیت آنها در فضای مورد مطالعه بستگی دارد. اگر همبستگی داده‌ها فقط تابعی از اندازه فاصله بین موقعیت آنها باشد، اصطلاحاً همسانگرد[۲] نامیده می‌شوند. برای برازش یک مدل تغییرنگار مناسب به داده‌ها معمولاً فاصله مورد مطالعه به تعدادی زیرفاصله یا لَگ[۳] تقسیم می‌شود و با استفاده از زوج داده‌هایی که در هر لگ قرار می‌گیرند مقدار تغییرنگار تجربی محاسبه می‌شود و به یکی از روشهای حداقل مربعات معمولی، حداقل مربعات موزون یا درستنمایی یک مدل معتبر به آنها برازش داده می‌شود. نظر به اینکه انتخاب تعداد لگ در میزان دقت

---

[۱] Variogram
[۲] Isotropic
[۳] Lag

برازش مدل تغییرنگار مؤثر است، محمدزاده و همکاران (۱۳۸۰) تعداد لگ بهینه را برای برازش مدل تغییرنگار معرفی کردند.

کرسی (۱۹۹۳) نشان داد چنانچه داده‌های فضایی دارای روند باشند، برآورد تغییرنگار اریب است و برای رفع این نقیصه برآورد تغییرنگار براساس داده‌های بدون روند شده را توصیه نمود. برای این کار یک منحنی رگرسیون چندجمله‌ای به موقعیتهای (x_i, y_i) برازش داده می‌شود و از تفاضل مقدار منحنی روند از هر مشاهده، داده‌های بدون روند شده به دست می‌آیند و تغییرنگار تجربی آنها محاسبه می‌شود. لیکن تغییرنگار داده‌های اصلی و داده‌های بدون روند شده یکسان نیست و به روش حذف روند از داده‌ها بسستگی دارد. اما محمدزاده (۱۳۷۸) نشان داد استفاده از تابع کوواریانس تعمیم یافتهٔ [۱] داده‌های بدون روند شده، مشکل تعیین ساختار همبستگی داده‌ها را تا حدی مرتفع می‌سازد.

ماترون (۱۹۶۳) تخمین فضایی در یک موقعیت مشخص $\boldsymbol{t}_0 = (x_0, y_0)$ براساس داده‌های $z(\boldsymbol{t}_1)$،...،$z(\boldsymbol{t}_n)$ را به نام دی. جی. کریگ، مهندس معدن آفریقای جنوبی، کریگینگ نامید. کریگینگ همان بهترین تخمین کنندهٔ خطی نااریب $(BLUP)$ به صورت $\sum_{i=1}^{n} \lambda_i z(\boldsymbol{t}_i)$ است، که در آن ضرایب λ_1،...،λ_n به گونه‌ای تعیین می‌شوند که تخمین کننده، نااریب و دارای کمترین واریانس باشد. ضرایب λ_i در کریگینگ به گونه‌ای خواهند بود که به مشاهدات نزدیک وزن بیشتر و به مشاهدات دور وزن کمتر اختصاص داده می‌شود. این امر به تبعیت از ماهیت داده‌های فضایی است که با افزایش فاصله، همبستگی آنها کاهش می‌یابد.

کریگینگ شامل سه نوع متداول ساده، عادی و عام است که ماترون (۱۹۶۹ و ۱۹۷۳) و کرسی (۱۹۹۳) آنها را مورد بررسی قرار داده‌اند. کریگینگ عادی در مواقعی به کار می‌رود که داده‌ها فاقد روند فضایی باشند و درصورت وجود روند از روش کریگینگ عام استفاده می‌شود. جزئیات بیشتر و محاسبات کریگینگ عام را می‌توان در کرسی (۱۹۹۳) و وبستر و اُلیور (۲۰۰۱) ملاحظه نمود.

۳ تعیین ساختار همبستگی میزان بروز سل ریوی

مختصات جغرافیایی مراکز شهرستانها نسبت به یک مبدأ واقع در جنوب غربی کشور به صورت $\boldsymbol{t}_i = (x_i, y_i)$ بر حسب کیلومتر از نقشه تقسیمات کشوری ۱۳۷۸ استخراج شده است. سپس از تقسیم تعداد موارد جدید سالیانه سل ریوی اسمیر مثبت بر برآورد جمعیت ۲۶۲ شهرستان کشور میزان بروز بیماری سل ریوی اسمیر مثبت براساس صد هزار نفر

[۱] Generalized Covariance Function

جدول ۱: مجموع مربعات موزون (WSS) برای مدلهای تغییرنگار برازش شده به میزانهای بروز سل ریوی

مدل									سال
Cubic	Gneiting	Powerd Exponential	Circular	Sphricaal	Gaussian	Wave	Matern	Exponential	
۴۲٫۷۹	۱۲۶/۴۵	۷۴٫۰۷	۱۰۱٫۰۱	۱۱۲٫۴۰	۵۴٫۸۳	۴۴٫۰۵	۸۰٫۱۲	۸۰٫۱۲	۱۳۷۷
۵۷٫۰۲	۱۵۵/۲۵	۹۴٫۲۹	۱۲۶٫۵۵	۱۲۳٫۸۲	۷۰٫۳۳	۵۹٫۴۲	۱۰۱٫۸۲	۱۰۱٫۸۲	۱۳۷۸

برای سالهای ۱۳۷۷ و ۱۳۷۸ محاسبه و با $z(\boldsymbol{t}_i)$ نشان داده شده است.

بدلیل وجود روند فضایی در داده‌های مذکور با برازش چندجمله‌ای درجه دو به صورت $\mu(\boldsymbol{t}_i) = a_0 + a_1 x_i + a_2 y_i + a_3 x_i^2 + a_4 y_i^2 + a_5 x_i y_i$، روند هر یک از مجموعه داده‌های مربوط به سالهای ۱۳۷۷ و ۱۳۷۸ تعیین و داده‌ها به صورت $R(\boldsymbol{t}_i) = z(\boldsymbol{t}_i) - \hat{\mu}(\boldsymbol{t}_i)$ بدون روند شده‌اند. سپس تغییرنگار تجربی همسانگرد داده‌های بدون روند شده در ۱۸ لگ (محمدزاده و همکاران، ۱۳۸۰) محاسبه شده است. هشت مدل متداول تغییرنگار (باری و همکاران، ۱۹۹۷) به روش حداقل مربعات موزون، به تغییرنگار تجربی برازش داده و مجموع مربعات موزون (WSS) مدلهای برازش شده به تفکیک سال ۱۳۷۷ و ۱۳۷۸ در جدول ۱ نشان داده شده است. معیار مجموع مربعات موزون، اختلاف مدل برازش شده با تغییرنگار تجربی را اندازه می‌گیرد و هر چه مقدار آن کمتر باشد مدل برازش شده بهتر خواهد بود. با توجه به نتایج مندرج در جدول ۱ مدل شبه تغییرنگار Cubic همسانگرد

$$\gamma(|\boldsymbol{h}|) = \begin{cases} c_0 + c\left[7(\frac{|h|}{a})^2 - 8/75(\frac{|h|}{a})^3 + 3/5(\frac{|h|}{a})^5 - 0/75(\frac{|h|}{a})^7\right] & , \quad |\boldsymbol{h}| < a \\ c_0 + c & , \quad |\boldsymbol{h}| \geq a \end{cases}$$

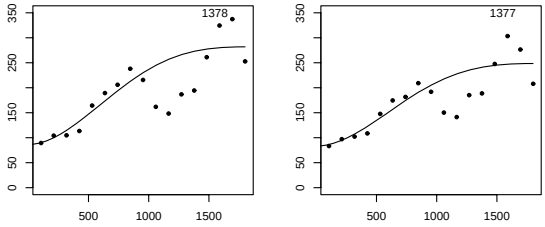
به عنوان بهترین مدل انتخاب شد، که مقادیر پارامترهای c_0، c و a به همراه مجموع مربعات موزون (WSS) به تفکیک سال در جدول ۲ نشان داده شده است. شکل ۱ نمودار تغییرنگار تجربی و مدل برازش شده Cubic همسانگرد را برای سالهای مختلف نشان می‌دهد.

۴ تخمین میزان بروز سل ریوی در چند شهرستان جدید

طبق تقسیمات کشوری سال ۱۳۷۸، جمهوری اسلامی ایران دارای ۲۸۲ شهرستان است که داده‌های سل ریوی ۲۶۲ شهرستان موجود و اطلاعی از میزان بیماری در ۲۰ شهرستان

جدول ۲: برآورد پارامترهای مدل تغییرنگار Cubic برای میزانهای بروز سل ریوی

سال	$\hat{c}_\circ$	$\hat{c}$	$\hat{a}$	WSS
۱۳۷۷	۴۲٫۸۸	۸۱٫۱۶	۱۹۶۷٫۲۵	۴۲٫۷۹
۱۳۷۸	۴۴٫۵۴	۹۵٫۳۲	۱۹۶۹٫۸۷	۵۷٫۰۲



شکل ۱: تغییرنگار تجربی و مدل Cubic برازش شده برای سالهای ۱۳۷۷ و ۱۳۷۸

دیگر در دسترس نمی‌باشد. به منظور نمایش قابلیتهای روش کریگینگ و تکمیل اطلاعات بیماری مطابق با تقسیمات کشوری، میزان بروز بیماری سل در این شهرستانها برای سالهای ۱۳۷۷ و ۱۳۷۸ به روش کریگینگ عام تخمین زده شده و به همراه انحراف معیار آنها در جدول ۳ نشان داده شده است.

هنگامی که در چند موقعیت جدید تخمین انجام می‌شود، چون در این موقعیتها هیچ مشاهده‌ای در دسترس نیست نمی‌توان خطای پیش‌بینی و یا خطای کلی مدل تغییرنگار را ارزیابی کرد. در اینگونه موارد می‌توان به کمک اعتبارسنجی متقابل[۱] (استون، ۱۹۷۴) معیاری را برای دقت تخمینها ارائه داد. در این شیوه یک مشاهده حذف و مقدار آن براساس $n-1$ مشاهدهٔ باقیمانده به روش کریگینگ تخمین زده می‌شود و از اختلاف مقدار واقعی و تخمین آن، خطای کریگینگ در یک موقعیت محاسبه می‌شود. این عمل را برای تمام n موقعیت تکرار و جذر میانگین مربع اختلافهای استاندارد به صورت:

$$RMSP = \left\{ \frac{1}{n} \sum_{j=1}^{n} \left[\frac{z(\boldsymbol{t}_j) - \hat{z}_{-j}(\boldsymbol{t}_j)}{\sigma_{-j}(\boldsymbol{t}_j)} \right]^2 \right\}^{\frac{1}{2}}$$

محاسبه می‌شود، که در آن $z(\boldsymbol{t}_j)$ مقدار واقعی مشاهده در موقعیت $\boldsymbol{t}_j$ و $\hat{z}_{-j}(\boldsymbol{t}_j)$ تخمین آن براساس بقیه مشاهدات بجز $z(\boldsymbol{t}_j)$ می‌باشد و $\sigma_{-j}(\boldsymbol{t}_j)$ انحراف معیار تخمین $\hat{z}_{-j}(\boldsymbol{t}_j)$

[۱] Cross-Validation

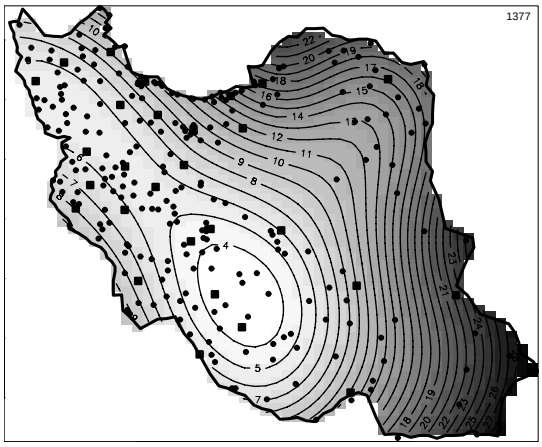
جدول ۳: تخمین فضایی میزان بروز بیماری سل ریوی در ۲۰ شهرستان

نام شهرستان	طول	عرض	سال ۱۳۷۷		سال ۱۳۷۸	
			میزان	انحراف معیار	میزان	انحراف معیار
اسکو	۲۱۹٫۲۰	۱۳۰۸٫۸۰	۶٫۱۲	۶٫۶۶	۴٫۸۸	۶٫۸۰
جلفا	۱۷۲٫۸۰	۱۴۲۰٫۸۰	۷٫۲۹	۶٫۸۰	۵٫۷۴	۶٫۹۴
چالدران	۵۹٫۲۰	۱۴۳۲٫۰۰	۶٫۸۵	۷٫۰۲	۴٫۶۲	۷٫۱۸
کوثر	۴۰۸٫۰۰	۱۲۸۴٫۸۰	۷٫۶۶	۶٫۶۴	۷٫۵۲	۶٫۷۷
نیر	۳۷۹٫۲۰	۱۳۳۲٫۸۰	۸٫۰۸	۶٫۶۶	۷٫۷۷	۶٫۷۹
ایجرود	۴۲۵٫۶۰	۱۱۶۰٫۰۰	۶٫۲۶	۶٫۶۳	۶٫۳۹	۶٫۷۶
خرمدره	۵۰۸٫۸۰	۱۱۳۶٫۰۰	۶٫۸۹	۶٫۶۳	۷٫۳۷	۶٫۷۶
املش	۵۸۷٫۲۰	۱۴۰۴٫۸۰	۱۴٫۰۳	۶٫۸۱	۱۴٫۴۲	۶٫۹۶
رضوانشهر	۴۸۸٫۰۰	۱۲۸۸٫۰۰	۹٫۱۱	۶٫۶۵	۹٫۲۸	۶٫۷۸
سیاهکل	۵۵۶٫۸۰	۱۲۴۴٫۸۰	۹٫۵۱	۶٫۶۴	۹٫۹۰	۶٫۶۷
ماسال	۴۸۸٫۰۰	۱۲۶۸٫۸۰	۸٫۷۳	۶٫۶۴	۸٫۹۱	۶٫۷۷
جویبار	۸۲۷٫۲۰	۱۲۱۹٫۲۰	۱۴٫۱۱	۶٫۶۶	۱۴٫۲۷	۶٫۷۹
بندرگز	۹۲۰٫۰۰	۱۲۴۴٫۸۰	۱۶٫۶۹	۶٫۶۹	۱۶٫۸۳	۶٫۸۲
پاکدشت	۷۶۸٫۰۰	۱۰۸۰٫۰۰	۹٫۷۸	۶٫۶۴	۱۰٫۲۵	۶٫۷۷
نیران و کرون	۷۲۰٫۰۰	۷۶۸٫۰۰	۴٫۵۱	۶٫۶۴	۵٫۰۲	۶٫۷۷
ارسنجان	۹۵۸٫۴۰	۴۸۶٫۶۰	۳٫۴۵	۶٫۶۸	۲٫۷۹	۶٫۸۱
خرم بید	۹۳۷٫۶۰	۵۶۰٫۰۰	۳٫۳۲	۶٫۶۸	۲٫۸۲	۶٫۸۱
جاجرم	۱۱۳۱٫۲۰	۱۲۹۴٫۴۰	۱۸٫۸۴	۶٫۷۶	۱۸٫۸۳	۶٫۸۹
ابوموسی	۱۱۸۵٫۶۰	۵۴٫۴۰	۱۱٫۱۰	۷٫۱۴	۱۰٫۱۷	۷٫۳۱
راور	۱۲۷۲٫۰۰	۶۷۶٫۸۰	۱۰٫۱۹	۶٫۷۸	۸٫۵۶	۶٫۹۳

است. وقتی در کریگینگ مدلهای متفاوتی برای تغییرنگاربه کار گرفته می‌شود، می‌توان میزان دقت تخمین حاصل از کریگینگ‌های مختلف را براساس معیار $RMSP$ با یکدیگر مقایسه نمود. بدیهی است هر چقدر مقدار این معیار به یک نزدیکتر باشد کریگینگ از دقت بیشتری برخوردار می‌باشد. این معیار برای میزانهای بروز سل ریوی ۱۳۷۷ برابر ۱٫۲۶ و برای سال ۱۳۷۸ برابر ۱٫۲۴ شده است، که به مقدار مورد انتظار یک نزدیک می‌باشند. نظر به اینکه تغییر مدل تغییرنگار سبب تغییر دقت تخمین‌ها می‌شود، معیار $RMSP$ به طور غیرمستقیم دقت مدل تغییرنگار برازش شده را نیز نشان می‌دهد، ممکن است بتوان با یافتن مدلهای تغییرنگاری که منجر به کاهش $RMSP$ شوند، تخمینهای دقیقتری را نیز به دست آورد.

۵ نقشه آماری

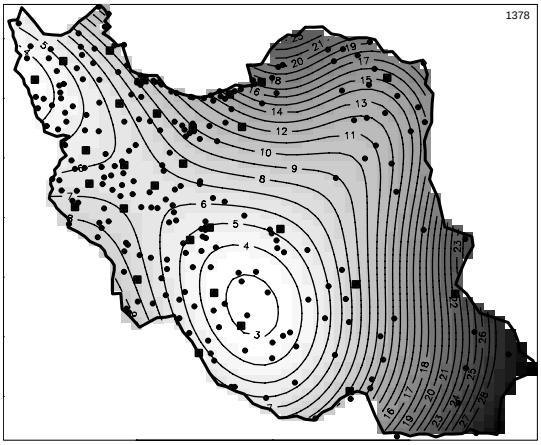
برای تهیه نقشه آماری بروز سل ریوی، کشور در یک شبکه منظم ٦٧ × ٥٠ با فواصل ۲۸/۸ کیلومتر محاط گردیده است. در تمام گره‌های شبکه واقع در داخل کشور میزان بروز بیماری سل ریوی به روش کریگینگ عام، با روند درجه دوم تخمین زده شده و انحراف معیار آنها محاسبه شده‌اند. شکلهای ۲ و ۳ نقشه‌های تخمین میزان بروز بیماری و شکلهای ۴ و ۵ نقشه‌های انحراف معیار تخمین را در سالهای ۱۳۷۷ و ۱۳۷۸ نشان می‌دهند. در این نقشه‌ها مراکز شهرستانها با نقطه و مراکز استانها با مربع نمایان شده‌اند. افزایش تیرگی رنگ در شکلهای ۲ و ۳ از مناطق مرکزی به سمت مرزهای شرقی کشور نشان‌دهندهٔ کمترین میزان بروز بیماری در مناطق مرکزی است، که به تدریج با نزدیک شدن به سمت مرزهای جنوب شرقی و قسمتی از شمال شرق به بیشترین مقدار می‌رسد.



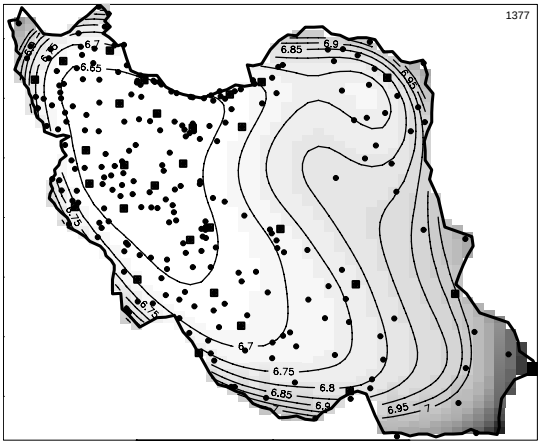
شکل ۲: نقشه میزان بروز سل ریوی، ۱۳۷۷

با توجه به شکلهای ۴ و ۵ ملاحظه می‌شود کمترین انحراف معیار مربوط به نیمه غربی کشور است که در آن تجمع شهرستانها بیشتر و در نتیجه داده‌های زیادتری در اختیار هستند. افزایش انحراف معیار به سمت شرق می‌تواند ناشی از پراکندگی شهرستانها و در نتیجه کم بودن تعداد داده‌ها و حتی ناشی از تغییرات زیاد بیماری در این مناطق باشد. بطور کلی انحراف معیار تخمین‌ها در سطح کشور نوسان زیادی ندارد و این نشان‌دهندهٔ آن است که دقت تخمین‌ها تقریباً یکسان است.

با توجه به شکلهای ۲ و ۳ در اغلب مناطق کشور میزان بیماری در سال ۱۳۷۸ نسبت به ۱۳۷۷ کاهش یافته است. به منظور بررسی روند جغرافیایی کاهش یا افزایش بیماری

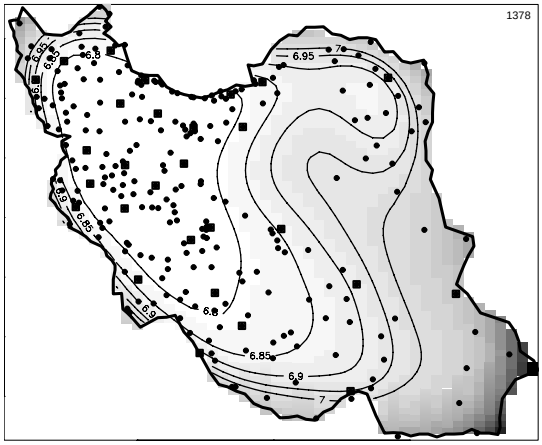


شکل ۳: نقشه میزان بروز سل ریوی، ۱۳۷۸

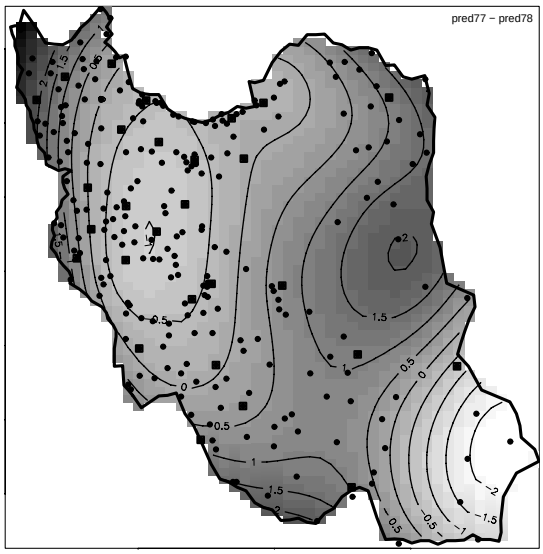


شکل ۴: نقشه انحراف معیار تخمین میزان بروز بیماری، ۱۳۷۷

با کم کردن تخمین میزان بیماری در سال ۱۳۷۸ از ۱۳۷۷، در گره‌های شبکه فوق‌الذکر، نقشه‌ای از تغییرات زمانی بیماری به دست آمده و در شکل ٦ ترسیم شده است. در این نقشه اعداد مثبت نشان‌دهندهٔ مقدار کاهش و اعداد منفی نشان‌دهندهٔ مقدار افزایش میزان بیماری (برحسب صدهزار نفر) می‌باشد. بنابراین در اکثر مناطق کشور، بجز قسمتی از مناطق مرکزی و جنوب شرقی کشور میزان بیماری در سال ۱۳۷۸ کاهش یافته است. در قسمت کوچکی از مناطق جنوب شرق کشور میزان بیماری سل ریوی افزایش یافته به گونه‌ای که در نزدیک مرز این افزایش به سه درصدهزار نفر می‌رسد.



شکل ۵: نقشه انحراف معیار تخمین میزان بروز بیماری، ۱۳۷۸



شکل ٦: نقشه کاهش میزان بروز بیماری از ۱۳۷۷ به ۱۳۷۸

٦ بحث و نتیجه‌گیری

دراین مقاله روش تخمین فضایی کریگینگ عام برای تجزیه و تحلیل داده‌ها در همه‌گیری‌شناسی جغرافیایی بیماریها و تهیه نقشه آماری آنها ارائه شد و داده‌های سل ریوی اسمیر مثبت سالهای ۱۳۷۷ و ۱۳۷۸ شهرستانهای کشور، به عنوان یک مطالعه موردی، مورد بررسی قرار گرفتند، که میزان دقت نتایج تا حدودی تحت تأثیر کم‌ثبتی بیماری ناشی از عدم

مراجعه بعضی بیماران به شبکه بهداشت یا عدم تشخیص بیماری می‌باشد. صرف‌نظر از آن نقشه‌های حاصل از روشهای ارائه شده، روند جغرافیایی بیماری را بطور کامل نشان می‌دهند. روند افزایشی میزان بروز بیماری در شرق ایران می‌تواند دلایل همه‌گیری‌شناسی خاصی داشته باشد، که در اینجا مورد بررسی قرار نگرفته‌اند.

در این مقاله کریگینگ براساس یک مدل تغییرنگار همسانگرد که به روش حداقل مربعات موزون به داده‌ها برازش داده شده صورت پذیرفته است. با توجه به امکان وجود ناهمسانگری تأثیر بسزایی در میزان دقت تخمین‌ها و نقشه بیماریها داشته باشد، که لازم است در مطالعات بعدی مورد توجه قرار گیرد.

کلیه محاسبات و تهیه نمودارها با برنامه‌نویسی در محیط نرم‌افزار SPLUS و به کمک مجموعه توابع GeoS (ریبیرو و دیگل، ۲۰۰۰) صورت گرفته است.

از همکاری صمیمانه واحد سل اداره کل مبارزه با بیماریها که داده‌ها و راهنماییهای فنی در مورد نحوه ثبت و جمع‌آوری آنها را ارائه نموده‌اند تشکر و قدردانی می‌شود.

مراجع

Barry, J., Crowder, M. and Diggle, P., (1997). Parametric Estimation of the Variogram, *Technical Report* ST-97-06. Dep. Maths and Stats, Lancaster University, Lancaster, UK.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. and Songini, M., (1995). Baysian Analysis of Space-Time Variation in Disease Risk, *Statistics in Medicine*,, **14**: 2433-2443.

Carrat, F. and Valleron A., J., (1992). Epidemiologic Mapping Using the Kriging Method: Application to an Influenza-Like Illness epidemic in France, *American Journal of Epidemiology*,, **135**: 1293-1300.

Clayton, D. and Bernardinelli, L., (1992). Baysian Methods for Mapping Disease Risk; In: *Geographical and Environmental Epidemiology*: Methods for Small-Area Studies (Eds., Elliott, P., Cuzick, J., English, D., and Stern R.), Oxford Press, Oxford.

Clayton, D. G. and Kaldor, J., (1987). Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping, *Biometrics,*, **43**: 671-691.

Cressie, N., (1993). *Spatial Statistics for Spatial Data,* New York, John Wiley.

De Boor C., (1978) *A Practical Guide to Splines.* New York: Springer-Verlag New York, Inc.

Glattre, E., (1989). *Atlas of Cancer Incidence in Norway 1970-1979.* Recent Results Cancer Res, **114**: 216-226.

Lawson, A. B., (2001). *Statistical Methods in Spatial Epidemiology,* John Wiley.

Lawson, A. B., Cressie, N., (2000). *Spatial Statistical Methods for Environmental Epidemiology,* Handbook of Statistics, **18**: 357-396.

Lawson, A. B. and Williams, L. R., (2001). *An Introductory Guide to Disease Mapping,* John Wiley.

Linde, V. D. A., Witzko, K. H. and Jaockel, K., H., (1995). Spatial-Temporal Analysis of Mortality Using Splines, *Biometrics,*, **51**: 1352-1360.

Marshall, R. A., (1991). Rview of Methods for the Statistical Analysis of Spatial Patterns of Disease, *Journal of Royal Statistical Society, Series A,*, **154**, 421-441.

Matheron, G., (1963). Principles of Geostatistics, *Economic Geology,*, **58**: 1246-1266.

Matheron, G., Le Krigeag, Universal, (1969). *Cahiers du Center de Morphologie Mathematique,*, No. 1 Fountainbleau, France.

Matheron, G., (1973). The Intrinsic Random Functions and their Applications, *Adv. Appl. Prob.,* No. 5, 439-468.

Mugglin, A. S., Cressie, N., and Gemmell, I., Hierarchial, (2000). Statistical Modeling of Influenza-Epidemic Dynamic in Space and Time, (Tentative title), in preparation.

Pukkala, E., (1989). *Cancer Maps of Finland: An Example of Small-Area Based Mapping.* Recent Results Cancer Res, **114**: 208-215.

Ribeiro, Jr. P. J. and Diggle, P. J., **geoR/geoS**: Functions for Geostatistical Analysis Using R or S-PLUS Technical Report: ST-99-09, Dep. Of Maths. And Stats. Lancaster University, Lancaster, UK. http://www.Maths.Lancs.ac.uk/~ribeiro/geoR.html.

Ripley, B., D., (1981). *Spatial Statistics,* New York: John Wiley & Son, 44-77.

Snow, J. (1954). *On the Mode of Communication of Cholera.* Churchill Livingstone, London, 2nd edition.

Stone, M. (1974). Cross-Validatory choice and Assement of Statistical Predictions. *Journal of the Royal Statistical Society,* B., **36**: 111-133.

Waston, G. S., (1972). Trend Surface Analysis and Spatial Correlation. *Geological Society of America,* (Special Paper), **146**: 39-46.

Watson, D., F. and Philip, G., M., (1985). A refinement of Inverse Distance Weighted Interpolation, *Geoprocessing,* **2**: 315-327.

Webster, R. and Oliver M., (2001). *Geostatistics for Environmental Scientists,* John Wiley.

محمدزاده محسن (۱۳۷۸)، تابع کوواریانس تعمیم‌یافته برای کریگیدن عام اندیشه آماری، شماره دوم، ۲۳—۱۸.

محمدزاده محسن، کاظم‌نژاد انوشیروان، فقیه‌زاده سقراط، واقعی یدا...، (۱۳۸۰)، تعداد لگ مناسب در مدلسازی تغییرنگار، ارائه شده به مجله علوم دانشگاه تربیت معلم.

شناسایی داده‌های دورافتادهٔ فضایی

محسن محمدزاده، علی محمدیان مصمم

P۱۳۰۲۶

گروه آمار، دانشگاه تربیت مدرس

چکیده: یکی از مسائل مهم در تجزیه و تحلیل داده‌های فضایی شناسایی داده‌های دور افتاده فضایی است. وقتی که بیش از یک داده دورافتاده در مجموعه داده‌ها باشد، اغلب روشهایی که تاکنون ارائه شده دچار مشکلاتی تحت عنوان درون آوری یا برون بری می‌شوند که منجر به شناسایی غلط یک داده معمولی بعنوان داده دورافتاده یا بالعکس می‌شوند. در این مقاله یک روش تحلیل اکتشافی شناسایی داده‌های دورافتادهٔ فضایی ارائه و به کمک یک الگوریتم جستجوی پیشرو، مشاهدات بر اساس سازگاریشان با یک مدل مشخص مرتب و داده‌های دورافتادهٔ فضایی شناسایی می‌شوند.

واژه‌های کلیدی: داده‌های فضایی، کریگینگ و داده‌های دورافتادهٔ فضایی.

۱ مقدمه

درحالتی که مشاهدات دارای نوعی وابستگی ناشی از موقعیت آنها در فضای مورد مطالعه باشند، داده‌های فضایی[۱] نامیده می‌شوند. هدف این مقاله ارائه روشی برای شناسایی داده‌های دورافتاده[۲] در مجموعه داده‌های فضایی است . معمولاً داده‌هایی که نزدیک هم قرار دارند، شبیه یکدیگرند و هر چقدر موقعیت جغرافیایی آنها از یکدیگر دور می‌شود از همبستگی و شباهت آنها کاسته می‌شود. در آمار فضایی، مشاهداتی که نسبت به مقادیر همسایگی خود، فرین[۳] باشند، داده دورافتاده فضایی نامیده می‌شود.

معمولاً برای تجزیه و تحلیل آماری داده‌ها ابتدا تحلیل اکتشافی[۴] برای شناخت خصوصیات آنها انجام می‌گیرد. هاسلت(۱۹۹۱)، کرسی(۱۹۹۳) و پاناتیر(۱۹۹۶) برای کشف داده‌های دورافتاده از ابزارهای گرافیکی ساده نظیر نمودارهای جعبه‌ای[۵] و ابرهای واریوگرام استفاده نموده‌اند. همچنین کریستنسن، جانسون و پیرسن (۱۹۹۲) و هاسلت و هایز(۱۹۹۸) نیز

[۱] Spatial Data
[۲] Outliers
[۳] Extreme
[۴] Exploratory Analysis
[۵] Pocket Plots

روش تشخیص نقاط دورافتاده را پس از برازش مدل به داده‌ها ارائه نمودند. در این مقاله داده‌های دورافتادهٔ فضایی براساس الگوریتم جستجوی پیشرو[۶]، که مشاهدات را بر اساس سازگاریشان با یک مدل مشخص مرتب می‌کند، شناسایی می‌شوند. برتری این روش نسبت به روشهای سنتی این است که بر اساس مدل پیشنهاد شده یک مرتب سازی برای داده‌ها فراهم می‌سازد و داده‌های دورافتادهٔ چندگانه را بدون تاثیر مسائل درون آوری[۷] و برون‌بری[۸]، کشف می‌نماید. چون شناسایی داده‌های دور افتاده بر اساس باقیمانده پیش بینی‌ها انجام می‌شود، روش پیش بینی کریگینگ در بخش ۲ شرح داده خواهد شد. در بخش ۳ داده‌های دورافتاده چندگانه معرفی و روشهای شناسایی و مشکلات درون آوری و برون بری آنها دربخش ۴ مورد بررسی قرار می‌گیرد و نهایتاً بحث و نتیجه گیری در بخش ۵ ارائه خواهد شد.

۲ کریگینگ

معمولاً در آمار فضایی، میدان تصادفی[۱]

$$\{Z(s) : s \in D \subset \Re^d\} \qquad (۱)$$

برای مدل سازی داده‌های فضایی در نظر گرفته می‌شود، که درآن D مجموعهٔ اندیس‌گذار، $\Re^d$ فضای اقلیدسی $d \geq ۱$ بعدی و برای هر $s \in D$، $Z(s)$ یک متغیر تصادفی است . برای تخمین مقادیر میدان تصادفی (۱) روش کریگینگ به عنوان بهترین پیش‌بینی‌کنندهٔ خطی نااریب[۲] مورد استفاده قرار می‌گیرد . کریگینگ معمولی[۳] روشی برای پیش‌بینی کمیت‌های غیرقابل مشاهده از میدان تصادفی(۱) بفرم

$$Z(s) = \mu + \delta(s) \qquad s \in D \qquad \mu \in \Re \qquad (۲)$$

است، که درآن μ ثابت نامعلوم و $\{\delta(s) : s \in D\}$ یک میدان تصادفی ایستای ذاتی با واریوگرام معلوم

$$۲\gamma(s-t) = Var(Z(s) - Z(s+h)); \qquad s \in D$$

---

[۶] Forward Search
[۷] Masking
[۸] Swamping
[۱] Random Field
[۲] Best Linear Unbiased Estimator
[۳] Ordinary Kriging

و میانگین صفر باشد. برای مشاهدات داده شده در مکانهای فضایی $S \equiv \{s_1, \ldots, s_n\}$ بهترین پیش‌بینی کنندهٔ خطی نا اریب در موقعیت $s_0 \in D$، بصورت

$$\hat{Z}(s_0) = \underline{\lambda}' Z$$

محاسبه می‌شود، که در آن

$$\underline{\lambda}' = (\underline{\gamma} + \underline{1}\frac{\underline{1} - \underline{1}'\Gamma^{-1}\underline{\gamma}}{\underline{1}'\Gamma^{-1}\underline{1}})'\Gamma^{-1} \qquad (۳)$$

بردار ضرایب، $\underline{\gamma}' = [\gamma(s_0 - s_1), \ldots, \gamma(s_0 - s_n)]$ بردار مشاهدات، $Z' = [Z(s_1), \ldots, Z(s_n)]$ بردار ستونی واحد، $\underline{1}$ و Γ یک ماتریس $n \times n$ بصورت زیر می باشد.

$$\Gamma = \begin{pmatrix} \gamma(s_1 - s_1) & \gamma(s_1 - s_2) & \ldots & \gamma(s_1 - s_n) \\ \gamma(s_2 - s_1) & \gamma(s_2 - s_2) & \ldots & \gamma(s_2 - s_n) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(s_n - s_1) & \gamma(s_n - s_2) & \ldots & \gamma(s_n - s_n) \end{pmatrix}$$

مقدار $\hat{Z}(s_0)$ از مینیمم کردن میانگین مجذور خطای پیش‌بینی به روش لاگرانژ در میان تمام توابع خطی $\lambda' Z$ بدست می آید، بطوریکه برای بردار n بعدی λ، $\mathrm{E}[\lambda' Z] = \mathrm{E}[Z(s_0)]$ باشد. در اینصورت واریانس کریگینگ در موقعیت s_0 نیز بصورت زیر خواهد بود.

$$\sigma^2(s_0) \equiv E[Z(s_0) - \hat{Z}(s_0)]^2 = \underline{\gamma}'\Gamma^{-1}\underline{\gamma} - \frac{(\underline{1}'\Gamma^{-1}\underline{\gamma} - \underline{1})^2}{\underline{1}'\Gamma^{-1}\underline{1}}$$

۳ داده دورافتادهٔ چندگانه

در مقالات روشهای متفاوتی برای شناسایی داده‌های دور افتاده فضایی ارائه گردیده‌اند. کرسی (۱۹۹۳) با قرار دادن $S_{-i} \equiv \{s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n\}$، $\hat{Z}_i(S_{-i})$ بعنوان پیش‌بینی‌کنندهٔ کریگینگ برای $Z(s_i)$ براساس مشاهدات در موقعیتهای S_{-i} و واریانس کریگینگ $\sigma_i^2(S_{-i})$، باقیمانده‌های استاندارد پیش‌بینی را بصورت

$$\tilde{e}_i(S_{-i}) = \frac{Z(s_i) - \hat{Z}_i(S_{-i})}{\sigma_i(S_{-i})} \qquad i = 1, \ldots, n. \qquad (۴)$$

تعریف نمود و استفاده از نمودارهای تشریحی باقیمانده‌های(۴) را برای کشف داده‌های دورافتادهٔ فضایی پیشنهاد کرد. کریستنسن، جانسون و پیرسون (۱۹۹۲) نیز با تجزیه میدان تصادفی بفرم $Z(s) = \mu + \delta(s) + \varepsilon(s) \quad s \in D,$، که در آن $\varepsilon(s)$ و $\delta(s)$ مستقل هستند،

روشهای تشخیص را بر اساس باقیمانده‌های (۴) مورد بررسی قرار دادند. روشهای فوق عمدتاً برای مواردی قابل استفاده هستند که تنها یک دادهٔ دورافتاده در داده‌ها وجود داشته باشد. در غیر اینصورت ممکن است یک داده دورافتاده موجب پنهان شدن داده‌های دورافتادهٔ دیگر شود، که به آن مشکل درون آوری گفته می‌شود. یا امکان دارد مشاهده معمولی به عنوان داده دورافتاده شناسایی شود، که مشکل برون‌بری نامیده می‌شود. برای نمایش این مشکلات یک میدان تصادفی به‌فرم $Z(s) = \mu + \delta(s) + \varepsilon(s) \quad s \in D$، با فرض گوسی بودن $\delta(\cdot)$ و $\varepsilon(\cdot)$ و با استفاده از شبه‌واریوگرام کروی [۱]

$$\gamma(h) = \begin{cases} ۰ & ||h|| = ۰, \\ \theta_۰ + \theta_۱[۱.۵||h||/\theta_۲ - ۰.۵(||h||/\theta_۲)^۳] & ۰ < ||h|| \le \theta_۲, \\ \theta_۰ + \theta_۱ & ||h|| \ge \theta_۲, \end{cases} \qquad (۵)$$

برای پارامترهای $\theta_۰ = ۲$، $\theta_۱ = ۴$، $\theta_۲ = ۸$، میانگین $\mu = ۱۰$ و واریانس $\mathrm{Var}(\varepsilon(s_i)) = ۰/۱$ تعداد $n = ۸۱$ داده در یک شبکهٔ منظم ۹ × ۹ شبیه‌سازی و در جدول ۱ نشان داده شده است. در موقعیتهای $s_۱ = (۱, ۱)$، $s_۲ = (۱, ۲)$ و $s_۳ = (۱, ۳)$ بترتیب مقادیر ۶ ، ۴ و ۵ اضافه شده‌اند، بطوریکه داده‌های حاصل در این نقاط نسبت به عمدهٔ داده‌ها، دور افتاده محسوب می‌شوند.

جدول ۱: موقعیتهای فضایی در شبکهٔ منظم ۹ × ۹

	۱	۲	۳	۴	۵	۶	۷	۸	۹
۱	۱۱/۴	۱۳/۴	۱۲/۸	۱۳/۲	۹/۸	۱۱/۴	۱۲/۰	۱۰/۸	۱۰/۵
۲	۱۱/۸	۱۲/۷	۱۰/۸	۱۳/۸	۱۴/۴	۱۱/۳	۸/۷	۶/۸	۷/۲
۳	۱۱/۷	۱۲/۳	۱۵/۶	۱۲/۳	۱۲/۳	۱۰/۲	۸/۹	۷/۷	۱۰/۶
۴	۱۲/۸	۹/۹	۱۲/۸	۹/۷۰	۱۰/۴	۸/۴	۵/۵	۷/۲	۱۰/۰
۵	۱۱/۶	۱۱/۹	۱۳/۸	۹/۹	۸/۹	۸/۵	۷/۵	۱۰/۶	۹/۱
۶	۱۱/۵	۱۱/۴	۱۴/۵	۱۲/۶	۱۱/۶	۱۱/۶	۸/۳	۸/۷	۱۰/۴
۷	۱۱/۴	۱۳/۱	۱۱/۳	۱۲/۴	۱۱/۳	۷/۴	۱۱/۰	۱۱/۴	۹/۵
۸	۱۵/۶	۱۲/۳	۱۲/۲	۱۴/۰	۱۱/۶	۱۱/۴	۱۰/۸	۱۱/۵	۱۰/۲
۹	۱۲/۷	۱۲/۶	۱۳/۸	۱۵/۰	۱۲/۸	۱۱/۶	۱۲/۲	۱۰/۹	۱۰/۸

کوک لصاوف مود هشیر و $\tilde{e}_i(S_{-i})$ یاهدنامیقاب یاه‌بعج یاهرادومن

$$C_i(S_{-i}) = -(\hat{\delta}(S_{-i}) - \hat{\delta})' \hat{\Gamma}^{-۱} (\hat{\delta}(S_{-i}) - \hat{\delta}) \qquad i = ۱, \ldots, n \qquad (۶)$$

[۱] Spherical Semivariogram

شکل ۱: نمودار جعبه‌ای الف: باقیمانده‌های پیش‌بینی شده ب: ریشهٔ دوم فواصل کوک

که در آن، $\hat{\delta}$ بردار مقادیر پیش‌بینی بر اساس S و $\hat{\delta}(S_{-i})$ بردار مقادیر پیش‌بینی بر اساس S_{-i} و $\hat{\Gamma}$ برآورد ماتریس Γ است، درشکل ۱ نمایش داده شده است. همانطور که ملاحظه می‌شود، هیچ یک از مقادیر ناخالص موجود در موقعیتهای s_1، s_2 و s_3 بعنوان داده دور افتاده تشخیص داده نمی‌شود و مشکل درون آوری بروز می‌کند. تنها داده دور افتاده‌ای که در هر دو نمودار الف و ب شکل ۱ نشان داده می‌شود، مشاهده موقعیت s_{12} است که دارای مقدار نسبتاً کوچک $\tilde{e}_{12}(S_{-12})$ و مقدار بزرگ $C_{12}(S_{-12})$ می‌باشد. چنین نتایج نامطلوب را می‌توان در تکنیکهای تشریحی که توسط کرسی (۱۹۹۳) برای پیداکردن بی‌نظمی‌های فضایی طراحی شده‌اند نیز ملاحظه نمود. در این تکنیک مقدار قدرمطلق تفاضلهای استاندارد شده در هر ردیف و ستون شبکه بصورت

$$u \equiv \sqrt{m}(\bar{Z} - \tilde{Z})/(0.5\hat{\Psi}), \qquad (۷)$$

محاسبه می شوند، که در آن m تعداد موقعیتهای دارای مشاهده در یک سطر یا ستون مشخص می‌باشد و $\bar{Z}$، $\tilde{Z}$ و $\hat{\Psi}$ بترتیب میانگین، میانه و دامنه میان چارکی همان سطر یا ستون هستند. اگر مقدار $|u|$ نزدیک یا بیشتر از ۳ باشد، آنگاه آن ردیف یا ستون می‌تواند دارای داده دور افتاده باشد. در اینجا سطرهای ۷ و ۸ بخاطر داشتن مقادیر ۴/۳ و ۳/۸ و ستون ۱ بخاطر وجود مقدار ۵/۱ مشکوک داشتن داده دور افتاده تشخیص داده می‌شوند، در حالیکه این سطرها دارای نقاط دور افتاده نیستند و تنها ستون ۱ دارای نقطهٔ دور افتاده است. با این روش ستونهای ۲ و ۳ و سطر ۱ که شامل نقاط دور افتاده هستند، انتخاب نشده‌اند. لذا تکنیکی که این‌گونه داده‌های دور افتاده را به خوبی شناسایی کند ضروری می‌باشد. بعلاوه در این روش اطلاعات زیادی از طریق مرتب‌کردن مشاهدات از بیشترین سازگاریشان با مدل فضایی مشخص تا آنهایی که سازگاری کمتری با آن دارند ارائه

جدول ۲: قدرمطلق تفاضلهای استاندارد شده u(میانگین منهای میانه)

شماره سطرها	۱	۲	۳	۴	۵	٦	۷	۸	۹
$\|u\|$	۱٫۱	۰٫٦	۱	۰٫۸	۰٫۵	۱٫۴	۴٫۳	۳٫۴	۰٫۵
شماره ستونها	۱	۲	۳	۴	۵	٦	۷	۸	۹
$\|u\|$	۵٫۱	۲٫۳	۰٫۴	۰٫۲	۰٫۴	۲	۱	۱٫۸	۲٫۱

می‌شود.

۴ روش جستجوی پیشرو

برای شناسایی نقاط دورافتادهٔ چندگانهٔ درمدلهای رگرسیونی با خطاهای مستقل و آنالیز چند متغیره روش‌هایی درمقالات هادی (۱۹۹۲)، هادی و سیمونوف (۱۹۹۳)، اتکینسن (۱۹۹۴) و ریانی و اتکینسن (۱۹۹۹) ارائه گردیده است. برای شناسایی داده‌های دورافتادهٔ چندگانهٔ درمدلهای پیش بینی فضایی از روش جستجوی پیشرو استفاده خواهد شد، که الگوریتم‌های آن با تعدیل روشهای فوق برای داده‌های فضایی طراحی می‌شوند. دراین روش سریولی و ریانی (۱۹۹۹) یک زیر مجموعهٔ مقدماتی از داده‌ها را که فاقد نقاط دور افتاده هستند اختیار نموده و بر اساس مراحل چهارگانه الگوریتم زیر آنرا به مجموعه‌ای بزرگتر توسعه داده‌اند، بگونه‌ای که نقاط دور افتاده خارج از این مجموعه قرار خواهند گرفت.

مرحلهٔ ۰ (برآورد ماتریس Γ) : در اغلب مسائل کاربردی تابع واریوگرام، که تعیین کنندهٔ ساختار وابستگی داده‌های فضایی است، نامعلوم می باشد. لذا بایستی این تابع بر اساس مشاهدات برآورد شود. کرسی و هاوکینز(۱۹۸۰) یک برآوردگر مقاوم برای واریوگرام بصورت

$$۲\tilde{\gamma}(h) = \frac{(|N(h)|^{-۱}\Sigma_{N(h)}\sqrt{|Z(s_i)-Z(s_j)|})^{۴}}{.۴۵۷ + .۴۹۴/|N(h)|} \qquad h \in \Re^{۲}$$

معرفی نمودند، که درآن $N(h) = \{(s_i, s_j); s_i - s_j = h \quad i,j = ۱,\ldots,n\}$ و $|N(h)|$ تعداد زوجها با تأخیر h و $\sum_{N(h)}$ مجموع روی چنین زوجهایی است. هاوکینز و کرسی(۱۹۸۴) همچنین $۲\hat{\gamma}^{*}(h) = \frac{[\bar{Z}(h)]^{۴}}{۰.۴۵۷}$ را معرفی کردند که درآن $\bar{Z}(h)$ میانه‌ریشهٔ دوم تفاضلهای $\sqrt{|Z(s_i)-Z(s_j)|}$ روی زوجهای با تأخیر h و ۰٫۴۵۷ ضریب تصحیح

اریبی است. کرسی (۱۹۹۳) نشان داد $2\gamma(h)$ یک تابع معین منفی[1] است. برای اینکه برآوردهای $2\gamma(h)$ درشرط معین منفی صدق کنند، مدل‌های پارامتری برای برازش $\tilde{\gamma}(h)$ استفاده می‌شوند و ماتریس Γ بصورت $\hat{\Gamma} = [\hat{\gamma}(s_i - s_j)]$ برآورد می‌گردد.

مرحلهٔ ۱ (انتخاب زیر مجموعهٔ مقدماتی): وقتی مدل مورد استفاده برای پیش‌بینی شامل p پارامتر باشد، این الگوریتم با انتخاب یک زیرمجموعه شامل p موقعیت فضایی که عاری از نقاط دورافتاده باشند، شروع می‌شود. برای شناسایی این موقعیتها از نمودارهای جعبه‌ای دو متغیره استفاده می‌شود. برای $p \leq n$ می‌توان p تایی‌های مجزای

$$S^{(p)}_{i_1,\ldots,i_p} \equiv \{s_{i_1}, \ldots, s_{i_p}\};\ 1 \leq i_1, \ldots, i_p \leq n \text{ و } i_j \neq i_{j'}$$

را که تعداد آنها $\binom{n}{p}$ می‌باشد، انتخاب کرد. فرض کنید $\ell' = [i_1, \ldots, i_p]$ و $\hat{Z}_i(S^{(p)}_\ell)$ یک پیش‌بینی کنندهٔ کریگینگ در محل s_i براساس مشاهدات در $S^{(p)}_\ell$ باشد، که در اینصورت باقیمانده‌های استاندارد $\tilde{e}_i(S^{(p)}_\ell)$ خواهد بود. اکنون زیر مجموعهٔ مقدماتی p تایی $S^{(p)}_*$ بگونه‌ای انتخاب می‌شود که در رابطه

$$\tilde{e}^2_{[med]}(S^{(p)}_*) = \min_\ell [\tilde{e}^2_{[med]}(S^{(p)}_\ell)] \qquad (8)$$

صدق کند، که در آن $\tilde{e}^2_{[\ell]}(S^{(p)}_\ell)$ ، نشان دهنده مجذور ℓ امین باقیماندهٔ مرتب شده در میان باقیمانده‌های $med = p + [\frac{n-p}{2}]$ است و $\tilde{e}^2_i(S^{(p)}_\ell), i = 1, \ldots, n$ می باشد، بطوری که $[\frac{n-p}{2}]$ نشان دهنده جزٔ صحیح $\frac{n-p}{2}$ است .

معیار (۸) که به منظور برازش یک پیش بینی کننده نیرومند به داده‌ها انتخاب شده است، شبیه روش حداقل مجذور میانه است که برای مدل‌های رگرسیونی با خطاهای مستقل استفاده می‌شود، جز اینکه در اینجا خطاها بر اساس $\sigma_i(S^{(p)}_\ell)$ استاندارد می‌شوند. اگر n یا p بزرگ باشند، این فرایند ممکن است محاسبات زیادی داشته‌باشد. اتکینسن (۱۹۹۴) نشان‌داد که جستجوی پیشرو را می‌توان با انتخاب تصادفی زیرمجموعهٔ مقدماتی انجام داد.

مرحلهٔ ۲ (پیشروی در جستجوی پیشرو): اگر بعد زیرمجموعهٔ $S^{(m)}_*$، $m \geq p$ باشد، جستجوی پیشرو برای بعد $m+1$ با انتخاب $m+1$ مکان فضایی بطوریکه مجذور باقیمانده‌ها حداقل باشد با مرتب‌کردن مجذور باقیمانده‌های $\tilde{e}^2_i(S^{(m)}_*), i = 1, \ldots, n$ انتخاب می‌شوند. دراکثر انتقالات از m به $m+1$، فقط یک موقعیت جدید به زیرمجموعه ملحق می‌شود. اما برای مدل (۲) ممکن است دو یا چند موقعیت به $S^{(m)}_*$ ملحق یا چند موقعیت از آن جدا شوند. هر چند طبق تجربه چنین رخدادهایی وقتی‌که الگوریتم از $S^{(p)}_*$

[1] Negative Definite

شروع می‌شود، غیرطبیعی هستند و تنها وقتی اتفاق خواهد افتاد که جستجو شامل یک موقعیت متعلق به مجموعهٔ داده دورافتادهٔ فضایی باشد. می‌توان جستجو را بوسیله مرتب کردن مجذور باقیمانده‌ها که بصورت زیر تعریف می‌شوند،

$$r_i^{۲}(S_*^{(m)}) = \begin{cases} e_i^{۲}(S_*^{(m)}) & \text{اگر} \quad s_i \in S_*^{(m)} \\ \tilde{e}_i^{۲}(S_*^{(m)}) & \text{اگر} \quad s_i \notin S_*^{(m)} \end{cases}$$

ادامه داد، که درآن $e_i(S_*^{(m)}) \equiv Z(s_i) - \hat{Z}_i(S_*^{(m)})$ است. در اینصورت با استفاده از $r_i^{۲}(S_*^{(m)})$ بجای $\tilde{e}_i^{۲}(S_*^{(m)})$ احتمال اینکه دو یا چند موقعیت به $S_*^{(m)}$ در اولین مراحل جستجوی پیشرو ملحق شوند تا حدی افزایش می‌یابد. بنابر این توصیه می‌شود این روش وقتی‌که زیرمجموعهٔ مقدماتی بطور تصادفی انتخاب می‌شود، به منظور افزایش شانس دفع داده دورافتاده از زیرمجموعهٔ مقدماتی اتخاذ شود.

مرحلهٔ ۳ (مرتب سازی داده‌های فضایی) : تا زمانیکه تمام مکانها وارد زیرمجموعه مقدماتی شوند، مرحلهٔ ۲ تکرار می‌شود. اگر در هر انتقال فقط یک مکان به $S_*^{(m)}$ وارد شود، الگوریتم یک مرتب سازی از داده‌ها را فراهم می‌کند. نقاط دورافتادهٔ فضایی و همچنین مشاهدات دیگری که بی‌نظمی‌هایی دارند را می توان توسط تشریح گرافیکی از تغییرات آماره‌های موجود در جستجوی پیشرو را کشف نمود. برای این منظور نمودار باقیمانده‌های $\tilde{e}_i^{۲}(S_*^{(m)})$ در مقابل m رسم می‌شوند.

$$\tilde{e}_m^{(m+۱)} = \sqrt{\tilde{e}_{[m+۱]}^{۲}(S_*^{(m)})}, \qquad \sigma_m^{(m+۱)} = \sigma_{[m+۱]}^{۲}(S_*^{(m)}) \qquad m = p+۱, \ldots, n-۱$$

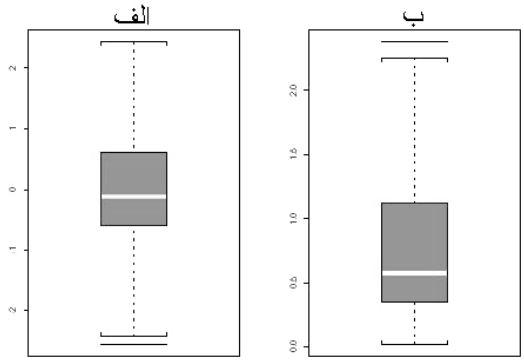
$$\tilde{e}_m^{(max)} = \sqrt{\tilde{e}_{[n]}^{۲}(S_*^{(m)})}, \qquad \sigma_m^{(max)} = \sigma_{[n]}^{۲}(S_*^{(m)}) \qquad m = p+۱, \ldots, n-۱$$

اگر یک یا چند مشاهدهٔ بی‌قاعده یا داده دورافتاده وجود داشته باشد، نمودارهای $\tilde{e}_m^{(m+۱)}$ و $\sigma_m^{(m+۱)}$ بایستی یک قله در مرحلهٔ قبلی نسبت به اولین داده دورافتاده نشان دهند.

در رابطه با یک طبقه از داده دورافتاده، وقتی اولین داده دورافتاده به $S_*^{(m)}$ ملحق می‌شود منحنی‌های $\tilde{e}_m^{(max)}$ و $\sigma_m^{(max)}$ بایستی یک شیب نزولی برای سهم اثر درون‌آوری داشته باشند. شکل ۲ نمودار باقیمانده‌های آماره‌های فوق را نشان می‌دهد. همانطور که ملاحظه می‌شود با نمایش توأم این نمودارها می‌توان اثر هر مشاهده را برحسب ویژگیهای داده‌های اصلی تعبیرکرد.

۵ نتیجه‌گیری

برای تشخیص داده‌های دورافتاده وقتی داده‌ها تنها شامل یک داده دورافتاده باشند روشهای کلاسیک قابل استفاده هستند. اما چنانچه بیش از یک داده دورافتاده در مجموعه داده‌ها



شکل ۲: ، منحنی‌های الف: $\tilde{e}_m^{(m+1)}$، ب: $\tilde{e}_m^{(max)}$، ج: $\sigma_m^{(m+1)}$ د: $\sigma_m^{(max)}$.

باشد، توان این روشها شدیداً کاهش می‌یابد. بعلاوه همانطورکه نشان داده شد، این روشها در تشخیص داده‌های دورافتاده همواره موفق نیستند، چون با مشکلات درون آوری و برون‌بری مواجه می‌شوند. روش ارائه شده بر اساس الگوریتم جستجوی پیشرو داده‌های دورافتاده چندگانه را بدون اینکه تحت تأثیر اینگونه مسائل قرار گیرند شناسایی می‌کند. بعلاوه می‌تواند برای کشف غیر ایستایی و دیگر بی‌نظمی های موضعی نیز بکار رود. چون روش ارائه شده یک مرتب سازی برای داده‌ها فراهم می‌سازد، حتی اگر داده‌ها شامل نقاط دورافتاده نباشند، این روش می‌تواند اطلاعات اکتشافی مفیدی را فراهم نماید.

مراجع

Atkinson, A. C. (1986), *Masking Unmasked* ", Biometrika,**73**, 533-541.

Atkinson, A. C. (1994), *Fast Very Roubust Method for the Detection of Multivariate Outliers*, Journal of the American Statistcal Association,**89**, 1329-1339.

Christensen, R., Johnson, W., and Pearson, L. M. (1992), *Prediction Diagnostics for Spatial Linear Models*, Biometrika,**79**, 583-591.

Cerioli, A. (1999), *The Ordering of Spatial Data and Detection of Multiple Outliers*, Journal of Computational and Graphical Statistics,**8**, 239-258.

Cressie, N. (1993), *Statistics for Spatial Data*, New York: Wiley.

Cressie, N. and Hawkins, D. M. (1980), *Roust Estimation of the Variogram:I,* Mathematical Geology,**12**, 115-125.

Hadi, A. S. (1992), *Identifying Multiple Outliers in Multivariate Data*, Journal of the Royal Statistical Society, Ser. B,**54**, 761-771.

Hadi, A. S., and Simonoff, J. S. (1993), *Procedures for the Identification of Multiple Outliers in Linear Models*, Journal of the American Statistical Association,**88**, 1264-1272.

Haslett, J., Bradley, R., Craig, P., Unwin, A. and Wills, G. (1991), *Dynamic Graphics for Exploring Spatial Data With Application to Locating Global and Local Anomalies*, The American Statistician,**45**, 234-242.

Haslett, J. and Hayes, K. (1998), *Residuals for the Linear Model whit Covariance Structure*, Journal of the Royal Statistical Society, Ser. B,**60**, 201-215.

Hawkins, D. M. (1993), *The Accuracy of Elemental Sets Approximations for Regression*, Journal of the American Statistical Association,**88**, 580-589.

Riani and Atkinson, A. C. (1999), *A Unified Approach to Multivariate Transformations and Multiple Outliers*, Submitted for Publication.

سرعت همگرایی الگوریتم EM و تسریع آن

سیّد محسن میرحسینی [۱]، دکتر عین‌اله پاشا [۲]

P۱۱۲۲۳

[۱] دانشکده ریاضی، دانشگاه یزد
[۲] دانشگاه تربیت معلم تهران

چکیده: یکی از روشهای برآورد پارامتر بعد از مشاهده،روش ماکزیمم درستنمایی است؛ اگر بردار مشاهدات غیرکامل باشند که فقدان قسمتی از داده ها ممکن است از مکانیزم سانسورکردن یا از مقادیر گمشده و... باشد، می‌توان از الگوریتم EM استفاده نمود. دراین مقاله، ابتدا به بررسی اجمالی الگوریتم EM و بیان برخی ویژگی ها و انتقادهای وارد بر آن می‌پردازیم. سپس مطالبی دربارهٔ همگرایی الگوریتم EM بیان می‌کنیم. با توجه به اینکه همگرایی الگوریتم EM به آرامی صورت می‌گیرد، با روش ایتکن،همگرایی آن را تسریع می‌بخشیم و در ادامه به ارائه چند مثال در این زمینه می‌پردازیم.

واژه‌های کلیدی: تابع درستنمایی ، الگوریتم EM،همگرایی الگوریتم EM[۱]، تسریع همگرایی ایتکن[۲]

۱ بخش اول: فرمول‌بندی الگوریتم EM

فرض کنید Y یک بردار تصادفی P بعدی با تابع چگالی احتمال $g(y;\Theta)$ $(p.d.f)$ باشد که $\Theta = (\theta_1, \theta_2, \cdots, \theta_d)^T$ بردار پارامترهای نامعلوم در تابع چگالی احتمال Y است. Tنشان دهندهٔ ترانسپوزهٔ بردار یا ماتریس است. فضای پارامتر را با Ω نمایش میدهیم. بردار Θ را می‌توان به روش ماکزیمم درستنمایی برآورد کرد. تابع درستنمایی برای Θ که از مقادیر مشاهده شده y تشکیل شده است برابر است با:

$$L(\Theta) = g(y;\Theta)$$

در بعضی مسائل،بردار yشامل داده های غیرکامل است که فقدان قسمتی از داده ها ممکن است از مقادیر گمشده و یا از مکانیزم سانسور کردن و... ناشی شده باشد، بدست

[۱] convergence of EM algorithm
[۲] Aitken's acceleration method

آوردن برآورد ماکزیمم درستنمایی (MLE)،کاری بس دشوار است. الگوریتم EM، یکی از ابزاری است که در این زمینه کارآیی وسیعی دارد. $g_c(x;\Theta)$ تابع چگالی احتمال بردار تصادفی X که نشان دهندهٔ بردار داده های کامل x است. بنابراین لگاریتم درستنمایی داده های کامل را با

$$\log L_c(\Theta) = \log g_c(x;\Theta)$$

نمایش می دهیم.
الگوریتم EM یک روش باتکرار برای برآورد پارامتر در داده‌های غیرکامل با استفاده از روش ماکزیمم درستنمایی است. باتوجه به اینکه $\log L(\Theta)$ شامل داده‌های غیرکاملی است، از تابع درستنمایی $\log L_c(\Theta)$استفاده می‌کنیم. چون بردار داده‌های کامل x شامل تمام مشاهدات نمی‌باشد، بنابراین به‌جای $\log L_c(\Theta)$، امید ریاضی شرطی آن به شرط y را جایگزین می‌کنیم. به عبارت دیگر فرض کنید $\Theta^{(k)}$، مقداری از Θ بعداز k امین تکرار باشد. در گام $(k+1)$ام گامهای Eو M به صورت زیرند:

۱) گام E (گام امید ریاضی): محاسبه $Q(\Theta;\Theta^{(k)})$ که

$$Q(\Theta;\Theta^{(k)}) = E_{\Theta^{(k)}}\{\log L_c(\Theta)|y\}$$

۲) گام M(گام ماکزیمم سازی): انتخاب $\Theta^{(k+1)}$ برای هر مقداری از $\Theta \in \Omega$ که $Q(\Theta;\Theta^{(k)})$ رانسبت به Θ ماکزیمم کند. به عبارت دیگر:

$$Q(\Theta^{(k+1)};\Theta^{(k)}) \geq Q(\Theta;\Theta^{(k)}) \quad \forall\Theta \in \Omega.$$

گامهای Eو M مکرراً تکرار میشوند تادنباله $L(\Theta^{(k)})$ به یک مقداری همگرا شود.

۲ بخش دّوم:همگرایی یک دنباله EM به یک مقدار مانا

در سال ۱۹۷۷، $Rubin, Dempster$ و $Laird$ نشان دادند که تابع درستنمایی داده‌های غیرکامل($L(\Theta)$) بعد از یک تکرار، کاهشی نیست[۶], [۱]، بنابراین دنباله ازبالا کراندار مقادیر درستنمایی $\{L(\Theta^{(k)}\}$، به طور یکنواخت به تعدادی مقدار L^* همگرا می‌شود که در اغلب موارد L^* یک مقدار مانا است که Θ^* در رابطه $\partial L(\Theta)/\partial\Theta = 0$ یا به طور

معادل $\partial \log L(\Theta)/\partial\Theta = 0$ صدق می‌کند.در برخی دیگر از مواقع ممکن است L^* یک ماکزیمم نسبی باشد. در مواقع نادری ممکن است Θ^* یک ماکزیمم‌کننده مطلق یا نسبی نباشد بلکه به عنوان مثال، یک نقطه زینی باشد که به نقطه شروع $\Theta^{(0)}$ بستگی دارد.

۳ بخش سّوم:نگاشت EM

الگوریتم EM رامی‌توان باتعریف نگاشتی از فضای پارامتر $\Theta(\Omega)$ به خودش تعریف نمود:

$$\Theta^{(k+1)} = M(\Theta^{(k)}) \qquad (k = 0, 1.2, \cdots)$$

اگر $\Theta^{(k)}$ به برخی مقادیر Θ^* همگرا شود و $M(\Theta)$ پیوسته باشد، آنگاه Θ^* باید در $\Theta^* = M(\Theta^*)$ صدق کند. بنابراین Θ^* نقطه ثابت شده نگاشت M است. با بسط تیلور $\Theta^{(k+1)} = M(\Theta^{(k)})$ در همسایگی Θ^* داریم

$$\Theta^{(k+1)} - \Theta^* \approx J(\Theta^*)(\Theta^{(k)} - \Theta^*) \qquad (1)$$

که $J(\Theta)$ ماتریس ژاکوبی $d \times d$ برای

$$M(\Theta) = (M_1(\Theta), M_2(\Theta), \cdots, M_d(\Theta))^T$$

است که (i,j)امین درایهٔ آن برابر است با $J_{ij}(\Theta) = \frac{\partial M_i(\Theta)}{\partial \Theta_j}$. بنابراین دریک همسایگی Θ^*، الگوریتم EM اساساً یک تکرار خطی با ماتریس نرخ همگرایی $J(\Theta^*)$ است، بنابراین $J(\Theta^*)$ نوعاً ناصفر است. از این رو فرض می‌کنیم که $\Theta^{(k)}$ یک دنباله EM است به طوری که برای $\Theta = \Theta^{(k+1)}$ داریم

$$\frac{\partial Q(\Theta, \Theta^{(k)})}{\partial \Theta} = 0$$

در سال ۱۹۷۷، $Dempster$ و همکارانش نشان دادند که اگر $\Theta^{(k)}$ همگرا به نقطه Θ^* باشد آنگاه

$$J(\Theta^*) = I - \mathcal{I}_c^{-1}(\Theta^*; y) I(\Theta^*; y) \qquad (2)$$

که

$$I(\Theta; y) = \frac{-\partial^2 \log L(\Theta)}{\partial\Theta \partial\Theta^T}$$

و

$$\mathcal{I}_c(\Theta; y) = E_\Theta(I_c(\Theta; X)|y)$$

و

$$I_c(\Theta; x) = \frac{-\partial^2 \log L_c(\Theta)}{\partial\Theta\partial\Theta^T}.$$

همچنین I یک ماتریس $d \times d$ است.

همگرایی الگوریتم EM به یک مقدار ماکزیمم نسبی برای $L(\Theta)$ وقتی رخ می‌دهد که $J(\Theta^*)$ دارای مقادیر ویژه بزرگتر از یک باشند.یک مقدار ویژه یک برای $J(\Theta^*)$ در همسایگی Θ^* نقطه مرزی $L(\Theta^*)$را نشان می‌دهد. در کل نشان داده می‌شود [۳]، $\partial^2 \log L(\Theta^*)/\partial\Theta\partial\Theta^T$ نیمه معین منفی است اگر معین منفی نباشد که دراین حالت مقادیر ویژه $J(\Theta^*)$، به ترتیب همگی در فواصل $[0, 1)$ یا $(0, 1]$ می‌افتند.

۴ بخش چهارم: برخی ویژگیها و معایب الگوریتم EM

تعدادی از ویژگیها ومحاسن الگوریتم EM به صورت زیر است:

۱) با هر تکرار افزایشی درستنمایی EM، به طور عددی پایدار است.
۲) تحت شرایط کلی مناسبی، حتماً به یک مقداری همگراست.
۳) به آسانی قابل استفاده است،زیرا روی محاسبات داده‌های کامل تکیه می‌کند.
۴) عموماً برای برنامه نویسی آسان است؛ چراکه هیچ نوع ارزیابی تابع درستنمایی و مشتقاتش در آن دخیل نیست.
۵) به فضای ذخیره سازی کمی نیازدارد و عموماً می‌توان بایک کامپیوتر کوچک هم انجام داد.
۶) هزینه تکرار عموماً پایین است به طوری که می‌توان با تعداد تکرارهای بیشتر مورد نیاز، الگوریتم EM را در مقایسه با دیگر روشهای مشابه توازن بخشید.
۷) کار آنالیزی مورد نیاز در این الگوریتم، خیلی ساده تر از روشهای مشابه است
۸) بامشاهده افزایش یکنوا دردرستنمایی روی تکرارها، به سادگی می‌توان همگرایی و خطاهای برنامه‌ریزی را برانداز نمود.
۹) الگوریتم EM می‌تواند برای داده‌های گمشده، سانسور شده و... به کاررود.

برخی معایب و انتقادهای وارد بر این الگوریتم از این قرارند:

۱) الگوریتم EM ممکن است حتی در مسائلی که به ظاهر بی‌دردسرند؛ به آرامی همگرا شود.
۲) برخلاف اکثر روشها، الگوریتم EM یک روش کلاسیک برای بدست آوردن یک تخمین از ماتریس برآورد پارامترها ارائه نداده است.

۳) الگوریتم EM نیز مانن سایر روشهای دیگر همگرایی به یک مقدار ماکزیمم مطلق راتضمین نمی‌کند و ممکن است به یک مقدار ماکزیمم نسبی یا نقطه زینی نیز همگرا شود.
۴) در بعضی مسائل، گام E ممکن است به طور آنالیزی به سادگی قابل کنترل نباشد، هرچند در چنین وضعیتهایی تأثیر پذیری از یک روش مونت کارلو وجود دارد.

۵ بخش پنجم: روش تسریع ایتکن چند متغیره

یکی از انتقادهایی که بر الگوریتم EM وارد است این است که همگرایی آن کاملاً کند صورت می‌گیرد. برخی روشها جهت تسریع آن وجود دارد که یکی از آنها که‌اغلب ازآن استفاده می‌شود؛ روش ایتکن است که در اینجا بیان می‌کنیم.
فرض کنید $\Theta^{(k)} \to \Theta^*$، وقتی که $k \to \infty$. بنابراین Θ^* را می‌توان به صورت زیر نوشت

$$\Theta^* = \Theta^{(k)} + \sum_{h=1}^{\infty} \left(\Theta^{(h+k)} - \Theta^{(h+k-1)}\right). \qquad (۳)$$

همچنین داریم:

$$\Theta^{(k+h)} - \Theta^{(h+k-1)} = M\left(\Theta^{k+h-1}\right) - M\left(\Theta^{(h+k-2)}\right). \qquad (۴)$$

$$\approx J(\Theta^{(h+k-2)})\left(\Theta^{(h+k-1)} - \Theta^{(h+k-2)}\right) \qquad (۵)$$

$$\approx J(\Theta^*)\left(\Theta^{(h+k-1)} - \Theta^{(h+k-2)}\right) \qquad (۶)$$

بنابراین برای kهای به اندازهٔ کافی بزرگ، $J(\Theta^{(h+k)}) = J(\Theta^*)$.تقریب (۴)و (۵) با بسط سری تیلور خطی $M\left(\Theta^{(k+h-1)}\right)$ حول نقطه $\Theta^{(h+k-2)}$ بدست آمده است. باتکرار عمل (۶) در (۳) داریم:

$$\begin{aligned}\Theta^* &= \Theta^{(k)} + \sum_{h=0}^{\infty}\{J(\Theta^*)\}^h(\Theta^{(k+1)} - \Theta^{(k)}) \\ &= \Theta^{(k)} + \{I - J(\Theta^*)\}^{-1}(\Theta^{(k+1)} - \Theta^{(k)}) \qquad (۷)\end{aligned}$$

بنابراین سری توانی $\sum_{h=0}^{\infty}\{J(\Theta^*)\}^h$ به $\{I - J(\Theta^*)\}$ همگرا می‌شود اگر همه مقادیر ویژه $J(\Theta^*)$بین صفر و یک باشد.
حال می‌خواهیم با استفاده از روش ایتکن، سرعت همگرایی دنباله $\{\Theta_A^{(k)}\}$ را تسریع بخشیم که $\Theta_A^{(k)}$ به صورت زیر تعریف می‌شود

$$\Theta_A^{(k+1)} = \Theta_A^{(k)} + \{I - J(\Theta_A^{(k)})\}^{-1}\left(\Theta_{EMA}^{(k+1)} - \Theta_A^{(k)}\right) \qquad (۸)$$

که $\Theta_{EMA}^{(k+1)}$ با استفاده از $\Theta_A^{(k)}$ در تکرار EM بدست آمده است. این روش را تا زمانی که در تکرار $(k+1)$ام ، به تولید اولین مقدار $\Theta_{EMA}^{(k+1)}$ در فرآیند EM با استفاده از $\Theta^{(k)}$ برسیم، ادامه می‌دهیم.

این روش در سال ۱۹۸۲ توسط لوئیس $(Louis)$ برای تسریع بخشیدن به همگرایی الگوریتم EM پیشنهاد شد[۲].

لوئیس در سال ۱۹۸۲، برای برآورد$J(\Theta_A^{(k)})$ موفق به یافتن ارتباط بین (۲) و (۸)شد که داریم

$$\Theta_A^{(k+1)} = \Theta_A^{(k)} + I^{-1}(\Theta_A^{(k)}; y)\mathcal{I}_c(\Theta_A^{(k)}; y)\left(\Theta_{EMA}^{(k+1)} - \Theta_A^{(k)}\right). \qquad (۹)$$

ارتباط (۲) برای مقادیر Θ به غیر از Θ^* به عنوان تقریب نسبتاً مفیدی برای $MLE\Theta^*$ است و همچنین نباید برخی تکرارهای EM به طور عادت انجام شود.استفاده از (۹) به عنوان تقریبی معادل استفاده از الگوریتم نیوتن رافسون برای پیداکردن صفر آماره نمرهٔ (داده‌های غیر کامل) $S(y;\Theta) = \partial \log L(\Theta)/\partial\Theta$ توسط $meilijson$ در سال ۱۹۸۹ مورد توجه قرار گرفت.

می‌توان نشان داد[۳] که

$$S(y;\Theta_A^{(k)}) \approx \mathcal{I}_c(\Theta_A^{(k)}; y)\left(\Theta_{EMA}^{(k+1)} - \Theta_A^{(k)}\right) \qquad (۱۰)$$

با جایگذاری این نتایج در (۹) داریم:

$$\Theta_A^{(k+1)} \approx \Theta_A^{(k)} + I^{-1}(\Theta_A^{(k)}; y)s(y;\Theta_A^{(k)}). \qquad (۱۱)$$

طرف راست (۱۱)، فرآیند بدست آمده در $(k+1)$امین تکرار روش نیوتن رافسون بکاررفته برای یافتن صفر $S(y;\Theta)$ است.

بنابراین استفاده از روش ایتکن به فرم استاندارد شده به طور اساسی معادل با روش نیوتن رافسون بکاررفته برای $S(y;\Theta)$ است که توسط لوئیس (۱۹۸۲) بیان شد.

در سال ۱۹۸۹، $Meilijson$ نشان داد که در حالتی که مشاهدات $w_n, \cdots, w_2, w_1$ مستقل و همتوزیع باشند، در روش نیوتن رافسون (۱۱)، می‌توان $I(\Theta; y)$ را با ماتریس اطلاع تجربی[۱] زیر تقریب کرد:

$$I_e(\Theta; y) = \sum_{j=1}^{n} s(w_j;\Theta)s(w_j;\Theta)^T - n^{-1}S(y;\Theta)S^T(y;\Theta) \qquad (۱۲)$$

که $s(w_j;\Theta)$ آماره نمره بر اساس مشاهده تکی w_j است.

[۱] empirical information matrix

٦ بخش ششم: مثالهایی از تسریع الگوریتم EM

۱.٦ مثال ۱– یک آزمایش چند جمله‌ای

یک مثال جهت نشان دادن چگونگی استفاده از روش تسریع ایتکن که توسط لوئیس در سال ۱۹۸۲ بیان شد؛ مثال معروف داده های یک توزیع چندجمله ای است که در مقاله DLR[۱]بیان شده است که بردار مشاهده شده $y = (y_۱, y_۲, y_۳, y_۴)^T$ از یک توزیع چندجمله‌ای به ترتیب با احتمالهای

$$\frac{۱}{۲} + \frac{۱}{۴}\theta, \frac{۱}{۴}(۱-\theta), \frac{۱}{۴}(۱-\theta), \frac{۱}{۴}\theta$$

گرفته شده است. پارامتر θ براساس بردار y برآورد می‌شود؛در صورتی که بردار $y = (۱۲۵, ۱۸, ۲۰, ۳۴)^T$ باشد آنگاه $\hat{\theta} = ۰/٦۲٦۸۲۱۵$ بدست می آید.حال اگر اولین خانه از خانه‌های چهارگانه اصلی را به دو خانه با احتمالهای $\frac{۱}{۴}$ و $\frac{۱}{۴}\theta$ تقسیم کنیم و $y_{۱۱}$ و $y_{۱۲}$ نشاندهندهٔ تقسیم $y_۱$ باشد به طوری که $y_۱ = y_{۱۲} + y_{۱۱}$. کاملاً واضح است که برآورد θ به سادگی قابل محاسبه نیست.پس از به کاربردن الگوریتم EM[٦]، در گامهای E و M داریم:

$$\theta^{(k+۱)} = \frac{y_{۱۲}^{(k)} + y_۴}{n - (y_{۱۱}^{(k)}}$$

که

$$y_{۱۱}^{(k)} = \frac{\frac{۱}{۲}y_۱}{(\frac{۱}{۲} + \frac{۱}{۴}\theta^{(k)})}$$

و

$$y_{۱۲}^{(k)} = y_۱ - y_{۱۱}^{(k)}.$$

با انتخاب $\theta^{(۰)}$ مناسب و با جایگذاری $y_{۱۱}^{(۰)}$ و $y_{۱۲}^{(۰)}$ (امید ریاضی شرطی $Y_{۱۱}$ و $Y_{۱۲}$ به شرط $y_۱$) در دنباله $\{\theta^{(k)}\}$ بعد از هشت مرحله تکرار به مقدار $\theta^{(۸)} = ۰/٦۲٦۸۲۱۴۸۴$ دست می‌یابیم.جدول نتایج الگوریتم EM برای این مثال در زیر آورده شده است: برای تسریع همگرایی آن، با بکاربردن فرمول (۹) بعد از دو مرحله تکرار EM داریم:

$$\theta_A^{(۲)} = \theta^{(۲)} = ۰/٦۲٦۳۳۸$$

$$I^{-۱}(\theta_A^{(۲)}; y)\mathcal{I}_c(\theta_A^{(۲)}; y) = \frac{۴۳۴/۷۹}{۳۷٦/۹۵}$$

$$= ۱/۱۵۳۴۴۲.$$

تکرار kام	$\theta^{(k)}$	$\log L(\theta^{(k)})$
۰	۰/۵۰۰۰۰۰۰۰۰	۶۴/۶۲۹۷۴
۱	۰/۶۰۸۲۴۷۴۲۳	۶۷/۳۲۰۱۷
۲	۰/۶۲۴۳۲۱۰۵۱	۶۷/۳۸۲۹۲
۳	۰/۶۲۶۴۸۸۸۷۹	۶۷/۳۸۴۰۸
۴	۰/۶۲۶۷۷۷۳۲۳	۶۷/۳۸۴۱۰
۵	۰/۶۲۶۸۱۵۶۳۲	۶۷/۳۸۴۱۰
۶	۰/۶۲۶۸۲۰۷۱۹	۶۷/۳۸۴۱۰
۷	۰/۶۲۶۸۲۱۳۹۵	۶۷/۳۸۴۱۰
۸	۰/۶۲۶۸۲۱۴۸۴	۶۷/۳۸۴۱۰

جدول ۱: نتایج الگوریتم EM برای داده های مدل چند جمله‌ای

$$\theta_A^{(۳)} = ۰/۶۲۶۳۳۸ + ۱/۱۵۳۴۴۲(\theta_{EMA}^{(۳)} - ۰/۶۲۶۳۳۸) = ۰/۶۲۶۸۲۱۶$$

و که $\theta_{EMA}^{(۳)} = ۰/۶۲۶۷۵۷$. می‌توان ملاحظه نمود که $\theta_A^{(۳)}$ از $\theta^{(۴)} = ۰/۶۲۶۸۱۲$ که چهارمین تکرار با مقادیر اصلی (بدون تسریع)الگوریتم EM است به $\hat{\theta} = ۰/۶۲۶۸۲۱۵$ نزدیکتر است.

۲.۶ مثال ۲. آمیزه هندسی

در سال ۱۹۸۹، $Meilijson$ مثالی دیگر در مورد آمیزه هندسی برای برآورد ML بردار پارامتر

$$\Theta = (\pi_۱, p_۱, p_۲)^T$$

بیان داشت. در آمیزه هندسی دو پارامتری

$$f(w;\Theta) = \sum_{i=۱}^{۲} \pi_i f(w_i; p_i)$$

که به ازای $i = ۱, ۲$ داریم:

$$f(w;p_i) = p_i(۱-p_i)^{w-۱} \qquad w = ۱, ۲, ۳, \cdots (۰ \le p_i \le ۱).$$

در حالت کلی می‌توان بردار داده کامل x را به دو قسمت داده‌های مشاهده شده(y)و داده‌های گمشده(z) تقسیم بندی نمود یعنی

$$x = \left(y^T, z^T\right)^T$$

که

$$z = (z_۱^T, z_۲^T, \cdots, z_n^T)^T.$$

$$y = (w_۱, w_۲, \cdots, w_n)^T$$

که $z_{ij} = (z_j)_i$ است که وقتی j امین مقدار مشاهده شده از i امین مؤلفه معلوم باشد مقداریک و وقتی گمشده باشد مقدار صفر را می‌گیرد$(i = ۱, ۲; j = ۱, ۲, \cdots, n)$.

لگاریتم درستنمایی داده های کامل به صورت زیر است:

$$\log L_c(\Theta) = \sum_{i=۱}^{۲} \sum_{j=۱}^{n} z_{ij} \{\log \pi_i + \log p_i + (w_j - ۱) \log(۱ - p_i)\} \quad (۱۳)$$

$$= \sum_{i=۱}^{۲} \{n_i(\log \pi_i + \log p_i) + \left(\sum_{j=۱}^{n} z_{ij} w_j - n_i \right) \log(۱ - p_i)\}, \quad (۱۴)$$

که

$$.n_i = \sum_{j=۱}^{n} z_{ij}$$

مطابق گام E در تکرار $(k+۱)$ام الگوریتم EM، تابع Q به صورت زیر است:

$$Q(\Theta; \Theta^{(k)}) = \sum_{i=۱}^{۲} \{n_i^{(k)} (\log \pi_i + \log p_i) + \left(\sum_{j=۱}^{n} z_{ij}^{(k)} w_j - n_i^{(k)} \right) \log(۱ - p_i)\} \quad (۱۵)$$

که

$$n_i^{(k)} = \sum_{j=۱}^{n} z_{ij}^{(k)}$$

و

$$z_{ij}^{(k)} = \frac{\pi_i^{(k)} p_i^{(k)} (۱ - p_i^{(k)})^{w_j - ۱}}{\sum_{h=۱}^{۲} \pi_h^{(k)} p_h^{(k)} (۱ - p_h^{(k)})^{w_j - ۱}}$$

امید ریاضی شرطی Z_{ij} به شرط yاست.با داشتن

$$\pi_i^{(k+۱)} = \frac{n_i^{(k)}}{n}$$

و

$$p_i(k+۱) = \left(\sum_{j=۱}^{n} z_{ij}^{(k)} w_j / n_i^{(k)} \right)^{-۱} \quad (i = ۱, ۲)$$

تکرار kام	$\pi_1^{(k)}$	$p_1^{(k)}$	$p_2^{(k)}$
۰	۰/۲	۰/۱	۰/۲
۵	۰/۱۴۶	۰/۲۴۸	۰/۷۰۴
۱۰	۰/۲۰۱	۰/۲۶۴	۰/۷۶۷
۱۰۰	۰/۴۰۴۶۵	۰/۳۵۳۹۵	۰/۹۰۲۲۰

جدول ۲: نتایج الگوریتم EM برای مدل آمیزه هندسی

برای محاسبه ماتریس اطلاع تجربی $I_e(\Theta;y)$ با استفاده از تعریف (۱۲) باید $s(w_j)$ را محاسبه کنیم که $s(w_j)$ آماره نمره داده غیرکامل بر اساس مشاهده تکی jام ($j = 1,2,\ldots,n$) است. داریم:

$$\begin{aligned} s(w_j;\Theta) &= \partial \log L_j(\Theta)/\partial\Theta \\ &= E_{\Theta^{(k)}}\left\{\partial \log L_{cj}(\Theta)/\partial\Theta|y\right\} \\ &= \partial Q_j(\Theta,\Theta^{(k)})/\partial\Theta \end{aligned}$$

که

$$Q_j(\Theta,\Theta^{(k)}) = E_{\Theta^{(k)}}\left\{\log L_{cj}(\Theta)|y\right\}.$$

بعد از محاسبات، ماتریس اطلاع تجربی $I_e(\Theta;y)$ به صورت زیر است:

$$s_1(w_j;\Theta) = (z_{1j}^{(k)} - \pi_1)/(\pi_1\pi_2)$$

$$s_{i+1}(w_j;\Theta) = z_{ij}^{(k)}(1 - w_j p_i)/\{p_i(1-p_i)\} \quad (i = 1,2).$$

$Meilijson$(۱۹۸۹) روشهای مختلفی را برای برازش مدل آمیزه هندسی به وسیله تابع درستنمایی با ۵۰۰ مشاهده انجام داد که $MLE\Theta$ بدست آمده به صورت زیر شد:

$$\hat{\Theta} = (0/40647, 0/35466, 0/90334)^T.$$

مقدار شروع Θ برای الگوریتم EM در تکرارهای پنجم، دهم و صدم در جدول ۲ آورده شده است.همچنین مقادیر تسریع الگوریتم EM نیز درجدول ۳ داده شده است.

تکرار kام	$\pi_1^{(k)}$	$p_1^{(k)}$	$p_2^{(k)}$
۰	۰/۱۴۶	۰/۲۴۸	۰/۷۰۴
۱	۰/۴۳	۰/۵۲	۰/۸۷
۲	۰/۶۰	۰/۴۹	۰/۹۸۶
۳	۰/۴۵	۰/۴۲	۰/۹۱
۴	۰/۴۲	۰/۳۷	۰/۹۱
۵	۰/۴۰۶	۰/۳۵۶	۰/۹۰۳
.	.	.	.
.	.	.	.
.	.	.	.
۸	۰/۴۰۶۴۷	۰/۳۵۴۶۶	۰/۹۰۳۳۴

جدول ۳: نتایج تسریع الگوریتم EM برای مدل آمیزه هندسی

مراجع

Dempster, A.P.,Laird, N.M., and Rubin, D.B. (1977). *Maximum Likelihood from Incomplete Data Via the EM Algorithm (with Discussion).* Journal of the Royal Statistical Society Series B 39, 1-38.

Louis, T.A. (1982). *Finding the obseved information matrix when using the EM algorithm* ., Journal of the Royal Statistical Society Series B 44, 226-233.

Mclachlan, G.J. , and Krishnan, T. (1995). *The EM Algorithm and Extensions.*, New York: Wiley. To be published.

Mclachlan, G.J. *On Aitken's Method and Other Approaches for Accelerating Convergence of the EM Algorithm.*

Meilijson, I. (1989).*A Fast Improvement to the EM Algorithm on Its Own Terms.* Journal of the Royal Statistical Society Series B 51, 127-138.

میرحسینی، سید محسن، الگوریتم EM و برآورد پارامترها در داده‌های غیرکامل، پایان نامه کارشناسی ارشد، ۱۳۷۸.

برآوردگر مینیماکس مجاز پارامتر مقیاس خانوادهٔ کای ـ دو تبدیل یافته در فضای پارامتری کراندار

نادر نعمت الهی [۱]، **محمد جعفری جوزانی** [۲]

$P۱۱۰۸۶$

[۱] دانشگاه علامه طباطبایی
[۲] دانشگاه شهید بهشتی

چکیده: روشهای معمول برآوردیابی از قبیل روش حداکثر درستنمایی در فضاهای پارامتری کراندار تحت تابع زیان مربع خطا، عموماً منجر به برآوردگرهای غیر مجاز می‌گردند. برآوردیابی به روش مینیماکس روش متداول دیگری است که در چنین فضاهایی مورد استفاده قرار می‌گیرد و معمولاً برآوردگرهای مجاز را به دست می دهد. دراین مقاله نتایج به دست آمده توسط ایدن (۱۹۹۵) در رابطه با برآوردگر پارامتر مقیاس توزیع گاما در فضای پارامتری کراندار تحت تابع زیان مربع خطای پایای مقیاس را به خانوادهٔ کای ـ دو تبدیل یافته تعمیم می دهیم. در این خانواده برآورد مینیماکس مجاز پارامتر مقیاس را در فضاهای پارامتری کراندار تحت تابع زیان مربع خطای پایای مقیاس به دست آورده و چند مثال ارایه می دهیم .

واژه‌های کلیدی: تابع زیان مربع خطای پایای مقیاس ، برآوردگر مینیماکس مجاز ، فضای پارامتری کراندار ، خانواده کای ـ دو تبدیل یافته .

۱ مقدمه

یکی از مسائلی که بیشتر مواقع در عمل با آن روبرو هستیم مسئله استنباط کردن و به ویژه برآوردیابی در فضای پارامتری کراندار است. روشهای معمول برآوردیابی از قبیل روش حداکثر درستنمایی در فضاهای پارامتری کراندار تحت تابع زیان مربع خطا عموماً منجر به برآوردگرهایی غیر مجاز می‌گردند. برآوردیابی به روش مینیماکس روش متداول دیگری است که در چنین فضاهایی مورد استفاده قرار می‌گیرد.

مسئله برآوردیابی در فضاهای پارامتری کراندار در ابتدا توسط برانک (۱۹۵۵) و ون ایدن (۱۹۵۷) مورد بررسی قرار گرفت . این نویسندگان توجه خود را روی وجود برآوردگرهای

حداکثر درستنمایی معطوف کردند. یکی از اولین کسانی که به مطالعه دربارهٔ برآوردگر مینیماکس مجاز در فضاهای پارامتری کراندار پرداخت کاتز(۱۹۶۱) بود. کاتز توانست یک برآوردگر مینیماکس مجاز برای میانگین توزیع نرمال در فضای پارامتری از پایین کراندار، تحت تابع زیان مربع خطا به‌دست آورد. بعدها بیکل، کسلا و استرادرمن (۱۹۸۱) مطالعاتی دربارهٔ برآوردیابی به روش مینیماکس در فضای پارامتری کراندار انجام دادند و بری (۱۹۹۳) برآوردگر مینیماکس مجاز برای توزیع نمایی با تکیه گاه $[\theta, \infty)$ که $\theta \in [a, b]$ یا $\theta \in [b, \infty)$ یا $(-\infty, a]$ را به دست آورد. تابع زیان مورد استفاده در بیشتر مطالعات انجام شده، تابع زیان مربع خطا بوده است و برآوردیابی تحت تابع زیان مربع خطای وزنی و در حالت خاص، مربع خطای پایای مقیاس مورد توجه کمتری قرار گرفته است.

تحت تابع زیان مربع خطای پایای مقیاس، ون ایدن (۱۹۹۵ و ۲۰۰۰) برآوردگر مینیماکس مجاز برای پارامتر مقیاس توزیع گاما و برآوردگر مینیماکس برای پارامتر مقیاس توزیع F در فضای پارامتری کراندار را به دست آورد.

در این مقاله نتایج به دست آمده توسط ون ایدن (۱۹۹۵) را به خانواده کای ـ دو تبدیل یافته که بوسیله رحمان [۱] و گوپتا [۲](۱۹۹۳) معرفی شده است، تعمیم می دهیم و در این خانواده تحت فضای پارامتری کراندار برآورد مینیماکس مجاز پارامتر مقیاس را تحت تابع زیان مربع خطای مقیاس به دست می آوریم. بدین منظور در بخش ۲ نتایج به دست آمده توسط ون ایدن (۱۹۹۵) را ارایه می دهیم. در بخش ۳ خانواده کای ـ دو تبدیل یافته را معرفی کرده و چند توزیع متعلق به این خانواده را مشخص می کنیم. در بخش ۴ برآوردگر مینیماکس مجاز پارامتر مقیاس خانواده کای ـ دو تبدیل یافته در فضای پارامتری کراندار را به دست آورده و چند مثال ارایه می دهیم.

۲ برآوردگر مینیماکس مجاز پارامتر مقیاس توزیع گاما در فضای پارامتری از پایین کراندار

در این بخش نتایج به دست آمده توسط ون ایدن(۱۹۹۵) در رابطه با برآوردگر مینیماکس مجاز پارامتر مقیاس توزیع گاما در فضای پارامتری از پایین کراندار، تحت تابع زیان مربع خطای پایای مقیاس را می آوریم. برای اثبات این نتایج می توان به ون ایدن (۱۹۹۵) مراجعه نمود.

---

[۱] Rahman
[۲] Gupta

فرض کنید X دارای توزیع گاما با تابع چگالی احتمال زیر باشد:

$$f_\theta(x) = \frac{۱}{\Gamma(\alpha)}.\frac{x^{\alpha-۱}}{\theta^\alpha}.e^{\frac{-x}{\theta}}, x > ۰ \quad (۱)$$

که در آن $\alpha > ۰$ معلوم، θ پارامتر مقیاس نامعلوم، $\theta \in [a, \infty)$ و $a > ۰$ معلوم است. تابع زیان مورد بررسی ، تابع زیان مربع خطای پایای مقیاس است. یعنی

$$L(\theta, \delta) = (\frac{\delta}{\theta} - ۱)^۲ \quad , \quad \delta \geq a, \theta \geq a \quad (۲)$$

همچنین فرض کنید $\hat{\theta}$ برآوردگری برای θ باشد که در شرط زیر صدق کند

$$P_\theta(\hat{\theta} \geq a) = ۱ \quad , \quad \forall \theta \geq a \qquad (۳)$$

برآوردگر $\hat{\theta}$ که در این بخش معرفی می‌شود به صورت زیر می باشد

$$\delta(x) = \lim_{n\to\infty} \delta_n(x) \qquad (۴)$$

که در آن $\delta_n(x)$، برای هر $n = ۱, ۲, \cdots$ یک برآوردگر بیز نسبت به توزیع پیشین زیر است

$$\begin{aligned} \pi_n(\theta) &= \frac{۱}{n\theta}\left(\frac{a}{\theta}\right)^{\frac{۱}{n}} \\ &= \frac{a^{\frac{۱}{n}}}{n\theta^{۱+\frac{۱}{n}}}, \forall \theta \geq a \end{aligned} \qquad (۵)$$

با به کار بردن (۱) و (۵) توزیع پسین θ به شرط $X = x$ متناسب است با

$$\pi(\theta|x) \propto \frac{۱}{\theta^{\alpha_n+۱}} e^{-\frac{x}{\theta}} \quad , \quad x > ۰ \quad , \quad \theta \geq a \quad (۶)$$

که در آن $\alpha_n = \alpha + \frac{۱}{n}$. ون ایدن (۱۹۹۵) توسط قضایای زیر برآورد مینیماکس مجاز پارامترθ را تحت تابع زیان (۲) به دست آورد.

قضیه ۱ برای تابع چگالی احتمال (۱) برآورد بیز θ ، برای هر $x > ۰$ ، نسبت به پیشین (۵) ، تحت تابع زیان (۲) عبارت است از

$$\delta_n(x) = \frac{x}{\alpha+۱+\frac{۱}{n}}\left(۱ + \frac{\left(\frac{x}{a}\right)^{\alpha+۱+\frac{۱}{n}} e^{-\frac{x}{a}}}{\int_۰^{\frac{x}{a}} t^{\alpha+۱+\frac{۱}{n}} e^{-t} dt}\right) \qquad (۷)$$

بعلاوه

$$r_n(\delta_n) < \infty \quad , \quad \forall n = ۱, ۲, \cdots$$

$$\lim_{n\to\infty} r_n(\delta_n) = \frac{۱}{\alpha+۱}$$

که در آن، r_n مخاطرهٔ بیز متناظر با δ_n است.

قضیه ۲ برای تابع چگالی احتمال (۱) با پارامتر مقیاس نامعلوم θ و پارامتر شکل معلوم $\alpha > ۰$ که در آن $\theta \geq a$ و $a > ۰$ معلوم است، برآورد

$$\delta(x) = \frac{x}{\alpha+۱}\left\{۱ + \frac{(\frac{x}{a})^{\alpha+۱}e^{-\frac{x}{a}}}{\int_۰^{\frac{x}{a}} t^{\alpha+۱}e^{-t}dt}\right\} \qquad (۸)$$

تحت تابع زیان (۲) ، یک برآورد مینیماکس مجاز برای θ است. بعلاوه مقدار مینیماکس برابر $\frac{۱}{\alpha+۱}$ می‌باشد که در نقطهٔ $\theta = a$ و یا هنگامی که θ به بینهایت میل کند به دست می آید.

۳ خانواده کای ـ دو تبدیل یافته و چند حالت خاص آن

در این بخش خانواده کا ی ـ دو تبدیل یافته که توسط رحمان و گوپتا (۱۹۹۳) معرفی شده است را مورد بررسی قرار می دهیم . فرض کنید $\underset{\sim}{X} = (X_۱, X_۲, ..., X_m)$ برداری تصادفی متعلق به خانوادهٔ نمایی یک پارامتری از توزیعها با تابع چگالی احتمال توأم زیر باشد

$$f(\underset{\sim}{x}, \eta) = exp\{a(\underset{\sim}{x})b(\eta) + c(\eta) + h(\underset{\sim}{x})\} \qquad (۹)$$

برای مثال توزیعهای پواسن ، دوجمله ای ، نرمال و گاما با اختیار $c(\eta), b(\eta), a(\underset{\sim}{x})$ و $h(\underset{\sim}{x})$ مناسب به خانوادهٔ فوق تعلق دارند. در قضیهٔ زیر نشان می دهیم تحت وجود بعضی شرایط $-۲a(\underset{\sim}{X})b(\eta)$ دارای توزیع گاما است .

قضیهٔ ۳: درخانوادهٔ نمایی یک پارامتری (۹) عبارت $-۲a(\underset{\sim}{X})b(\eta)$ دارای توزیع گاما با پارامترهای $\frac{k}{۲}$ و ۲ است اگر و تنها اگر

$$\frac{۲c'(\eta)b(\eta)}{b'(\eta)} = k \qquad (۱۰)$$

که در آن k مثبت و مستقل از η می باشد.

برهان : فرض کنید $k = \frac{۲c'(\eta)b(\eta)}{b'(\eta)}$ از آنجا که (۹) یک تابع چگالی احتمال است ، بنابراین

$$\int e^{a(\underset{\sim}{x})b(\eta)+c(\eta)+h(\underset{\sim}{x})}d\underset{\sim}{x} = ۱$$

و یا به طور معادل

$$\int e^{a(\underset{\sim}{x})b(\eta)+h(\underset{\sim}{x})}d\underset{\sim}{x} = e^{-c(\eta)} \qquad (۱۱)$$

از طرفی با استفاده از رابطهٔ (۱۰) داریم

$$c'(\eta) = \frac{k}{۲}\frac{b'(\eta)}{b(\eta)}$$

و باانتگرال گیری از طرفین رابطهٔ فوق نسبت به η داریم

$$c(\eta) = \frac{k}{۲}Ln|b(\eta)| + k_۱$$

که در آن $k_۱$ ثابت انتگرال گیری است .بنابراین می توان رابطهٔ (۱۱) را به صورت زیر بازنویسی کرد

$$\int e^{a(\underset{\sim}{x})b(\eta)+h(\underset{\sim}{x})}d\underset{\sim}{x} = e^{-\frac{k}{۲}Ln|b(\eta)|-k_۱}$$

قرار می دهیم $Y = -۲a(\underset{\sim}{X})b(\eta)$ تابع مشخصهٔ Y عبارت است از

$$\begin{aligned}\varphi_Y(t) &= E\left[e^{-۲ita(\underset{\sim}{X})b(\eta)}\right] \\ &= e^{c(\eta)}\int e^{a(\underset{\sim}{x})b(\eta)(۱-۲it)+h(\underset{\sim}{x})}d\underset{\sim}{x} \\ &= e^{\frac{k}{۲}ln|b(\eta)|-k_۱}e^{\frac{k}{۲}ln|b(\eta)|(۱-۲it)+k_۱} \\ &= (۱-۲it)^{-\frac{k}{۲}}\end{aligned}$$

که همان تابع مشخصهٔ توزیع گاما با پارامترهای $\frac{k}{۲}$ و ۲ است . از آنجا که تابع مشخصه به طور یکتا توزیع متغیر تصادفی را مشخص می کند لذا $-۲a(\underset{\sim}{X})b(\eta)$ دارای توزیع گاما با پارامترهای مذکور است .

برعکس فرض کنید $Y = -۲a(\underset{\sim}{X})b(\eta)$ دارای توزیع گاما با پارامترهای $\frac{k}{۲}$ و ۲ باشد. تابع چگالی احتمال Y عبارت است از

$$f_\eta(y) = \frac{e^{-\frac{y}{۲}}y^{\frac{k}{۲}-۱}}{\Gamma(\frac{k}{۲})۲^{\frac{k}{۲}}}$$

بنابراین تابع چگالی احتمال $a(\underset{\sim}{X})$ به صورت زیر است

$$f_\eta(a(\underset{\sim}{x})) = \frac{e^{a(\underset{\sim}{x})b(\eta)}\ \{-b(\eta)\}^{\frac{k}{2}}\ \{a(\underset{\sim}{x})\}^{\frac{k}{2}-1}}{\Gamma(\frac{k}{2})}$$
$$= e^{\{a(\underset{\sim}{x})b(\eta)+\frac{k}{2}Ln|b(\eta)|+(\frac{k}{2}-1)Ln\{a(\underset{\sim}{x})\}-Ln\{\Gamma(\frac{k}{2})\}\}}$$

که متعلق به خانوادهٔ نمایی یک پارامتری (۹) است و در آن

$$c(\eta) = \frac{k}{2}Ln|b(\eta)| + k_1 \qquad (۱۲)$$

لازم به ذکر این نکته است که اگر $a(\underset{\sim}{X}) > 0$ آنگاه می بایست $b(\eta)$ منفی باشد . در رابطهٔ بالا k_1 ثابت انتگرالگیری می باشد و بامشتق گیری از طرفین رابطهٔ (۱۲) داریم

$$\frac{2c'(\eta)b(\eta)}{b'(\eta)} = k$$

تعریف ۱: خانوادهٔ توزیعهای نمایی یک پارامتری با تابع چگالی احتمال (۹) که در شرط (۱۰) صدق می کند هنگامی که k یک عدد صحیح مثبت باشد را خانوادهٔ توزیعهای کا ی ـ دو تبدیل یافته گویند.

برای مثال توزیع توأم یک نمونه تصادفی m تایی از توزیع لاگ نرمال $LN(0,\sigma^2)$ عبارت است از

$$f_{\sigma^2}(\underset{\sim}{x}) = (2\pi)^{-\frac{m}{2}}(\sigma^2)^{-\frac{m}{2}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(lnx_i)^2}$$

که متعلق به خانواده کای ـ دو تبدیل یافته با

$$k = m \quad , \quad a(\underset{\sim}{x}) = \frac{1}{2}\sum_{i=1}^{m}(lnx_i)^2$$

$$b(\sigma) = -\frac{1}{\sigma^2}, c(\sigma) = -mln\sigma, -2a(\underset{\sim}{x})b(\sigma) = \frac{1}{\sigma^2}\sum_{i=1}^{m}(lnx_i)^2$$

می باشد. همچنین توزیع توأم یک نمونه تصادفی m تایی از توزیع وایبل $W(\alpha,\beta)$ با α معلوم عبارت است از

$$f_\beta(\underset{\sim}{x}) = \beta^m\alpha^m e^{-\beta\sum_{i=1}^{m}{x_i}^\alpha}(\Pi_{i=1}^{m}{x_i}^{\alpha-1})$$

که متعلق به خانواده کای ـ دو تبدیل یافته با

$$k = ۲m, a(\underset{\sim}{x}) = \sum_{i=۱}^{m} {x_i}^{\alpha}, b(\beta) = -\beta, c(\beta) = mln\beta, -۲a(\underset{\sim}{x})b(\beta) = ۲\beta \sum_{i=۱}^{m} {x_i}^{\alpha}$$

می باشد. در جدول (۱) بعضی از توزیعهای متعلق به خانواده کای ـ دو تبدیل یافته نشان داده شده است .

۴ برآوردگر مینیماکس مجاز پارامتر مقیاس در خانواده کای ـ دو تبدیل یافته

در این بخش برآوردگر مینیماکس مجاز پارامتر مقیاس θ که در آن $\theta \geq a$ و $a > ۰$ معلوم است در خانواده کای ـ دو تبدیل یافته تحت تابع زیان مربع خطای پایای مقیاس را به دست می آوریم .

در رابطه (۱۲) قرار می دهیم $\theta = -\frac{۱}{b(\eta)} > ۰$ در این صورت

$$\begin{aligned} e^{c(\eta)} &= \left[-\frac{۱}{b(\eta)} \right]^{-\frac{k}{۲}} e^{k_۱} \\ &= \theta^{-\frac{k}{۲}} e^{k_۱} \end{aligned}$$

بنابر این خانواده توزیعهای نمایی (۹) (کای ـ دو تبدیل یافته) را می توان به صورت زیر نوشت

$$\begin{aligned} e^{-\frac{a(\underset{\sim}{x})}{[-۱/b(\eta)]} + c(\eta) + h(\underset{\sim}{x})} &= e^{h(\underset{\sim}{x}) + k_۱} e^{c(\eta)} e^{-\frac{a(\underset{\sim}{x})}{[-۱/b(\eta)]}} \\ &= c(\underset{\sim}{x}, m) \theta^{-\frac{k}{۲}} e^{-\frac{a(\underset{\sim}{x})}{\theta}} \end{aligned}$$

که در آن $c(\underset{\sim}{x}, m) = e^{h(\underset{\sim}{x}) + k_۱}$ است. همچنین با توجه به اینکه

$$-۲a(\underset{\sim}{X})b(\eta) = \frac{۲a(\underset{\sim}{X})}{\theta} \sim \Gamma(\frac{k}{۲}, ۲)$$

بنابر این $a(\underset{\sim}{X}) \sim \Gamma(\frac{k}{۲}, \theta)$. در نتیجه خانواده کای ـ دو تبدیل یافته به صورت زیر تبدیل می شود

$$f_\theta(\underset{\sim}{x}) = c(\underset{\sim}{x}, m)\theta^{-\frac{k}{۲}} e^{-\frac{a(\underset{\sim}{x})}{\theta}} \quad , \quad k > ۰ \ , \ \theta > ۰ \ , \ a(\underset{\sim}{X}) \sim \Gamma(\frac{k}{۲}, \theta) \quad (۱۳)$$

در این بخش با در نظر گرفتن توزیع پیشین (۵) و تابع زیان به فرم (۲) و شرایط (۳) و (۴) روی برآوردگرها ، می خواهیم با استفاده از نتایج به دست آمده توسط ون ایدن (۱۹۹۵) ، برآوردگر مینیماکس مجاز پارامتر θ را در خانواده کای ـ دو تبدیل یافته (۱۳) به دست آوریم.

قضیه ۴ برای خانواده توزیعهای کای ـ دو تبدیل یافته به فرم (۱۳) با $\theta \geq a$ و $a > ۰$ معلوم، برآورد بیز $\theta = -\frac{۱}{b(\eta)}$ برای هر $x > ۰$، نسبت به‌توزیع پیشین (۵)، تحت تابع زیان (۲) عبارت است از

$$\delta_n(\underset{\sim}{x}) = \frac{a(\underset{\sim}{x})}{\frac{k}{۲}+۱+\frac{۱}{n}} \left(۱ + \frac{\left(\frac{a(\underset{\sim}{x})}{a}\right)^{\frac{k}{۲}+۱+\frac{۱}{n}} e^{-\frac{a(\underset{\sim}{x})}{a}}}{\int_۰^{\frac{a(\underset{\sim}{x})}{a}} t^{\frac{k}{۲}+۱+\frac{۱}{n}} e^{-t} dt} \right) \qquad (۱۴)$$

به علاوه

$$r_n(\delta_n) < \infty \quad , \quad \forall n = ۱, ۲, \cdots$$
$$\lim_{n\to\infty} r_n(\delta_n) = \frac{۲}{k+۲} \qquad (۱۵)$$

که در آن، r_n مخاطره‌بئیز متناظر با δ_n است.

اثبات : با استفاده از (۱۳) و (۵) ، تابع چگالی پسین θ به شرط $\underset{\sim}{x}$ متناسب است با

$$\pi(\theta|\underset{\sim}{x}) \propto \frac{۱}{\theta^{\alpha_n+۱}} e^{-\frac{a(\underset{\sim}{x})}{\theta}} \quad , \quad x > ۰ \ , \quad \theta \geq a$$

که در آن $\alpha_n = \frac{k}{۲} + \frac{۱}{n}$. این تابع چگالی پسین همانند چگالی پسین (۶) می باشد که در آن x و α به ترتیب به $a(\underset{\sim}{x})$ و $\frac{k}{۲}$ تبدیل شده‌اند . بنابراین برآوردگر بیز (۱۴) و خواص آن را می توان به موازات اثبات قضیه (۱) ارایه شده توسط ون ایدن (۱۹۹۵) به دست آورد.

از قضیه (۴) برای هر $x > ۰$ به دست می آوریم که

$$\delta(\underset{\sim}{x}) = \lim_{n\to\infty} \delta_n(\underset{\sim}{x}) = \frac{a(\underset{\sim}{x})}{\frac{k}{۲}+۱}\left\{ ۱ + \frac{(\frac{a(\underset{\sim}{x})}{a})^{\frac{k}{۲}+۱} e^{-\frac{a(\underset{\sim}{x})}{a}}}{\int_۰^{\frac{a(\underset{\sim}{x})}{a}} t^{\frac{k}{۲}+۱} e^{-t} dt} \right\} \qquad (۱۶)$$

مینیماکس و مجاز بودن برآورد (۱۶) در قضیه زیر آورده شده است . اثبات این قضیه شبیه اثبات قضیه (۲) می باشد که در ون ایدن (۱۹۹۵) آمده است.

قضیه ۵ : برای خانواده توزیعهای تبدیل یافته به فرم (۱۳) با $\theta \geq a$ و $a > ۰$ معلوم، برآورد مینیماکس مجاز $\theta = -\frac{۱}{b(\eta)}$ برای هر $a(\underset{\sim}{x}) > ۰$، تحت تابع زیان به فرم (۲) به صورت (۱۶) می باشد. به علاوه مقدار مینیماکس برابر $\frac{۲}{k+۲}$ است. این مقدار در $\theta = a$ یا هنگامی که θ به سمت بینهایت میل کند به دست می آید.

مثال ۱ : فرض کنید X یک متغیر تصادفی با توزیع $\Gamma(\alpha, \beta)$ باشد که در آن $\alpha > ۰$ مقداری معلوم است در این صورت

$$f_\beta(x) = \frac{۱}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-۱} e^{-\frac{x}{\beta}} \quad , \quad x > ۰$$

که متعلق به خانواده کای ـ دو تبدیل یافته با

$$k = ۲\alpha, a(\underset{\sim}{x}) = x, b(\beta) = -\frac{۱}{\beta}, c(\beta) = -\alpha ln\beta, -۲a(\underset{\sim}{x})b(\beta) = \frac{۲x}{\beta}, \theta = \beta$$

است . همچنین $۲\alpha = \frac{۲c'(\beta)b(\beta)}{b'(\beta)}$ بنابر این طبق قضیه (۵) برآورد مینیماکس مجاز پارامتر $\theta = \beta$ با شرط $\beta \geq a$ عبارت است از

$$\delta(x) = \frac{x}{\alpha+۱}\left\{ ۱ + \frac{(\frac{x}{a})^{\alpha+۱} e^{-\frac{x}{a}}}{\int_۰^{\frac{x}{a}} t^{\alpha+۱} e^{-t} dt} \right\}$$

که همان برآورد به دست آمده در قضیه ۲ می باشد که توسط ون ایدن (۱۹۹۵) ارایه گردیده است.

مثال ۲ : فرض کنید $X_1, X_2, ..., X_m$ یک نمونه تصادفی از $N(0, \sigma^2)$ باشند. در این صورت

$$f_{\sigma^2}(\underset{\sim}{x}) = (2\pi)^{-\frac{m}{2}}(\sigma^2)^{-\frac{m}{2}} e^{-\frac{\sum_{i=1}^{m} x_i^2}{2\sigma^2}}$$

که متعلق به خانواده کای ـ دو تبدیل یافته با

$$k = m, a(\underset{\sim}{x}) = \frac{1}{2}\sum_{i=1}^{m} {x_i}^2, b(\sigma) = -\frac{1}{\sigma^2},$$

$$c(\sigma) = -mln\sigma, -2a(\underset{\sim}{x})b(\sigma) = \frac{\sum_{i=1}^{m} {x_i}^2}{\sigma^2}, \theta = \sigma^2$$

است. همچنین $m = \frac{2c'(\sigma)b(\sigma)}{b'(\sigma)}$ بنابراین طبق قضیه (۵) برآورد مینیماکس مجاز پارامتر $\theta = \sigma^2$ با شرط $\sigma^2 \geq a$ عبارت است از

$$\delta(\underset{\sim}{x}) = \frac{\sum_{i=1}^{m} {x_i}^2}{m+2} \left\{ 1 + \frac{(\frac{\sum_{i=1}^{m} {x_i}^2}{2a})^{\frac{m}{2}+1} e^{-\frac{1}{2a}\sum_{i=1}^{m} {x_i}^2}}{\int_0^{-\frac{1}{2a}\sum_{i=1}^{m} {x_i}^2} t^{\frac{m}{2}+1} e^{-t} dt} \right\}$$

مثال ۳ : فرض کنید $X_1, X_2, ..., X_m$ یک نمونه تصادفی از $IG(\infty, \lambda)$ باشند.در این صورت

$$f_\lambda(\underset{\sim}{x}) = (\Pi_{i=1}^{m} 2\pi x_i^3)^{-\frac{1}{2}} \lambda^{\frac{m}{2}} e^{-\frac{\lambda}{2}\sum_{i=1}^{m} \frac{1}{x_i}}$$

که متعلق به خانواده کای ـ دو تبدیل یافته با

$$k = m, a(\underset{\sim}{x}) = \frac{1}{2}\sum_{i=1}^{m} \frac{1}{x_i}, b(\lambda) = -\lambda,$$

$$c(\lambda) = \frac{m}{2} ln\lambda, -2a(\underset{\sim}{x})b(\lambda) = \lambda \sum_{i=1}^{m} \frac{1}{x_i}, \theta = \frac{1}{\lambda}$$

است. همچنین $m = \frac{۲c'(\lambda)b(\lambda)}{b'(\lambda)}$ بنابر این طبق قضیه (۵) برآورد مینیماکس مجاز پارامتر $\theta = \frac{۱}{\lambda}$ با شرط $a \geq \frac{۱}{\lambda}$ عبارت است از

$$\delta(\underset{\sim}{x}) = \frac{\sum_{i=۱}^{m} \frac{۱}{x_i}}{m+۲} \left\{ ۱ + \frac{(\frac{\sum_{i=۱}^{m} \frac{۱}{x_i}}{۲a})^{\frac{m}{۲}+۱} e^{-\frac{۱}{۲a}\sum_{i=۱}^{m}\frac{۱}{x_i}}}{\int_{۰}^{-\frac{۱}{۲a}\sum_{i=۱}^{m}\frac{۱}{x_i}} t^{\frac{m}{۲}+۱} e^{-t} dt} \right\}$$

در جدول (۱) بعضی از توزیعهای متعلق به خانواده توزیعهای کای ـ دو تبدیل یافته به همراه $a(\underset{\sim}{x})$ و $\theta = -\frac{۱}{b(\eta)}$ و k آورده شده است. در این توزیعها برآوردگر (۱۶) برای θ مینیماکس و مجاز است .

مراجع

Berry J. C. (1993). *Minimax Estimation of a Restricted Exponential location Parameter.* Statist. Decisions, 11, 307-316.

Brunk, H. D. (1955). *Maximum Likelihood Estimates of Monotone Parameters.* Ann. Math. Statist., 26, 607-616.

Katz, M.W. (1961). *Admissible and Minimax Estimates of Parameter in Truncated Spaces.* Ann. Math. Statist., 32, 136-142.

Rahman, M.S and Gupta, R.P. (1993) *Family of Transformed Chi-square Distributions.* Commun. Statist. Theory Meth., 22 (1), 135-146.

van Eeden, C. (1957). *Maximum Likelihood Estimation of Partially or Completely Ordered Parameters.* Proc. Kon. Ned. Akad . V.Wet. 60A, 128-136 and 201-211.

van Eeden, C. (1995). *Minimax Estimation of a Lower Bounded Scale-parameter of a Gamma Distribution for Scale-invariant Squared-error Loss.* The Canadian Journal of Statistics, 23 (3), 245-256.

van Eeden, C. (2000). *Minimax Estimation of a Lower-bounded Scale-parameter of an F-distribution.* Statistics & Probability Letters ,46 , 283-286.

Name of distribution with p.d.f	$a(X)$	$b(\cdot)$	$c(\cdot)$	$-۲a(X)b(\cdot)$	k	θ
$Normal$ $\frac{e^{-\frac{x^۲}{۲\sigma^۲}}}{\sigma\sqrt{(۲\pi)}}$	$\frac{۱}{۲}\sum_{i=۱}^{m} X_i^۲$	$-\frac{۱}{\sigma^۲}$	$-m\ Ln\ \sigma$	$\frac{\sum_{i=۱}^{m} X_i^۲}{\sigma^۲}$	m	$\sigma^۲$
$Lognormal$ $\frac{e^{-\frac{(Ln\ x)^۲}{۲\sigma^۲}}}{\sigma\ x\sqrt{(۲\pi)}}$	$\frac{۱}{۲}\sum_{i=۱}^{m}(Ln\ X_i)^۲$	$-\frac{۱}{\sigma^۲}$	$-m\ Ln\ \sigma$	$\frac{\sum_{i=۱}^{m}(Ln\ X_i)^۲}{\sigma^۲}$	m	$\sigma^۲$
$Exponential$ $\frac{۱}{\lambda}\ e^{-\frac{x}{\lambda}}$	$\sum_{i=۱}^{m} X_i$	$-\frac{۱}{\lambda}$	$-m\ Ln\ \lambda$	$\frac{۲\sum_{i=۱}^{m} X_i}{\lambda}$	$۲m$	λ
$Gamma$ $x^{\alpha-۱}\ \frac{e^{-\frac{x}{\beta}}}{\beta^\alpha\Gamma(\alpha)}$ with known α	$\sum_{i=۱}^{m} X_i$	$-\frac{۱}{\beta}$	$-m\alpha\ Ln\ \beta$	$\frac{۲\sum_{i=۱}^{m} X_i}{\beta}$	$۲\alpha m$	β
$Rayleigh$ $x\ \frac{e^{-\frac{x^۲}{۲\beta^۲}}}{\beta^۲}$	$\frac{۱}{۲}\sum_{i=۱}^{m} X_i^۲$	$-\frac{۱}{\beta^۲}$	$-۲mLn\ \beta$	$\frac{\sum_{i=۱}^{m} X_i^۲}{\beta^۲}$	$۲m$	$\beta^۲$
$Pareto$ $\alpha x^{-(\alpha+۱)}$	$\sum_{i=۱}^{m} Ln\ X_i$	$-\alpha$	$mLn\ \alpha$	$۲\alpha\sum_{i=۱}^{m} Ln\ X_i$	$۲m$	$\frac{۱}{\alpha}$
$Weibull$ $\beta\alpha x^{\alpha-۱}\ e^{-\beta x^\alpha}$ with known α	$\sum_{i=۱}^{m} X_i^\alpha$	$-\beta$	$mLn\ \beta$	$۲\beta\sum_{i=۱}^{m} X_i^\alpha$	$۲m$	$\frac{۱}{\beta}$
$Maxwell$ $\sqrt{\frac{۲}{\pi}}\ \beta^{\frac{۳}{۲}}\ x^۲ e^{-\frac{\beta}{۲}x^۲}$	$\frac{۱}{۲}\sum_{i=۱}^{m} X_i^۲$	$-\beta$	$\frac{۳m}{۲}Ln\ \beta$	$\beta\sum_{i=۱}^{m} X_i^۲$	$۳m$	$\frac{۱}{\beta}$
$InverseGaussian$ $\sqrt{\frac{\lambda}{۲\pi x^۳}}\ e^{-\frac{\lambda(x-\mu)^۲}{۲\mu^۲ x}}$ with known μ	$\frac{۱}{۲}\sum_{i=۱}^{m}\frac{(X_i-\mu)^۲}{\mu^۲ X_i}$	$-\lambda$	$m\frac{Ln\ \lambda}{۲}$	$\frac{۱}{۲}\sum_{i=۱}^{m}\frac{\lambda(X_i-\mu)^۲}{\mu^۲ X_i}$	m	$\frac{۱}{\lambda}$

بعضی از توزیعهای متعلق به خانواده کای دو تبدیل یافته :1جدول

مقدمه‌ای بر افشای داده‌ها

دکتر حمیدرضا نواب پور، محمد بردبار

P۱۴۲۷۲

دانشگاه علامه طباطبایی

چکیده: امروزه سازمانهای آماری در کشورهای مختلف اطلاعات بسیار زیادی در مورد جامعه خود جمع‌آوری می‌نمایند. این اطلاعات مواردی همچون جرائم، وضعیت مالی افراد، میزان درآمد شرکتها، موضوعاتی مثل وضعیت استخدامی و وضعیت بهداشت جامعه و ... را شامل می‌شوند. داده‌های جمع‌آوری شده جمع‌بندی و تحلیل می‌گردند و نتایج آن به دو صورت جدول یا داده‌های خرد[۱] منتشر می‌شوند تا کسانیکه برای تحقیق و برنامه‌ریزی به این اطلاعات نیاز دارند، به‌آنها دسترسی پیدا کنند. در تمامی قوانین عمومی آمار، اطلاعات فردی محرمانه تلقی می‌گردد زیرا که ممکن است در صورت انتشار به دلایل مختلف مورد سوءاستفاده قرار گیرند. این عمل نتایج زیان‌باری را برای پاسخ‌دهندگان و در نهایت برای جامعه در پی خواهد داشت. به همین دلیل سازمانهای آماری سعی می‌کنند که از راههای گوناگون جنبهٔ محرمانه‌بودن اطلاعات دریافتی از پاسخ‌دهندگان خود را حفظ نمایند و تا حد ممکن خطر افشای اطلاعات فردی را کاهش دهند. در این مقاله سعی بر آن است تا مفهوم کلی افشای داده‌ها، روشهای گوناگون ارزیابی خطر آن و راههای کاهش این خطر که امروزه به موضوع بحث‌انگیز و مهمی تبدیل شده است مورد بحث قرار گیرد.

واژه‌های کلیدی: افشاء[۲]، داده‌های جدولی[۳]، داده‌های خرد.

۱ مقدمه

محرمانه بودن اطلاعات شخصی و محدودسازی افشای آنها موضوعاتی ذاتاً آماری می‌باشند که تا این اواخر توجه تعداد معدودی از محققان آماری را به خود جلب کرده‌بودند. این وضعیت در دههٔ اخیر تغییر یافته است. قبلاً این مباحث به بخش غیر آماری تنزل داده می‌شدند و در نتیجه شیوه‌هایی که برای پرداختن به آنها مورد استفاده قرار می‌گرفت قدیمی و غیرآماری بودند. بحث روی این موضوعات با یک مقاله توسط دالنیوس در سال ۱۹۷۷

[۱] Microdata
[۲] Disclosure
[۳] Tabular Data

و یک گزارش مشروح از زیر کمیتهٔ روش شناختی محدودسازی افشای اطلاعات به سال ۱۹۷۸ در آمریکا بصورت جدی شروع شد و آمارشناسان به‌آهستگی به تحقیق در این موضوع پرداختند. با اینکه تا به حال تحقیقات زیادی روی این موضوعات صورت گرفته است اما هنوز هم باید برای رسیدن به سطح مطلوب تلاش کرد. ابتدا برخی مفاهیم مربوط به ادبیات افشای داده‌ها مطرح می شود و در ادامه چند روش محدودسازی افشاء را مورد بررسی قرار می‌دهیم.

۲ تعاریف

۱.۲ حفظ محرمانه بودن

بسیاری از داده‌هایی که توسط سازمانهای آماری جمع آوری می‌گردند یا مستقیماً از پاسخ‌دهندگان بدست می آیند که می‌تواند به صورت مصاحبهٔ حضوری، مصاحبهٔ تلفنی، مصاحبهٔ پستی، مصاحبهٔ اینترنتی و... باشد و یا اینکه برای کاهش زحمت پاسخگویان و مینمم کردن هزینه‌های جمع آوری داده‌ها از رکوردهای اداری (ثبتی) استفاده شود. در هر صورت، داده‌های جمع آوری شده به جنبه‌هایی از زندگی شخصی یا تجاری پاسخ‌دهندگان که به نظرشان این جنبه‌ها محرمانه تلقی می‌شود، مربوط می‌شوند. حفاظت از جنبهٔ محرمانه‌بودن داده‌ها یک مسئولیت مشخص و انکار ناپذیر سازمانهای آماری می‌باشد. این مسئولیت به ۳ دلیل زیراست:

الف— حفاظت از جنبهٔ محرمانه‌بودن داده‌ها یک رفتار اخلاقی برای آمارشناسان در حرفهٔ آمار می‌باشد که به تصویب بین‌المللی رسیده است.[۱]

ب— حفاظت از جنبهٔ محرمانه‌بودن داده‌ها اغلب مورد توجه قوانین دولتی می‌باشد.

ج— حقوقدانان معتقدند که اغلب پاسخگویان، بدون تضمین اینکه پاسخهایشان برای یک نهاد سوم فاش نمی‌شوند، حاضر به پاسخگویی نیستند و یا بدرستی پاسخ نمی‌دهند. اگر برای حفظ محرمانه‌بودن داده‌های جمع آوری شده تضمینی ارائه نشود، اعتماد عمومی پاسخ‌دهندگان به سازمانهای آماری کاهش خواهد یافت و این امر باعث کاهش میزان پاسخگویی در آمارگیریها می‌شود که نتیجهٔ آن افت کیفی سرشماریها و تحقیقات بعدی خواهد بود.

۲.۲ افشاء

بحث افشاء مربوط به برداشت نابجا (نامناسب) از اطلاعات موجود در داده‌های مربوط به یک شخص یا سازمان می‌باشد. دانکن در سال ۱۹۹۳ تعریفی از افشاء ارائه کرد که ۳

[۱] International Statistical Institute 1986 and American Statistical Association 1989

نوع از افشاء را در بر دارد.

افشاء هنگامی اتفاق می‌افتد که یک موضوع داده‌ای از یک فایل منتشر شدهٔ داده‌ها مشخص شود(افشای هویت[۱])، یااینکه اطلاعات حساس در مورد یک موضوع داده‌ای توسط یک فایل داده‌های منتشرشده آشکار گردند (افشای صفت[۲])، یا اینکه داده‌های منتشرشده این امکان را بوجود آورد که مقدار بعضی مشخصه‌های یک فرد یا واحد آماری با دقتی بیشتر نسبت به دیگر حالات ممکن ، تعیین شود (افشای استنباطی[۳]). افشای استنباطی ممکن است شامل افشای هویت یا افشای صفت نیز باشد. انتشار داده‌های آماری قطعاً بعضی اطلاعات را در مورد موضوعات شخصی آشکار می‌کنند. افشاء هنگامی اتفاق می‌افتد که اطلاعاتی که جنبهٔ محرمانه دارند آشکار شوند. گاهی اوقات افشاء تنها بر پایهٔ داده‌های منتشرشده اتفاق می‌افتد. گاهی اوقات افشاء، نتیجهٔ ترکیب داده‌های منتشرشده با اطلاعات عمومی در دسترس می‌باشد و گاهی اوقات افشاء نتیجهٔ ترکیب داده‌های منتشرشده با منابع دادهٔ بیرونی است که می‌تواند در دسترس عموم باشند.

۳.۲ دسترسی مقید و داده‌های مقید

برای برخورد با بحث افشای داده‌ها دو فلسفه وجود دارد: یکی محدودکردن امکان دسترسی به داده‌ها و دیگری محدودکردن مقدار و شکل داده‌های منتشرشده است. کمبود اطلاعات در یک قلمرو معین می‌تواند فاجعه‌آفرین باشد. با ورود به عصر اطلاعات تمایل برای بدست‌آوردن هر چه بیشتر اطلاعات بطور سرسام‌آوری افزایش یافته است . با پیشرفت فوق‌العاده وسایل ارتباط‌جمعی در آینده نیز این تمایل با شدت بیشتری ادامه خواهد یافت. به سبب اینکه داده‌های آماری که با هزینهٔ دولت جمع‌آوری می‌شوند، مایملک عمومی محسوب می‌گردند، عدم انتشار اطلاعات با موازین یک جامعهٔ آزاد مغایرت دارد و در کل منطق قابل قبولی برای پیشگیری از افشای اطلاعات محسوب نمی‌شود. محدودکردن دسترسی به اطلاعات موجود در داده‌ها تنها در صورتی باید موجه باشد که از راههای دیگر نتوان جنبهٔ محرمانه بودن اطلاعات را حفظ کرد. معمولاً سازمانهای آماری از گزارش مطالب کلیدی که موجب شناخت دقیق افراد می‌شود (شناسه‌های مستقیم)اجتناب می‌کنند. بعلاوه بسیاری از سازمانهای آماری به منظور کاهش خطر افشای اطلاعات تنها نمونه‌هایی از اطلاعات سرشماری و زیرنمونه‌هایی از نمونه‌گیریهای گسترده را گزارش می‌دهند. با اتخاذ چنین سیاستهایی و در کنار یک سری اقدامات قانونی و نیز اقداماتی در جهت افزایش امنیت سازمانی و کامپیوتری ، شمار روزافزونی از سازمانهای آماری به خط‌مشی‌هایی

---

[۱] Identity Disclosure
[۲] Attribute Disclosure
[۳] Inferential Disclosure

روی می آورند که نیازمند تعدیل داده‌های آماری است. این اقدامات موضوع محدودسازی افشای اطلاعات آماری می‌باشد. پس از یک سو باید حداکثر اطلاعات در اختیار جامعه قرار گیرد و از سوی دیگر باید این تضمین داده شود که حریم شخصی افراد و سایر واحدهای آماری در حد معقولی حفظ می‌شود. پس باید یک توازن بین این دو مسأله برقرار شود. زیرمجموعه‌ای از آمار که در این زمینه فعالیت دارد تحت عناوین مختلفی همچون کنترل افشای آمار، جلوگیری از افشای آمار، حفظ اطلاعات آماری و محرمانه‌بودن آمار عنوان می‌گردد.

۴.۲ جدولها و داده‌های خرد

انتخاب روشهای محدودسازی افشای آماری به طبیعت محصولات داده‌ای که باید جنبهٔ محرمانه بودن آنها حفظ شود بستگی دارد. بیشتر داده‌های آماری به دو شکل جدولها یا فایلهای داده‌خرد منتشر می‌شوند. جدولها شامل دو گروه هستند: جدولهای فراوانی و جدولهای بزرگی[۱]. برای هر گروه داده‌ها می‌توانند به شکل تعداد، نسبت یا درصد بیان شوند.

یک فایل داده‌های خرد شامل رکوردهای شخصی است که هر مقدار آن شامل متغیرهای یک شخص ، یک مؤسسه تجاری یا یک واحد دیگر می‌باشد. بعضی از فایلهای داده‌ها شامل مشخص‌کننده‌های صریحی مثل نام، آدرس یا شمارهٔ کد ملی می‌باشد. حذف کردن هر کدام از این مشخص کننده‌ها از فایل ، یک اقدام اولیه بدیهی برای مهیاکردن زمینهٔ انتشار یک فایل داده‌هااست که بایستی جنبهٔ محرمانه‌بودن اطلاعات شخصی آن حفظ شود.

۵.۲ پوشاندن ماتریسی

یک رده کلی از روشهای محدودسازی افشاء با نام پوشاندن ماتریسی[۲] شناخته می‌شود که خلاصهٔ آن در زیر می آید.

فرض کنید ماتریس Z یک ماتریس nxp باشد که داده‌های مربوط به n شخص یا حالت را برای تعداد p متغیر یا صفت ارائه کرده است. پوشاندن ماتریسی بدین صورت عمل می‌کند: فرض کنید بتوان ماتریس Z را بصورت M=AZB+C به ماتریس M تبدیل کرد که در آن A ماتریسی است که حالات را تبدیل می‌کند و B ماتریس تبدیل صفات یا متغیرها می‌باشد و C ماتریسی است که اعضای Z را مبهم می‌سازد. پوشاندن ماتریسی

[۱] Tables of Magnitude
[۲] Matrix Masking

شامل راهبردهای استاندارد گوناگون گسترده‌ای می‌باشد که چند راهبرد مشهور آن در ادامه آمده است :

الف — انتشار زیرمجموعه‌ای از مشاهدات (حذف سطرهایی از Z).

ب — شبیه‌سازی داده‌ها(افزودن سطرهایی به Z).

پ — افزودن پرشیدگیهای تصادفی[۱] به Z.

ت — انتشار زیر مجموعه‌ای از متغیرها (حذف ستونهایی از Z).

ج — انتشار ماتریس واریانس — کوواریانس .

چ — مبادلهٔ مقادیر ستونی انتخاب شده‌ای با جفتهایی از سطرها.

ح — پنهان‌سازی خانه‌ای[۲] برای طبقه‌بندی‌های متقاطع از داده‌ها.

اگر چه در اینجا حالت دو بعدی داده‌ها را در نظر گرفته‌ایم و آن را بصورت یک ماتریس nxp نمایش دادیم ولی حقیقت این است که پوشاندن ماتریسی قابل تعمیم به ابعاد بالاتر نیز است.

۳ روشهای محدودسازی افشاء

۱.۳ جدولهای فراوانی

۱.۱.۳ نمونه‌گیری به عنوان یک روش محدودسازی افشای آماری

خطر افشاء در مورد داده‌های حاصل از یک بررسی مبتنی بر نمونه‌گیری تقریباً کمتر از خطر افشاء در مورد داده‌های حاصل از یک سرشماری می‌باشد. هر چه کسر نمونه‌گیری کوچکتر باشد احتمال کمتری وجود دارد که شخصی بتواند با استفاده از اطلاعات موجود در داده‌های نمونه به اطلاعات محرمانه افراد یا سازمانهایی در جامعه دست پیدا کند. البته این موضوع ارتباط نزدیکی با طرح نمونه‌گیری و توزیع مشخصهٔ مورد نظر دارد.

هرگاه جدولهای فراوانی را مستقیماً با استفاده از داده‌های سرشماری بسازیم ، بایستی روشهای محدودسازی افشاء را به کار ببریم. دو رده از قوانین محدودسازی افشاء برای جدولهای فراوانی وجود دارد. اولین رده شامل قواعد خاصی برای جدولهای معینی می‌باشد. دومین رده کلی‌تر است : اگر تعداد پاسخگویان در یک خانهٔ جدول از تعداد مشخص‌شده‌ای

---

[۱] Random Perturbatins

[۲] Cell Suppression

کمتر باشد، آن خانه، حساس تعریف میشود (قاعده آستانه‌ای)[۳].

۲.۱.۳ قواعد خاص

قواعد خاص روی سطح جزئیاتی که می‌تواند در یک جدول فراهم شود، محدودیتهایی وضع می‌کند. به عنوان مثال ، انتشار جدولهایی که مقدار یک خانه داخلی آنها با یک جمع حاشیه‌ای برابر است را ممنوع میکنند. این قواعد کلاً برای فراهم کردن داده‌هایی که از طرف سازمان آماری بصورت خاصی ، حساس تلقی می شود مطرح است و از سازمانی به سازمان دیگر متفاوت می‌باشد.

۳.۱.۳ قاعده آستانه‌ای

با استفاده از قاعدهٔ آستانه‌ای اگر تعداد پاسخگویان در خانهٔ جدولی از اعداد مشخص‌شده‌ای کمتر باشد، آن خانه را حساس می‌نامیم . برخی سازمانها خواستار وجود حداقل ۵ پاسخگو در یک خانه می‌باشند وبرخی ۳ پاسخگو و ... . یک سازمان می‌تواند جدولهایی بسازد و گروهها را مانند حالت قبل ترکیب کند یا اینکه از فنون پنهان‌سازی خانه‌ای، گردکردن تصادفی ، گردکردن کنترل‌شده و ویرایش محرمانه‌بودن استفاده کند. جدول ۱ یک جدول با افشاء می‌باشد که مربوط به نوجوانان بزهکار است. خانه‌های دارای کمتر از ۵ پاسخگو را حساس تعریف کرده‌ایم. به عنوان مثال با مراجعه به این جدول مشخص می‌شود که در ناحیه A تنها یک نوجوان بزهکار وجود دارد که میزان تحصیلات سرپرستش خیلی بالا است. با اطلاعات عمومی که ممکن است در مورد تحصیلات سرپرستان خانوارها در ناحیه A در دسترس باشد، این نوجوان به دقت مشخص می‌شود.

الف – پنهان‌سازی[۱]

یکی از عمومی‌ترین راه‌های مورد استفاده برای حفاظت خانه‌های حساس روش پنهان‌سازی است. واضح است که در یک سطر یا ستون با یک خانهٔ حساس پنهان‌شده، حداقل یک خانهٔ اضافی باید پنهان گردد، در غیر این صورت مقدار خانهٔ حساس می‌تواند از طریق جمع حاشیه‌ای متناظر محاسبه گردد. بدین دلیل خانه‌های معین دیگری باید پنهان شوند. این عمل را پنهان‌سازی مکمل گویند. تا هنگامیکه انتخاب خانه‌ها برای پنهان‌سازی مکمل بصورت دستی انجام گیرد، تضمین اینکه نتیجهٔ حاصله ، حفاظت کافی را انجام دهد مشکل است. جدول ۲ مثالی از یک سیستم خانه‌های پنهان‌شده را برای داده‌های جدول

[۳] Threshold Rule
[۱] Suppression

۱ ارائه می‌دهد که در هر سطر و ستون حداقل ۲ خانه پنهان‌شده دارد. به نظر می‌رسد که این جدول خانه‌های حساس را حفظ می‌کند ولی عملاً این‌گونه نیست. به عنوان مثال با استفاده از رابطه:

$$(۱۵ + D۱ + D۲ + D۳) + (۲۰ + D۴ + D۵ + ۱۵) - (D۱ + D۴ + ۱۰ + ۱۴) - (D۲ + D۵ + ۱۰ + ۷) = (۲۰ + ۵۵ - ۳۵ - ۳۰)$$

نتیجه می‌شود: $D۳ = ۱$. یعنی عملاً می‌توان مقدار D۳ را با استفاده از جمعهای حاشیه‌ای تعیین کرد. پس هنوز مسأله افشاء وجود دارد.

این مثال نشان می‌دهد که انتخاب خانه‌ها برای پنهان‌سازی مکمل پیچیده‌تر از آن چیزی است که در ابتدا تصور می‌شد. روشهای ریاضی برنامه‌ریزی خطی برای انتخاب خودکار خانه‌ها به منظور پنهان‌سازی مکمل و همچنین الگوی پنهان‌سازی مفروض به کار می‌روند. جدول ۳ نشان‌دهنده یک سیستم خانه های پنهان است که حفاظت کافی را برای خانه‌های حساس فراهم می‌کند. به هر حال جدول ۳ یکی از مشکلات پنهان‌سازی را نشان می‌دهد. از ۱۶ خانهٔ اولیه تنها ۷ خانه منتشر شده‌اند و ۹ خانهٔ دیگر پنهان شده‌اند.

ب ـ گردکردن تصادفی[۱]

به منظور کاهش تعداد داده‌هایی که در صورت استفاده از روش پنهان‌سازی از دست می‌دهیم روشهای دیگری برای حفاظت خانه‌های حساس در جدولهای فراوانی به کار می‌روند که به روشهای پرشیدگی[۲] موسومند. روش گردکردن تصادفی و گردکردن کنترل شده مثالهایی از روشهای پرشیدگی می‌باشند. در روش گردکردن تصادفی، مقادیر خانه‌ای گرد می‌شوند اما به جای استفاده از قراردادهای استاندارد گردکردن، یک تصمیم تصادفی برای گردکردن به بالا یا پایین اتخاذ می‌گردد. عمل گردکردن برای هر خانه در یک جدول بصورت جداگانه انجام می‌شود ولی جمعهای حاشیه‌ای بدون تغییر باقی می‌مانند. یک جدول که عمل گردکردن تصادفی در آن انجام گرفته‌است میتواند منجر به از دست دادن اطمینان عمومی به اعداد گردد.

ج ـ گردکردن کنترل شده[۳]

[۱] Random Rounding
[۲] Perturbation Methods
[۳] Controlled Rounding

برای حل مشکل روش گردکردن تصادفی، روش گردکردن کنترل‌شده توسعه یافته است. این روش همانند روش گردکردن تصادفی است اما تاکید بر این است که مجموع مقادیر منتشرشده در هر سطر و ستون برابر جمعهای حاشیه‌ای متناظر باشد. این روش در سالهای ۱۹۷۰ تا ۱۹۸۰ مورد تحقیق قرار گرفت و از آن در سرشماری سال ۱۹۹۰ آمریکا استفاده شد. یک مشکل این روش این است که به یک برنامه کامپیوتری مخصوص نیاز دارد که در حال حاضر این برنامه‌ها به گستردگی در دسترس نیست. مشکل دیگر آن این است که پاسخهای گردکردن کنترل‌شده برای جدولهای پیچیده ممکن است همواره وجود نداشته باشد.

د- ویرایش محرمانه‌بودن [۱]

در این روش دو راهبرد متفاوت وجود دارد. یکی برای داده‌های سرشماری استفاده می‌شود و دیگری برای داده‌های نمونه‌ای به کار می‌رود. هر دو فن، روشهای محدودسازی افشاء را برای فایلهای داده‌های خرد، قبل از اینکه این فایلها برای تنظیم جدولها به‌کار روند استفاده می‌کنند و خود این فایلها منتشر نمی‌شوند.

روش اول: برای فایل داده‌های خرد مربوط به سرشماریها ویرایش محرمانه‌بودن شامل تبادل داده‌ها یا معاوضه می‌باشد. مراحل ویرایش محرمانه‌بودن در زیر آمده‌است:

۱— گرفتن یک نمونه از رکوردها از فایل داده‌های خرد.

۲— پیدا کردن یک همانند برای این رکوردها در بعضی نواحی جغرافیایی دیگر (همانندیابی روی یک مجموعه مشخص از صفات مهم).

۳— تعویض تمامی صفات روی رکوردهای همانند شده.

ادارهٔ سرشماری برای حفاظت بیشتر از بلوکهای کوچک، کسر نمونه‌گیری را افزایش می‌دهد. پس از آماده‌شدن فایل، از آن برای آماده‌کردن جدولها استفاده می‌شود.

روش دوم: فایل داده‌های نمونه‌ای شامل داده‌های یک نمونه از جامعه می‌باشد و همانطور که قبلاً ذکر کردیم، انجام نمونه‌گیری تقریباً حفاظت افشاء را فراهم می‌کند. مطالعات نشان می‌دهد که این حفاظت به جز در نواحی کوچک جغرافیایی، کافی است . برای توضیح ویرایش محرمانه‌بودن به کار رفته در فایل داده‌های خرد سرشماری، از رکوردهای ساختگی برای ۲۰ شخص در ناحیه A که در جدولهای قبلی به کار رفته‌اند استفاده

[۱] Confidentiality Edit

می‌کنیم. جدول ۴، پنج متغیر را برای این اشخاص نشان می‌دهد.

برای به کاربردن روش ویرایش محرمانه‌بودن بایستی مراحل زیر را طی کنیم :

۱ — گرفتن یک نمونه از رکوردها از فایل داده‌های خرد (مثلاً یک نمونه ۱۰ درصدی).

۲— وقتی به جدولهایی با تفکیک ناحیه و سطح تحصیلات نیاز داریم، همانندی در ناحیه‌ای دیگر روی متغیرهای دیگری مثل رنگ پوست ، جنسیت و درآمد پیدا می‌کنیم . (در نتیجه جمعهای حاشیه‌ای برای این متغیرها با عمل تبادل بدون تغییر خواهند ماند). یک همانند برای رکورد شماره ۴ (پیت) در ناحیه B پیدا شد. شخص همانند، آلفُنسو می‌باشد که سرپرستش تحصیلات بالایی دارد. رکورد شماره ۱۷ (مایک) با جُرج از ناحیه D همانند شد که سرپرستش دارای تحصیلات متوسط است. همچنین بخشی از نمونهٔ ۱۰ درصدی انتخاب شده از نواحی دیگر دارای رکوردهای نظیری در ناحیه A می‌باشند. یک رکورد از ناحیه D جون با سرپرست دارای تحصیلات بالا با رکورد ویرجینیا همانند شد (شماره ۱۲). یک رکورد از ناحیه C (هیتر با سرپرست دارای تحصیلات پایین)با رکورد نانسی همانند شد (شماره ۲۰).

۳— پس از اینکه تمام همانندها ساخته شدند، صفات آنها روی رکوردهای همانند شده مبادله می‌شوند. فایل داده‌خرد مرتب‌شده در جدول ۵ آمده است.

۴— استفادهٔ مستقیم از فایل داده‌های مبادله‌شده برای تولید جدولها (جدول ٦).

اگر چه برخی از خانه‌های جدول دارای تعداد پاسخگوی کمتر از ۵ می‌باشد اما چون مبادله صورت گرفته است، نمی‌توان رکورد یک شخص را در جامعه بطور دقیق تعیین کرد.

روش ویرایش محرمانه‌بودن این مزیت را دارد که می‌تواند به آسانی برای جدولهای چندبعدی آماده شود و همیشه می‌تواند از افشای محرمانه‌بودن جلوگیری کند.

۲.۳ جدولهای بزرگی

جدولهای داده‌های بزرگی یک مجموعه منحصربه‌فرد از مسائل افشاء را دارند. داده‌های بزرگی عموماً کمیتهایی نامنفی می‌باشند که در تحقیقات یا سرشماریها از مؤسسه‌های تجاری، مزارع یا سازمانها گزارش می‌شوند. این داده‌ها ممکن است بر درآمد اشخاص، میزان فروش یا میزان درآمد مؤسسات دلالت کنند. توزیع این مقادیر معمولاً چوله می‌باشد که در آن تعداد کمی از اعضاء دارای مقادیر خیلی بزرگی هستند. محدودسازی افشاء در این جدولها برای اطمینان دادن اینکه داده‌های منتشرشده برای برآورد مقادیر گزارش شده

توسط بزرگترین واحد استفاده نخواهند شد، متمرکز شده‌است. در این حالت امکان اینکه نمونه‌گیری به تنهایی بتواند حفاظت افشاء را فراهم کند کمتر است، زیرا واحدهایی که به علت اندازه‌شان نمایانی بالایی دارند بوسیلهٔ نمونه‌گیری نمی‌توانند حفاظت شوند. قواعد پنهان‌سازی مقدماتی یا اندازه‌های حساس خطی برای تعیین اینکه آیا یک خانهٔ جدولی داده‌شده می‌تواند اطلاعات پاسخ‌دهندهٔ شخصی را نمایان کند، توسعه داده شده‌اند. چنین خانه‌ای، خانهٔ حساس نامیده می‌شود و نمی‌تواند منتشر شود. قواعد پنهان‌سازی مقدماتی که عموماً برای تشخیص خانه‌های حساس استفاده می‌شوند، قاعده (n,k)، قاعده p درصد و قاعده pq می‌باشد. همهٔ این قواعد بر این پایه استوار هستند که امکان برآورد دقیق مقدار گزارش شده توسط یک پاسخگو را برای شخصی دیگر مشکل نمایند. بزرگترین مقدار گزارش شده دارای بیشترین احتمال برای برآوردشدن دقیق می‌باشد. قواعد پنهان‌سازی مقدماتی برای داده‌های فراوانی نیز قابل استفاده می‌باشند. هرگاه خانه‌های حساس مشخص شوند تنها دو گزینه وجود خواهد داشت : بازسازی جدول و کاهش خانه‌ها تا اینکه خانهٔ حساسی در جدول باقی نماند یا پنهان‌سازی خانه‌ای.

هرگاه خانه‌های حساس مشخص شوند آنها را از انتشار کنار می‌گذاریم. این خانه‌ها پنهان‌سازی‌های مقدماتی نامیده می‌شوند. خانه‌های دیگر که پنهان‌سازی‌های مکمل نامیده می‌شوند، انتخاب می‌گردند و پنهان میشوند تا مقدار خانه‌های حساس باتوجه به جمعهای حاشیه‌ای منتشرشده قابل تعیین نباشند. مشکلات مربوط به پنهان‌سازی خانه‌ای در جدولهای بزرگی همانند جدولهای فراوانی می‌باشد. یک راه اجتناب از مشکلات مربوط به بکارگیری پنهان‌سازی خانه‌ای توسط تعدادی از سازمانهای آماری بکار گرفته شده‌است که عبارت است از گرفتن اجازهٔ کتبی از پاسخگویان برای انتشار یک خانهٔ حساس .

۳.۳ داده‌های خرد

اطلاعات جمع‌آوری شده دربارهٔ مؤسسات در ابتدا بصورت داده‌های بزرگی هستند. این داده‌ها به شدت چوله می‌باشند و پاسخ‌گویانی که چشم‌آمد هستند، بواسطهٔ اطلاعات دیگری که در دسترس عمومی می‌باشند، می‌توانند به آسانی مشخص شوند. یک نتیجه این می‌تواند باشد که فایلهای داده‌های خرد برای استفاده عموم منتشر نشوند. حفاظت یک فایل داده‌های خرد به علت وجود دیگر منابع داده‌ای بسیار مشکل می‌باشد. بعلاوه اندازه قابل قبولی از خطر افشاء برای یک فایل داده‌های خرد وجود ندارد. شیوه‌های حفاظت از فایلهای داده‌های خرد در ادامه توصیف شده‌اند. این شیوه‌ها توسط تمام سازمانهایی که فایلهای داده‌های خرد را برای استفاده عموم فراهم می‌کنند به‌کار می‌رود. به منظور کاهش خطر افشاء، تمام فایلهای داده‌های خرد مورد استفاده عمومی باید شرایط زیر را دارا باشند:

۱ — باید تنها شامل داده‌های یک نمونه از جامعه باشند.

۲— شامل متغیرهای چشم‌آمد [۱] نباشند.

۳— جزئیات جغرافیایی محدود شده باشند.

۴— تعداد متغیرهای موجود در فایل محدود شده باشند.

روشهای اضافه‌ای که برای متغیرهای چشم‌آمد استفاده می‌شوند عبارتند از:

۱— کدبندی به بالا یا کدبندی به پایین.

۲— بازکدبندی [۲] به بازه‌ها یا گردکردن .

۳— افزودن یا ضرب کردن اعداد تصادفی (نوفه [۳]).

۴— مبادله یا مبادلهٔ رتبه‌ای.

۵— انتخاب رکوردهایی به تصادف، حذف متغیرهای انتخاب‌شده و جانهی برای آنها.

٦— تجمع [۴] در طول گروههای کوچک پاسخ‌گویان و جایگذاری یک مقدار گزارش‌شدهٔ شخصی با مقدار متوسط پاسخهای گروههای انتخاب شده.

این روشها را بایک مثال ساختگی که در قسمتهای قبل استفاده کردیم توضیح می‌دهیم .

---

[۱] High Visibility Variables
[۲] Recoding
[۳] Noise
[۴] Aggregation

۱.۳.۳ نمونه‌گیری، حذف مشخص‌کننده‌ها و محدودسازی جزئیات جغرافیایی

الف — تنها داده‌های یک نمونه از جامعه را در نظر می‌گیریم. برای مثال از یک نمونهٔ ۱۰ درصدی از جامعه نوجوانان بزهکار استفاده نموده‌ایم. بخشی از جامعهٔ آماری (ناحیه) A در جدول ۴ نشان داده شده است .

ب — حذف مشخص‌کننده‌های صریح. در این مثال مشخص‌کننده نام نوجوان است.

ج — محدود کردن جزئیات جغرافیایی. در این مثال تصمیم گرفتیم داده‌های ناحیه‌ای شخصی مربوط به یک ناحیهٔ دارای کمتر از ۳۰ نوجوان بزهکار در جامعه را منتشر نکنیم. بنابراین نواحی A و C از جدول ۱ در جدول ۷ ترکیب شده و بصورت AC نشان داده شده است . در این مثال تنها ۵ متغیر برای هر نوجوان در نظر گرفته‌ایم . شاید به نظر برسد که این ۵ متغیر از مجموعه دادهٔ کاملتری شامل متغیرهای نام مادر، نام خواهر و برادر، تعداد خواهر و برادر، سن نوجوان، سن خواهر و برادر، آدرس مدرسه و غیره انتخاب شده‌اند. هر چه تعداد متغیرها درفایل داده‌های خرد مربوط به هر نوجوان بیشتر باشد، ترکیبات یکتای متغیرها احتمال تشخیص نوجوانی را با استفاده از دانش شخصی افزایش می‌دهد. محدود کردن تعداد متغیرها به ۵ متغیر احتمال چنین تشخیصی را کاهش خواهد داد.

۲.۳.۳ متغیرهای چشم‌آمد

ممکن است در جامعه اطلاعاتی موجود باشد که در دسترس دیگران است. مثلاً می‌توان با استفاده از داده‌های درآمد که در جدول ۷ آمده است، خانواده یک نوجوان مجرم را بصورت منحصر به فرد تشخیص داد. به عنوان مثال، کارفرمای سرپرست خانوار معمولاً حقوق دقیق او را می‌داند. چنین متغیرهایی، متغیرهای چشم آمد نامیده می‌شوند و احتیاج به محافظت اضافی دارند.

الف — کدبندی به بالا، کدبندی به پایین و بازکدبندی به بازه‌ها

مقادیر زیاد درآمد به بالا کدبندی می‌شوند، به این صورت که فقط نشان می‌دهیم که میزان درآمد بیشتر از ۱۰۰ هزار دلار در سال است. مقادیر کوچک درآمد به پایین کدبندی می‌شوند، بدین صورت که فقط مشخص می‌کنیم که درآمد کمتر از ۴۰ هزار دلار درسال می‌باشد و درآمدهای دیگر به بازه‌های ۱۰ هزار دلاری بازکدبندی می‌شوند. نتیجهٔ این اقدامات فایلی می‌باشد که در جدول ۸ برای استفادهٔ عموم آمده است . روشهای کدبندی به بالا، کدبندی به پایین و بازکدبندی به بازه‌ها عمومی‌ترین روشهای مورد استفاده برای فراهم‌کردن حفاظت متغیرهای چشم آمد در فایلهای داده‌های خرد می‌باشد.

ب — افزودن پرشیدگی تصادفی

یک روش دیگر برای مخفی‌کردن متغیرهای چشم‌آمد از قبیل درآمد، افزودن یک عدد تصادفی به مقادیر این متغیرها یا ضرب یک عدد تصادفی در مقادیر این متغیرها می‌باشد. به عنوان نمونه، در مثال قبل فرض کنید می‌خواهیم یک متغیر تصادفی از توزیع نرمال با میانگین صفر و انحراف استاندارد ۵ به درآمد اضافه کنیم. ممکن است این عمل در حین نمونه‌گیری، حذف مشخص‌کننده‌ها و محدودکردن جزئیات جغرافیایی انجام شود. نتیجه در جدول ۹ آمده است. برای تولید این جدول، ۱۴ عدد تصادفی از توزیع نرمال مشخص شده انتخاب گردید و به داده‌های درآمد در جدول ۷ اضافه شد.

ج – مبادله یا مبادلهٔ رتبه‌ای

مبادله‌کردن شامل انتخاب یک نمونه از رکوردها، یافتن یک همانند در بانک داده‌ها روی یک مجموعه از متغیرهای از پیش تعیین‌شده و مبادله‌کردن تمام متغیرها می‌باشد. مبادله‌کردن در بخش ویرایش محرمانه‌بودن برای جدولهای داده‌های اندازه مورد بحث قرار گرفت. در آن مثال رکوردها از نواحی مختلف مشخص بودند و روی متغیرهای رنگ پوست ، جنسیت و درآمد همانندیابی شدند، آنگاه نام کوچک نوجوان و میزان تحصیلات سرپرست خانوار مبادله شدند. به منظور فراهم کردن محافظت اضافی برای متغیر درآمد در یک فایل داده‌های خرد، می‌توان روی متغیرهای رنگ پوست و تحصیلات سرپرست همانندیابی کنیم و آنگاه متغیر درآمد را مبادله نماییم. مبادله‌کردن رتبه‌ای یک راه برای استفاده از متغیرهای پیوسته برای تعریف جفت رکوردهای قابل مبادله ارائه می‌کند. به جای اینکه دقیقاً همانند متغیرها را بیابیم، آنها را بطور تقریبی و براساس یک لیست مرتب‌شده تحت متغیری پیوسته تعریف می‌کنیم. رکوردهایی که بر اساس این متغیر مرتب شده‌اند و دارای رتبه نزدیکی هستند را به عنوان جفتهای قابل مبادله در نظر می‌گیریم. دراین روش، متغیری که برای مرتب کردن در نظر می‌گیریم اغلب یکی از متغیرهایی است که مبادله خواهد شد.

د– حذف و جانهی برای رکوردهای انتخاب شده تصادفی

حذف و جانهی شامل انتخاب تعدادی از رکوردهای فایل داده‌های خرد، حذف متغیرهای انتخاب‌شده و جایگذاری آنها با مقادیر انتسابی می‌باشد. این فن با استفاده از داده‌های جدول ۷ توضیح داده می‌شود. ابتدا از هر ناحیهٔ قابل انتشار یعنی AC ، B و D یک رکورد بصوررت تصادفی انتخاب می‌شود. در رکورد انتخاب شده مقدار درآمد با یک مقدار انتسابی جایگذاری می‌شود. اگر رکورد منتخب از ناحیه AC رکورد شماره ۲ و در ناحیه B رکورد شماره ٦ و در ناحیه D رکورد شماره ۱۳ باشد، مقدار انتسابی به این رکوردها ممکن است به ترتیب اعداد ٦۳ ، ۵۲ و ۴۹ باشد. این اعداد ساختگی هستند اما می‌توان دید که مقادیر انتسابی با متوسط میزان درآمد خانوارها در یک ناحیه با رنگ پوست و

سطح تحصیلات یکسان، برابر می‌باشد.

ه — نامشخص کردن

در روش نامشخص کردن یک مقدار گزارش‌شده با یک مقدار متوسط جایگذاری می‌شود. راههای زیادی برای نامشخص کردن وجود دارد. گروههایی از رکوردها که برای متوسط‌گیری انتخاب می‌کنیم شاید با همانندیابی روی متغیرهای دیگر تعیین شوند و یا از طریق مرتب‌کردن داده‌ها روی متغیر مورد علاقه‌مان حاصل شوند. تعداد رکوردهای یک گروه که داده‌های آنها متوسط‌گیری می‌شود ممکن است ثابت یا تصادفی باشند. مقدار متوسط مربوط به یک گروه خاص ممکن است به همه اعضای گروه تخصیص داده‌شود و یا اینکه به عضو میانی گروه تخصیص یابد. به عنوان مثال این فن را برای نامشخص کردن داده‌های درآمد در مثال قبل تشریح می‌کنیم. در فایل داده‌های خرد کامل ممکن است روی متغیرهای مهمی همچون ناحیه، رنگ پوست و دو گروه تحصیلی (خیلی بالا و بالا) و (متوسط و پایین)همانندیابی را انجام دهیم. آنگاه نامشخص کردن می‌تواند متوسط‌گیری خانوارها در هر گروه باشد. با توجه به جدول ۴ شاید این‌گونه برداشت شود که بایستی درآمد خانوار برای گروه شامل جان و سو با مقدار متوسط درآمدشان جایگذاری شود(۱۳۹)، درآمد خانوار برای گروه شامل جیم و پیت باید با مقدار متوسط درآمدشان جایگذاری شود(۸۲) و به همین صورت ادامه می‌دهیم. پس از نامشخص کردن، فایل داده‌ها می‌تواند به عنوان موضوعی برای نمونه‌گیری، حذف کردن مشخص‌کننده‌ها و محدودسازی جزئیات جغرافیایی به کار رود.

۴ خلاصه

در این مقاله سعی شد مفاهیم مربوط به ادبیات افشای داده‌ها مورد بررسی قرار گیرد و روشهای استاندارد محدودسازی افشاء که توسط بسیاری از سازمانهای آماری در مورد جدولها و فایلهای داده‌های خرد استفاده می‌شود، با مثالهایی ساده معرفی گردند. هر کدام از روشهای مطرح‌شده دارای ابعاد کاملتر و پیچیده‌تری است که نیازمند تحقیق و کار بسیار بیشتری می‌باشد.

ناحیه	پایین	متوسط	بالا	خیلی بالا	جمع
A	۱۵	۱	۳	۱	۲۰
B	۲۰	۱۰	۱۰	۱۵	۵۵
C	۳	۱۰	۱۰	۲	۲۵
D	۱۲	۱۴	۷	۲	۳۵
جمع	۵۰	۳۵	۳۰	۲۰	۱۳۵

جدول ۱ : تعداد نوجوانان بزهکار به تفکیک ناحیه و سطح تحصیلات سرپرست (با افشاء)

ناحیه	پایین	متوسط	بالا	خیلی بالا	جمع
A	۱۵	D۱	D۲	D۳	۲۰
B	۲۰	D۴	D۵	۱۵	۵۵
C	D٦	۱۰	۱۰	D۷	۲۵
D	D۸	۱۴	۷	D۹	۳۵
جمع	۵۰	۳۵	۳۰	۲۰	۱۳۵

جدول ۲ : تعداد نوجوانان بزهکار به تفکیک ناحیه و سطح تحصیلات سرپرست (با افشاء)

ناحیه	پایین	متوسط	بالا	خیلی بالا	جمع
A	۱۵	F	F	F	۲۰
B	۲۰	F	F	۱۵	۵۵
C	F	۱۰	۱۰	F	۲۵
D	F	۱۴	۷	F	۳۵
جمع	۵۰	۳۵	۳۰	۲۰	۱۳۵

جدول ۳: تعداد نوجوانان بزهکار به تفکیک ناحیه و سطح تحصیلات سرپرست (بدون افشاء)، خانه‌هایی که با F نشان داده شده‌اند پنهان شده‌اند

شماره	نوجوان	ناحیه	سطح تحصیلات سرپرست	درآمد سرپرست	رنگ پوست
۱	John	A	خیلی بالا	۲۰۱	سیاه
۲	Jim	A	بالا	۱۰۳	سفید
۳	Sue	A	بالا	۷۷	سیاه
۴	Pete	A	بالا	٦۱	سفید
۵	Ramesh	A	متوسط	۷۲	سفید
٦	Danis	A	پایین	۱۰۳	سفید
۷	Virgil	A	پایین	۹۱	سیاه
۸	Wanda	A	پایین	۸۴	سفید
۹	Stan	A	پایین	۷۵	سفید
۱۰	Irmi	A	پایین	٦۲	سیاه
۱۱	Renee	A	پایین	۵۸	سفید
۱۲	Virginia	A	پایین	۵٦	سیاه
۱۳	Mary	A	پایین	۵۴	سیاه
۱۴	Kim	A	پایین	۵۲	سفید
۱۵	Tom	A	پایین	۵۵	سیاه
۱٦	Ken	A	پایین	۴۸	سفید
۱۷	Mike	A	پایین	۴۸	سفید
۱۸	Joe	A	پایین	۴۱	سیاه
۱۹	Jeff	A	پایین	۴۴	سیاه
۲۰	Nancy	A	پایین	۳۷	سفید

جدول ۴: رکوردهای نوجوانان بزهکار در ناحیه A (درآمد بر حسب هزار دلار است)

شماره	نوجوان	ناحیه	سطح تحصیلات سرپرست	درآمد سرپرست	رنگ پوست
۱	John	A	خیلی بالا	۲۰۱	سیاه
۲	Jim	A	بالا	۱۰۳	سفید
۳	Sue	A	بالا	۷۵	سیاه
۴	Alfonso	A	خیلی بالا	٦۱	سفید
۵	Ramesh	A	متوسط	۷۲	سفید
٦	Danis	A	پایین	۱۰۳	سفید
۷	Virgil	A	پایین	۹۱	سیاه
۸	Wanda	A	پایین	۸۴	سفید
۹	Stan	A	پایین	۷۵	سفید
۱۰	Irmi	A	پایین	٦۲	سیاه
۱۱	Renee	A	پایین	۵۸	سفید
۱۲	June	A	بالا	۵٦	سیاه
۱۳	Mary	A	پایین	۵۴	سیاه
۱۴	Kim	A	پایین	۵۲	سفید
۱۵	Tom	A	پایین	۵۵	سیاه
۱٦	Ken	A	پایین	۴۸	سفید
۱۷	George	A	متوسط	۴۸	سفید
۱۸	Joe	A	پایین	۴۱	سیاه
۱۹	Jeff	A	پایین	۴۴	سیاه
۲۰	Heather	A	پایین	۳۷	سفید

جدول ۵: رکوردهای نوجوانان بزهکار در ناحیه A پس از مبادلهٔ رکوردها(درآمد بر حسب هزار دلار است)

ناحیه	پایین	متوسط	بالا	خیلی بالا	جمع
A	۱۳	۲	۳	۲	۲۰
B	۱۸	۱۲	۸	۱۷	۵۵
C	۵	۹	۱۱	۰	۲۵
D	۱۴	۱۲	۸	۱	۳۵
جمع	۵۰	۳۵	۳۰	۲۰	۱۳۵

جدول ٦: تعداد نوجوانان بزهکار به تفکیک ناحیه و سطح تحصیلات سرپرست (بدون افشاء)

شماره	ناحیه	سطح تحصیلات	درآمد سرپرست	رنگ پوست
۱	AC	بالا	٦۱	سفید
۲	AC	پایین	۴۸	سفید
۳	AC	متوسط	۳۰	سیاه
۴	AC	متوسط	۵۲	سفید
۵	AC	خیلی بالا	۱۱۷	سفید
٦	B	خیلی بالا	۱۳۸	سیاه
۷	B	خیلی بالا	۱۰۳	سفید
۸	B	پایین	۴۵	سفید
۹	B	متوسط	٦۲	سفید
۱۰	B	بالا	۸۵	سفید
۱۱	D	پایین	۳۳	سیاه
۱۲	D	متوسط	۵۱	سیاه
۱۳	D	متوسط	۵۹	سفید
۱۴	D	بالا	۷۲	سیاه

جدول ۷: رکوردهای نوجوانان بزهکار پس از حذف مشخص‌کننده‌ها(درآمد بر حسب هزار دلار است)

شماره	ناحیه	سطح تحصیلات	درآمد سرپرست	رنگ پوست
۱	AC	بالا	٦۰–٦۹	سفید
۲	AC	پایین	۴۰–۴۹	سفید
۳	AC	متوسط	<۴۰	سیاه
۴	AC	متوسط	۵۰–۵۹	سفید
۵	AC	خیلی بالا	>۱۰۰	سفید
٦	B	خیلی بالا	>۱۰۰	سیاه
۷	B	خیلی بالا	>۱۰۰	سفید
۸	B	پایین	۴۰–۴۹	سفید
۹	B	متوسط	٦۰–٦۹	سفید
۱۰	B	بالا	۸۰–۸۹	سفید
۱۱	D	پایین	<۴۰	سیاه
۱۲	D	متوسط	۵۰–۵۹	سیاه
۱۳	D	متوسط	۵۰–۵۹	سفید
۱۴	D	بالا	۷۰–۷۹	سیاه

جدول ۸: رکوردهای نوجوانان بزهکار پس از حذف مشخص‌کننده‌ها، محدودکردن جزئیات جغرافیایی و کدبندی (درآمد بر حسب هزار دلار است)

شماره	ناحیه	سطح تحصیلات	درآمد سرپرست	رنگ پوست
۱	AC	بالا	۶۱	سفید
۲	AC	پایین	۴۲	سفید
۳	AC	متوسط	۳۲	سیاه
۴	AC	متوسط	۵۲	سفید
۵	AC	خیلی بالا	۱۲۳	سفید
۶	B	خیلی بالا	۱۳۸	سیاه
۷	B	خیلی بالا	۹۴	سفید
۸	B	پایین	۴۶	سفید
۹	B	متوسط	۶۱	سفید
۱۰	B	بالا	۸۲	سفید
۱۱	D	پایین	۳۱	سیاه
۱۲	D	متوسط	۵۲	سیاه
۱۳	D	متوسط	۵۵	سفید
۱۴	D	بالا	۶۱	سیاه

جدول ۹: رکوردهای نوجوانان بزهکار پس از افزودن پرشیدگی تصادفی (درآمد بر حسب هزار دلار است)

مراجع

1- Bader , E. *Basic right information-Basic right to data protection: A contradiction*, Bulletin of the international statistical institute 49th session, 1993, p.63.

2- Cox, L.H. and Zayatz, L.V. *An Agenda for Research in Statistical Disclosure Limitation*, Journal of Official Statistics, Vol.11, No.2 , 1995, pp. 205-220.

3- Dalenius T. Towards a methodology for statistical disclosure control. Statistisk Tidskrift 1977.

4- Duncan G.T., Jabin T.B., de Wolf V.A. (eds). Private Lives and Public Policies: Confidentiality and Accessibility of Goverment Statistics. Panel on Confidentiality and Data Access, Committee on National Statistics. National Academy Press: Washington, D.C., 1993.

5- Fienberg, S.E. *Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality*, Journal of official Statistics Vol. 10, No. 2, 1994, pp.115-132.

6- Fienberg, S.E. and Makov, U.E. *Confidentiality , Uniqeness and Disclosure Limitation for Categorical Data*, Journal of official Statistics, Vol.14 No. 1998, pp.385-397.

7- Fienberg, S.E. *Confidentiality and data access in public health*, Statist. Med. 2001 ; 20: 1347-1356.

8- Subcommitee on Disclosure-Avoidance Techniques. Statistical Policy Working Paper No. 22: Report on statistical disclosure limitation methodology. Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, D.C., 1994.

9- Subcommitee on Disclosure-Avoidance Techniques. Statistical Policy Working Paper No. 2: Report on statistical disclosure and disclosure avoidance techniques. Federal Committee on Statistical Methodology, Office of Federal Policy and Standards, U.S. Department of Commerce, Washington, D.C., 1978.

توسعه کنترل فرآیند آماری با استفاده از رویکرد بیز تجربی

رسول نورالسناء[۱]، محمد رضا لطفی [۲]

$P۱۲۰۴۸$

[۱] دانشکده صنایع، دانشگاه علم و صنعت ایران
[۲] دانشگاه آزاد واحد علوم تحقیقات

چکیده: دراین‌تحقیق، کنترل فرآیند آماری جهت یک فرآیند با توزیع پواسن دارای پارامتر λ_t که این پارامتر در طی زمان تغییر می‌کند، توسعه داده می‌شود. برای این پارامتر متغیر، توزیع گاما بعنوان توزیع پیشین منظور شده و جهت بروز نمایی آن از مشاهدات اخیر، یک برآورد پسین ارائه می‌گردد. ساختار فوق شرایط استفاده از تئوری بیز تجربی را فراهم می‌سازد. این موضوع مدلی بهینه جهت کنترل فرآیند ایجاد می‌کند و با استفاده از روابط بازگشتی سرعت پردازش اطلاعات افزایش می‌یابد و حجم داده‌ها و ذخیره‌سازی اطلاعات را در شبکه‌های رایانه‌ای در فرآیندهای بزرگ و پیچیده کاهش می‌دهد.

واژه‌های کلیدی: کنترل فرآیند،بیز تجربی، توزیع پواسن،توزیع گاما، برآورد میانگین متحرک وزنی، برآوردهای بازگشتی.

۱ مقدمه

در کنترل فرآیند تولید با استفاده از روشهای آماری، برای مشخصه‌های کیفی محصول توزیع احتمال مناسب منظور می‌گردد و سعی می‌شود با کنترل نمودن پارامترهای آن، عملکرد فرآیند تولید را کنترل نمود و از این طریق میزان ضایعات و محصولات معیوب را کاهش داد. از دهه ۱۹۲۰ که آقای والترشوهارت کنترل فرآیند آماری را معرفی کرد معمولا فرض اصلی، ثابت بودن پارامترهای توزیع مفروض بوده‌است. به دلیل پیچیده شدن فرآیندهای تولید ضرورت می‌یابد که این فرض مورد تجدیدنظر قرار گیرد. در سالهای اخیر تحقیقات مهمی در زمینه کاربرد تئوری بیز تجربی در کنترل فرآیند بعمل آمده است. بویژه یوسری و دیگران (۱۹۹۱) از رویکرد بیزی برای یک توزیع دوجمله‌ای با پارامتر متغیر p_t استفاده کرده‌اند. فلتز و شیا (۲۰۰۱) این رویکرد را برای توزیع نرمال چند متغیره با پارامتر متغیرμ بکار بردند. ویلو و رودی (۲۰۰۱) از تئوری بیز درآزمون تایید قابلیت استفاده می‌کند. همچنین مارتز و والر (۱۹۸۹) در کتاب تجزیه و تحلیل قابلیت بیزی کاربرد تئوری بیزی را در تجزیه و تحلیل قابلیت اعتماد و دوام معرفی کردند.

جین و گریمناو (۱۹۹۳) کنترل کیفیت چند متغیره با رویکرد بیزی را ارائه دادند. چون مشخصه‌های کیفی در بسیاری از فرآیندها از توزیع پواسن پیروی می‌کند ما دراین مقاله ، کاربرد تئوری بیز در کنترل فرآیند آماری را با در نظر گرفتن توزیع پواسن با پارامتر متغیر λ توسعه می‌دهیم.

۲ مدل

اگر X_t تعداد نقص یا قطعات معیوب در زمان t و دارای توزیع پواسن با پارامتر λ_t باشد دراینصورت تابع احتمال X_t بشرح زیر است:

$$f(x_t|\lambda_t) = \frac{e^{-\lambda_t}\lambda_t^{x_t}}{x_t!} \qquad x_t = ۰, ۱, ۲, \ldots \qquad (۱)$$

در این تحقیق فرض می‌شود پارامتر λ_t ثابت نمی‌باشد و در طی زمان تغییر می‌کند و یک متغیر تصادفی است و دارای توزیع گاما با پارامترهای α و β است. بنابراین :

$$f(\lambda_t) = \frac{۱}{\beta^\alpha \Gamma(\alpha)}\lambda_t^{\alpha-۱}e^{-\lambda_t/\beta} \qquad \lambda_t > ۰, \ \ \alpha, \beta > ۰. \qquad (۲)$$

در اینصورت به این تابع احتمال، توزیع احتمال پیشین پارامتر λ_t گفته می‌شود. میانگین و واریانس این توزیع پیشین که به آن میانگین و واریانس فرآیند نیز می‌گویند عبارت است از :

$$\mu = \alpha\beta \qquad (۳)$$

$$\gamma^۲ = \alpha\beta^۲ \qquad (۴)$$

برای برآورد λ_t در زمان t و بروزنمایی آن از نمونه‌گیری و مشاهدات اخیراستفاده می‌شود. اگر X_t تعداد نقص در نمونه دوره t باشد، براساس تئوری بیز تجربی تابع چگالی احتمال λ_t به شرط X_t بشرح زیراست و به آن توزیع احتمال پسین پارامتر λ_t گفته می‌شود.

$$f(\lambda_t|x_t) = \frac{f(\lambda_t)f(x_t|\lambda_t)}{\int f(\lambda_t)f(x_t|\lambda_t)d\lambda_t} \qquad (۵)$$

می‌توان نشان داد که این توزیع احتمال، یک توزیع گاما با پارامترهای $x_t+\alpha$ و $\frac{\beta}{۱+\beta}$ است ، درنتیجه میانگین و واریانس پسین λ_t می‌شود:

$$\mu_{\lambda_t} = \frac{\beta(x_t+\alpha)}{\beta+۱}, \qquad \sigma^۲_{\lambda_t} = \frac{\beta^۲(x_t+\alpha)}{(\beta+۱)^۲} \qquad (۶)$$

اگر واریانس کل X_t با σ_T^2 نشان داده شود، این تغییر پذیری کل از دو منبع تغییر پذیری شامل تغییر پذیری یا واریانس فرآیند (γ^2) و تغییر پذیری حاصل از نمونه‌گیری که به آن واریانس نمونه‌ای می‌گویند و دراینجا با σ_S^2 نشان می‌دهیم، تشکیل شده است. بنابراین :

$$\begin{aligned}\sigma_T^2 &= \sigma_S^2 + \gamma^2 = E_{\lambda_t}\left[V(X_t|\lambda_t)\right] + V_{\lambda_t}\left[E(X_t|\lambda)\right] \\ &= E_{\lambda_t}(\lambda_t) + V_{\lambda_t}(\lambda_t) \\ &= \alpha\beta + \alpha\beta^2\end{aligned} \qquad (۷)$$

بطور مشابه، منطقی است که بخشی از میانگین پسین λ_t، تحت تاثیر نمونه‌گیری و بخشی دیگر ناشی از وضعیت فرآیند باشد، لذا می‌توان آنرا یک میانگین وزنی بصورت زیر در نظر گرفت:

$$\mu_{\lambda_t} = E(\lambda_t|x_t) = wx_t + (1-w)\mu \qquad (۸)$$

جهت تعیین مقدارمناسب برای w به این نکته توجه می‌شود که هر چقدر پراکندگی فرآیند یا نمونه‌گیری کمتر باشد باید اهمیت و وزن آن بیشتر درنظر گرفته شود، لذا نسبت زیرمناسب خواهد. بود.

$$w = \frac{\gamma^2}{\sigma_T^2} = \frac{\alpha\beta}{1+\beta} \qquad (۹)$$

با توجه به اینکه در همه موارد فوق پارامترهای α و β نقش اساسی دارند، برآورد آنها مهم خواهد بود.

۳ برآوردها

با توجه به روابط (۳) و (۷) جهت برآورد α و β می‌توان از میانگین توزیع پیشین (μ) و واریانس کل نمونه‌ای (σ_T^2) استفاده نمود، لذا ابتدا برآوردهای آنهاارائه می‌گردد.

۳−۱) برآوردμ

چون X_t یک برآورد نااریب برای μ است زیرا:

$$E(X_t) = E_{\lambda_t}\left[E(X_t|\lambda_t)\right] = E_{\lambda_t}(\lambda_t) = \alpha\beta = \mu$$

در نتیجه یک برآورد کننده مناسب و نااریب باواریانس کمتر برای μ را می‌توان $\bar{x_t}$ را در نظر گرفت:

$$\hat{\mu_T} = \bar{x_T} = \frac{1}{T}\sum_{t=1}^{T} x_t \qquad Var(\hat{\mu_T}) = \frac{\sigma_T^2}{T} \qquad (۱۰)$$

۳–۲) برآورد $\sigma_T^۲$

برآورد کننده زیر که می‌توان نشان داد یک برآورد کننده نااریب‌است، $\sigma_T^۲$ را برآورد می‌کند.

$$\hat{\sigma}_T^۲ = \frac{۱}{T-۱}[\sum_{t=۱}^{T}(x_t - \hat{\mu}_T)^۲] = \frac{۱}{T-۱}[\sum_{t=۱}^{T} x_t^۲ - T\hat{\mu}_T^۲] \qquad (۱۱)$$

۳–۳) برآورده‌ای α و β

بر اساس روابط (۳) و (۷) داریم:

$$\mu = \alpha\beta$$
$$\sigma_T^۲ = \alpha\beta + \alpha\beta^۲$$

در نتیجه :

$$\alpha = \frac{\mu^۲}{\sigma_T^۲ - \mu}$$
$$\beta = \frac{\sigma_T^۲ - \mu}{\mu}$$

و با قرار دادن برآوردهای μ و $\sigma_T^۲$ برآوردهای α و β بشرح زیر تعیین می‌گردد:

$$\hat{\alpha} = \frac{\hat{\mu}^۲}{\hat{\sigma}_T^۲ - \hat{\mu}}$$
$$\hat{\beta} = \frac{\hat{\sigma}_T^۲ - \hat{\mu}}{\hat{\mu}} \qquad (۱۲)$$

۳–۴) برآورد واریانس فرآیند

$$\gamma^۲ = \alpha\beta^۲ \rightarrow \hat{\gamma^۲} = \hat{\sigma_T^۲} - \hat{\mu_T} \qquad (۱۳)$$

با توجه به اینکه $\alpha > ۰$ و $\beta > ۰$ و $\gamma^۲ > ۰$ است، بدیهی است برآوردهای فوق وقتی معتبراست که:

$$\hat{\sigma}_T^۲ - \hat{\mu}_T > ۰ \rightarrow \hat{\sigma}_T^۲ > \hat{\mu}_T$$

۴ برآورد های میانگین متحرک وزنی نمایی (EWMA) و رابطه بازگشتی

در برآوردهای قبلی کلیه T مشاهده $x_T, \ldots, x_1$ دارای تاثیر گذاری واهمیت یکسان بوده‌است ، جهت واقعی تر نمودن و بهینه‌سازی بروزنمایی اطلاعات ، مناسب تراست که مشاهدات جدیدتر، اهمیت و وزن بیشتری داشته باشند، بدین منظوراز برآورد EWMA استفاده می‌گردد

$$\hat{\mu}_T = \frac{\sum_{t=0}^{T-1} r^t x_{T-t}}{\sum_{t=0}^{T-1} r^t} = \frac{(1-r)\sum_{t=0}^{T-1} r^t x_{T-t}}{1-r^T} \qquad (۱۴)$$

$$\hat{\sigma}_T^2 = \frac{\sum_{t=0}^{T-1} r^t (x_{T-t} - \hat{\mu}_T)^2}{\sum_{t=0}^{T-1} r^t} = \frac{(1-r)\sum_{t=0}^{T-1} r^t x_{T-t}}{1-r^T} - \hat{\mu}_T^2 \qquad (۱۵)$$

در محاسبات کامپیوتری بویژه در نمونه‌گیری با مشاهدات زیاد ، استفاده از روابط (۱۴) و (۱۵) زمان و حافظه و هزینه زیادی را شامل می‌گردد. جهت کاهش این موارد می‌توان روابط فوق را بصورت روابط بازگشتی زیر نوشت و ازآن استفاده نمود.

$$\hat{\mu}_T = \frac{1-r}{1-r^T} x_T + \frac{r-r^T}{1-r^T} \hat{\mu}_{T-1} \qquad (۱۶)$$

$$\hat{\sigma}_T^2 = \frac{1-r}{1-r^T} x_T^2 + \frac{r-r^T}{1-r^T} (\hat{\sigma}_{T-1}^2 + \hat{\mu}_{T-1}^2) - \hat{\mu}_T^2 \qquad (۱۷)$$

بنابراین برآورد μ در زمان T یک ترکیب خطی از مشاهده فعلی (x_T) و برآورد μ دریک دوره زمانی قبل و بطور مشابه برآورد σ_T^2 یک ترکیب خطی از x_T^2 و برآوردهای میانگین و واریانس دریک دوره قبل و میانگین دوره فعلی می‌باشد. درعمل برای اینکه مشاهدات جدید وزن بیشتری داشته‌باشند، $0/9 < r < 1$. درنظرگرفته می‌شود. در روابط فوق چون در هر دوره یک مشاهده به حجم نمونه‌اضافه می‌شود، می‌توان طول دوره را ثابت و بطول n درنظر گرفت دراین صورت روابط بالا بصورت زیر درمی آید.

$$\hat{\mu}_T = \frac{\sum_{t=0}^{n-1} r^t x_{T-t}}{\sum_{t=0}^{n-1} r^t} = \frac{1-r}{1-r^n} x_T + r\hat{\mu}_{T-1} + \frac{r^{n+1}-r^n}{1-r^n} x_{T-n} \qquad (۱۸)$$

$$\hat{\sigma}_T^2 = \frac{\sum_{t=0}^{n-1} r^t (x_{T-t} - \hat{\mu}_T)^2}{\sum_{t=0}^{n-1} r^t} = \frac{1-r}{1-r^n} x_T^2 + r(\hat{\sigma}_{T-1}^2 + \hat{\mu}_{T-1}^2) + \frac{r^{n+1}-r^n}{1-r^n} x_{T-n}^2 - \hat{\mu}_T^2 \qquad (۱۹)$$

۵ برآوردهای حدی

وقتی تعداد مشاهدات (nیاT) زیاد باشد، به عبارت دیگر به سمت بینهایت میل‌کند در اینصورت روابط (۱۶) و (۱۸) به رابطه (۲۰) و روابط (۱۷) و (۱۹) به رابطه (۲۱) گرایش می‌یابد.

$$\hat{\mu}_T = (1-r)x_T + r\hat{\mu}_{T-1} \qquad (20)$$

$$\hat{\sigma}_T^2 = (1-r)x_T^2 + r(\hat{\sigma}_{T-1}^2 + \hat{\mu}_{T-1}^2) - \hat{\mu}_T^2 \qquad (21)$$

بنابراین برآورد کننده‌های حدی یک ترکیب محدب از مشاهدات یا برآوردهای فعلی و یک دوره قبل می‌باشند.

۶ برآوردهای کوتاه مدت و بلند مدت

دربرآوردهای EWMA و بازگشتی، اگر r نزدیک به۱ باشد، ضرایب r^t دیرتر به صفر میل می‌کند و از مشاهدات دورتر و گذشته بیشتراستفاده می‌گردد، لذا به این حالت، برآورد کننده بلند مدت گفته می‌شود. واگر r نزدیک به۰/۹ باشد ضرایب r^t زودتر به صفر گرایش دارد و مشاهدات گذشته کمتر منظور می‌گردد و به این حالت برآوردکننده کوتاه مدت گفته می‌شود. مثال زیر یک نمونه از کاربردهای نتایج این تحقیق در بروزنمایی پارامتر λ_t جهت تجزیه وتحلیل آماری و تشکیل نمودارهای کنترل درفرایندهای تولیدی می باشد.

مثال: اگر در یک فرآیند تولید طی ۲۰ روز گذشته تعداد قطعات معیوب به ترتیب ۵۲، ۶۰، ۷۰، ۴۸، ۵۵، ۶۳، ۸۲، ۶۷، ۵۰، ۴۵، ۶۴، ۶۸، ۵۷، ۶۵، ۵۹، ۶۱، ۷۸، ۸۰، ۷۳، ۶۹ قطعه در روز باشد.براساس این مشاهدات ۲۰ روزه و مفروضات و روابط حاصله دراین تحقیق، برآورد پارامترها و مشخصه‌های کیفی فرآیند تولید برای روزهای بعدی براساس برآوردهای بلند مدت (r=۰/۹۹) و کوتاه مدت (r=۰/۹۰) و دوره ثابت بطول n=۲۰ (روابط (۱۸) و (۱۹)) بشرح زیراست.

روز	مشاهده	برآورد بلند مدت ($r = ۰/۹۹$)					
T	x_i	$\hat{\mu}_T$	$\hat{\sigma}^2_T$	$\hat{\alpha}$	$\hat{\beta}$	$\mu_{\hat{\lambda}_t}$	$\hat{\sigma}^2_{\lambda_t}$
۲۰	۶۹	۶۳/۳	۱۰۹	۸۷/۱	۰/۷۲	۶۵/۳	۲۷/۴
۲۱	۶۰	۶۳/۶	۱۰۶/۲	۹۴/۳	۰/۶۷	۶۱/۹	۲۴/۸
۲۲	۷۳	۶۴/۳	۱۰۶/۲	۹۸/۷	۰/۶۵	۶۷/۶	۲۶/۶
۲۳	۸۸	۶۵/۳	۱۳۹/۶	۵۷/۴	۱/۱۴	۷۷/۵	۴۱/۳
۲۴	۵۳	۶۵/۴	۱۳۳/۳	۶۳	۱/۰۴	۵۹/۱	۳۰/۱
۲۵	۹۷	۶۷/۶	۱۷۸	۴۱/۴	۱/۶۳	۸۵/۸	۵۳/۲

روز	مشاهده	برآورد کوتاه مدت ($r = ۰/۹۰$)					
T	x_i	$\hat{\mu}_T$	$\hat{\sigma}^2_T$	$\hat{\alpha}$	$\hat{\beta}$	$\mu_{\hat{\lambda}_t}$	$\hat{\sigma}^2_{\lambda_t}$
۲۰	۶۹	۶۳/۳	۱۰۹	۸۷/۱	۰/۷۲	۶۵/۳	۲۷/۴
۲۱	۶۰	۶۲/۹	۱۲۱	۶۸/۱	۰/۹۲	۶۱/۴	۲۹/۴
۲۲	۷۳	۶۴/۱	۱۱۸/۷	۷۵/۳	۰/۸۵	۶۸/۱	۳۱/۳
۲۳	۸۸	۶۶/۸	۱۵۷/۷	۴۹/۱	۱/۳۶	۷۹	۴۵/۵
۲۴	۵۳	۶۵/۵	۱۵۶/۱	۴۷/۴	۱/۳۸	۵۸/۲	۳۳/۸
۲۵	۹۷	۶۹/۲	۲۴۴	۲۷/۴	۲/۵۳	۸۹/۲	۶۳/۹

مراجع

Yousry, Strum, Feltz and Noorossana (1991), Process Monitoring in Real Time : Empirical Bayes approach-Discrete Case. Qual. Reliab. Engng.Int.,7,123-132.

Feltz and shiau (2001) statistical process monitoring using an Empirical Bayes Multivaraite process contorol chart, Qual. Reliab. Engng. Int.,17,119-124.

Weilu and Rudy (2001), Reliability Demonstraction Test for a Finite population; Qual. Reliab. Engng.Int.,17,33-38.

Martz and Waller (1989) Bayesian Reliabity Analysis, Krieger, New York, NY.

Jain, Alt and Grimsnaw (1993), Multivariate quality contorl-A Bayesian approach, ASQC Quality congress Transactions, Boston : 645-641.

تصمیم آماری بیزی بوسیله برآوردگرهای MLM در فضاهای غیر حقیقی

حسین نوری امامزاده، مهدی دوست پرست

$P۱۳۰۵۵$

گروه آمار، دانشگاه فردوسی مشهد

چکیده: زمانی که از برآوردکردن و برآورد پارامتری مجهول سخن گفته می‌شود، آنچه در اذهان تجلّی می‌یابد حقیقی بودن پارامتر مجهول است. امّا باید متذّکر شویم که در بسیاری از موارد پارامتر واقعی درفضای اقلیدسی نیست . دراین گونه مسائل معمولاً از دو برآوردگر ماکزیمم چگالی پسین (MAP) که بالاترین چگالی احتمال پسین را انتخاب می کند و مینیمم مربعات خطا $(MMSE)$ که متوسط مربع فاصله از مقادیر واقعی پارامتر را کمینه می‌کند، استفاده می‌شود.

در این مقاله برآوردگر جدیدی را با نام برآوردگر ماکزیمم جرم موضعی (MLM) که روی چگالی احتمال موضعی انتگرال می‌گیرد، معرفی کرده و ضمن مقایسهٔ آنها با یکدیگر، نشان می‌دهیم که در اکثر موارد، این برآوردگر کارآتر از برآوردگرهای MAP و $MMSE$ است و حتی از الگوریتم‌های استاندارد مورد استفاده، مناسب تراست. مثالهایی که ازفیزیک نور ارائه می‌شود صحت مطالب فوق را تأیید می‌کند .

واژه‌های کلیدی: برآوردگر ماکزیمم پسین، برآوردگر حداقل مربعات، برآوردگر ماکزیمم جرم موضعی، چگالی پسین، چگالی پیشین، الگوریتم.

۱ مقدمه

اغلب الگوریتمهای شهودی در چارچوب آمار بیزی بررسی می‌شوند. بطور کلی دو برآوردگر مورد استفاده است که عبارتند از: برآوردگر پسین ماکزیمم (MAP) و برآوردگر مینیمم میانگین مربعات خطا $(MMSE)$ ما نشان می‌دهیم که هیچکدام از این دو برآوردگر مناسب نیست چرا که برآوردگر MAP ازساختارتوزیع احتمال پسین استفاده نمی کندو برآوردگر $MMSE$ نوع خطا را منعکس نمی‌کند. در این مقاله برآوردگر جدیدی تحت عنوان ماکزیمم جرم موضعی (MLM) را معرفی می‌کنیم که با انتگرال گیری نسبت به چگالی احتمال موضعی بدست می آید. برآوردگر MLM نسبت به ساختار موضعی چگالی پسین حساس است در حالیکه MAP این گ‌ونه نیست. MLM محتمل ترین جواب به

طور تقریبی صحیح را می‌یابد و در حالتی که مشاهدات دارای اغتشاش کم هستند نیز تقریب کارایی فراهم می‌کند. این برآوردگر جدید را در یک مسئله فیزیک نور ($color\ constancy$) به کار می‌گیریم. فرض کنید یک پرتوی تابش مجهول به یک سطح مجهول تابیده می‌شود و خصوصیات پرتوی بازتابیده توسط سنسور ثبت می‌شود. می‌توانیم براساس اطلاعات بدست آمده از سنسور خصوصیات پرتوی تابش و سطح بازتابش را مشخص کنیم که این کار را با الگوریتم MLM انجام داده و نشان می‌دهیم مناسب تر از دو برآوردگر قبلی است.

در این مقاله، در بخش ۲ مسأله اساسی را مطرح کرده و معایب MAP و $MMSE$ را بیان می‌کنیم. بخش ۳ شامل ایده کلی جهت بدست آوردن برآوردگر مناسب است و همچنین با یک مثال ساده توزیعهای احتمالی پیشین و پسین و تابع زیان را معرفی می‌کنیم. دربخش ۴ برآورد MLM رابرای یک مفهوم عمیق تر شهودی یعنی $color\ constancy$ بیان و به کارمی گیریم. در بخش ۵ یک مقایسه کلی بین سایر الگوریتمهای رایج و الگوریتم MLM صورت گرفته است. دربخش ٦شامل نتایج بدست آمده است.

۲ مسأله کلی

هدف استنباط در مورد خصوصیات جهان طبیعی از قبیل: اشکال، رنگهاو... است که با توجه سنسورهای خاصی اندازه‌گیری می‌شوند. روش کلی در برخورد با این مسأله آنالیز بیزی است به طوریکه داده‌های شهودی با توزیع احتمالات پیشین برای یافتن توزیع احتمال پسین خواص طبیعت ترکیب می‌شود که آن را پارامترهای منظری ($sceneparameters$) می‌نامیم. نوعاً، هدف انتخاب بهترین برآورد پارامتر منظری است. برای این منظور یک معیار بهینگی انتخاب شده و از توزیع احتمال پیشین برای یافتن برآورد بهینه استفاده می‌کنیم. دو قاعده تصمیم که بطور عمومی مورد استفاده‌اند MAP و $MMSE$ هستند.

قاعده MAP مقادیری از پارامترهای منظری را انتخاب می‌کند که چگالی احتمال پسین را ماکزیمم می‌کند. این برآوردگر رابطه تنگاتنگی با روش درستنمایی ماکزیمم دارد و در بسیاری از الگوریتمها از آن استفاده می‌شود.

قاعده $MMSE$ مقادیری از پارامترهای منظری را انتخاب می‌کند که میانگین مربعات فاصله از مقادیر واقعی پارامتر منظری را مینیمم می‌کند.به سادگی می‌توان نشان داد که میانگین توزیع پسین، $MMSE$ است. همچنین این برآوردگر بطور وسیعی مورد استفاده قرار گرفته است.

هنگامی که چگالی احتمال پسین فضای پارامتر منظری را بخوبی مشخص کند قاعده MAP و $MMSE$ تقریباً جواب یکسانی می‌دهند و در صورتیکه این چگالی کمی ساده

تر باشد این دو برآوردگر رضایت بخش نیستند. برآوردگر MAP به جزئیات ساختار چگالی پسین حساس نیست زیرا در آن، نقطه ماکزیمم چگالی احتمال مد نظر است. برآوردگر $MMSE$ ممکن است جوابی خارج از فضای پارامتر ارائه دهد. بعلاوه در محاسبات این برآوردگر ممکن است نیاز به حل یک انتگرال روی فضای پارامتر پیدا کنیم.

۳ ایده

جهت مقایسه MAP و $MMSE$ و بدست آوردن برآوردگر جدید، تابع زیان در این فضا معرفی می‌کنیم. برای یک پارامتر منظری θ تابع زیان را با $L(\theta,\hat{\theta})$ نشان داده و آن را زیان حاصل از برآورد $\hat{\theta}$ وقتی که پارامتر واقعی θ باشد، می‌نامیم. تابع زیان منجر به روشی برای برآورد پارامترهای منظری از روی توزیع پسین می‌شود بگونه‌ای که پارامترهایی که متوسط زیان را مینیمم می‌کنند را انتخاب می‌کند. قاعده MAP تابع زیان را بصورت $L(\theta,\hat{\theta}) = -\delta(\theta-\hat{\theta})$ تعریف می‌کند که ما آن را $-\delta$ می‌نامند. قاعده $MMSE$ تابع زیان را بصورت $L(\theta,\hat{\theta}) = (\theta-\hat{\theta})^2$ در نظر می‌گیرد.
معتقدیم، هیچکدام از این توابع زیان، تقریب مناسبی برای زیان حاصل از برآورد نادرست نمی‌دهد. تابع زیان $-\delta$ اشاره دارد که خطاهای کوچک هم به اندازه خطاهای بزرگ زیان حاصل می‌کنند. تابع زیان مربع خطا بیان می‌کند که باافزایش اندازه خطای برآورد میزان زیان حاصل بشدت افزایش می‌یابد لذا به دنبال تابع زیانی هستیم که معایب توابع زیان فوق را نداشته باشد یعنی خطاهای کوچک به اندازه خطاهای بزرگ نباشد و افزایش خطا، زیان حاصل را به شدت و غیر معقول افزایش ندهد زیرا در طبیعت نیز با مسائلی روبرو هستیم که خطاهای کوچک و بزرگ زیانهای متفاوتی به ما می‌رساند مثلاً در یک مسابقه فوتبال و در یک ضربه پنالتی با یک خطای کم ممکن است موفق به دفع توپ شویم اما با خطای زیاد ممکن است حتی جهت توپ را تشخیص ندهیم.
ما یک تابع زیان معرفی می‌کنیم که برای خیلی از مسائل شهودی نیز مناسب باشد این تابع زیان با انتگرال گیری روی چگالی احتمال برای یافتن جرم احتمال موضعی در یک ناحیه مشخص می‌شود که ما آنرا تابع زیان جرم موضعی نامیده و برآوردگر حاصل را برآوردگر ماکزیمم جرم موضعی (MLM) می‌نامیم.

۱.۳ مثال

فرض کنیدیک عدد (y) مشاهده شده که حاصل ضرب دو عدد دیگراست یعنی $y=ab$.
هدف تعیین مقادیر a و b است. درارتباط باداده های بصری (شهودی) y می توانبه

عنوان یک تصویرحاصل ازدوعامل دیگر(یاپارامترمنظری)درنظرگرفته شود. ازنظر ریاضی این یک مسأله غیرخطی وباجواب نامتناهی است.
می توان فرض کرد $y = ۱$ و $۰ < a, b < ۴$ برقراراست.اگربخواهیم‌به تنهایی از معلومات هندسی استفاده کنیم فقط می توان گفت که جواب در امتداد یک جزیی از هذلولی $ab - ۱ = ۰$ قرارمی گیرد. یک آنالیزاحتمالی جواب دقیق تری ارایه می‌دهد. می‌خواهیم توزیع احتمالی پسین $p(\theta|y)$ رابیابیم که $\theta = (a \quad b)^T$. با استفاده از قضیه بیز می‌توان نوشت :

$$P(\theta|y) = \frac{P(y|\theta)\Pi(\theta)}{P_y(y)} \qquad (۱)$$

ثابت نرمال ساز $P_Y(y)$ مستقل از پارامتر θ است که به دنبال برآورد آن هستیم. $\Pi(\theta)$ توزیع پیشین پارامتر منظری θ است و $p(y|\theta)$ تابع درستنمایی است. اجازه می‌دهیم مشاهدات همراه بااغتشاش باشد، لذا مدل احتمالی است. تابع تحویلی $f(\theta)$ مشاهدات اختصاصی پارامترهای منظری θ بدون اغتشاش را بیان می‌کند. در این مثال تابع تحویلی عبارت است از

$$f(\theta) = a.b = \theta_۱\theta_۲$$

فرض می‌کنیم اغتشاش مشاهدات دارای توزیع نرمال با میانگین ۰ و واریانس $\sigma^۲$ باشد. پس تابع درستنمایی به صورت

$$P(y|\theta) = \frac{۱}{(\sqrt{۲\pi\sigma^۲})^N} e^{\frac{-||y-f(\theta)||^۲}{۲\sigma^۲}} \qquad (۲)$$

چون $۰ < \theta_۱, \theta_۲ < ۴$ توزیع پیشین را بصورت

$$P_\theta(\theta) = \frac{۱}{۱۶} \qquad (۳)$$

روی ناحیه $[۰, ۴] \times [۰, ۴]$ و برابر ۰ در سایر نقاط، تعریف می کنیم. با جایگذاری (۲), (۳) در (۱) داریم

$$P(\theta|y = ۱) = \begin{cases} Ce^{\frac{-(۱-ab)^۲}{۲\sigma^۲}} & ۰ < a, b < ۴ \\ ۰ & \text{درغیر اینصورت} \end{cases}$$

که درآن C یک ثابت است.درشکل (۱) این توزیع پسین نمایش داده شده‌است.نقاط درطول هذلولی $ab = ۱$ تشکیل یک مرزبابالاترین احتمال می دهند چون مشاهدات دارای اغتشاش هستند جفت پارامتردارای احتمال ناصفراند. باتوجه به شکل $۱(b)$ توجه می کنیم که هرچند همه نقاط مرز دارای ارتفاع یکسانی هستند اما نزدیک نقطه $(۱, ۱)$

عریض تراز سایرنقاط درامتداد هذلولی است.
برآورد بهینه θ بامی نیمم کردن متوسط زیان $R(\hat{\theta}|y)$ به دست می آیدکه آن راریسک بیزی می نامیم. داریم

$$R(\hat{\theta}|y) = \int L(\theta, \hat{\theta}) P(\theta|y) d\theta \qquad (۴)$$

درحالت خاصی که $L(\theta, \hat{\theta}) = L(\theta - \hat{\theta})$ ریسک بیزی بصورت

$$R(\hat{\theta}|y) = \int L(\theta - \hat{\theta}) P(\theta|y) d\theta = L(-\hat{\theta})$$

می‌باشد.
شکل ۲ زیان حاصل شده باقاعده MAP و ریسک بیزی آن را برای مسأله $y = ab$ با توزیع پیشین یکنواخت را بدست می دهد. برای برآوردگر MAP ریسک بیزی قرینه تابع احتمال پسین است. جهت واضح نشان دادن و مقایسه باشکل (۱) آن رابه طرف بالا رسم کرده‌ایم. هرنقطه واقع بر مرز هذلولی دارای چگالی احتمال یکسان است. پس قاعده MAP برآورد منحصربه فردی نمی دهد. همچنین این قاعده ازهمه تغییرات درعرض چشم پوشی می کندکه می تواند منبع معنی داری از اطلاعات باشد.
شکل ۳ تابع زیان برآوردگر $MMSE$ و ریسک بیزی آن رانشان می دهد. توجه داریم که برآورد $MMSE$ حاصل برابر $a = b = ۱.۳$ است که نتیجه عجیبی است. چرا که $a \times b \neq ۱$
تابع زیانی را که برای قاعده MLM تعریف می‌کنیم نقص دو برآوردگر MAP و $MMSE$ را برطرف می‌کند.
تابع زیان جرم موضعی را با علامت منفی به فرم نرمال با همبستگی کم بصورت

$$L(\theta, \hat{\theta}) = -exp\{-|K_L^{-\frac{۱}{۲}}(\theta - \hat{\theta})|^۲\}$$

تعریف می کنیم که درآن :

$$|K^{-\frac{۱}{۲}}\theta|^۲ = \theta^T K^{-۱}\theta$$

چون مقادیر ویژه ماتریس K_L به قدرکافی کوچک هستند. این تابع زیان تقریباً جواب صحیح می دهد.(شکل ۴). همچنین شکل ۴ متوسط زیان به روش MLM برای مسأله $ab = ۱$ را نشان می‌دهد که برآورد MLM حاصل به صورت $(۱, ۱)$ است. چنانچه مشاهده می شوداین برآورد عیب های دوبرآوردگر قبلی راندارد.

۲.۳ نتیجه

این مثال ساده چندین مطلب رانشان می دهد.
—تابع زیان دارای اهمیت زیادی است که می تواند برآورد بهینه راتغییردهد.
—برآورد MAP ازعرض های مرز ماکزیمم درستنمایی چشم پوشی می کند.
—برآورد $MMSE$ که ازنظر میانگین مربعات خطا، بهینه است جواب نادرست می دهد. $(۱.۳ \times ۱.۳ \neq ۱)$
—برآورد MLM ساختار جرم احتمال موضعی رابه کارمی گیرد و برآورد بهینه راارائه می دهد. و همانطورکه مشاهده شد از دو برآورد قبلی بهتربود.

۴ کاربرد در $color\ constancy$

دربرخورد پرتو تابش به برخی سطوح پرتوی بازتابش تلفیقی از رنگها است اما در برخی سطوح دیگر فقط شکست نورحاصل می شود. در اینجا هدف برآورد مشخصه‌های نور تابیده شده از روی اطلاعاتی است که سنسور از پرتو بازتابش می‌دهد، می‌باشد. به این معنی که نه پرتو تابش و نه سطح بازتابش معلوم هستند.

فرض کنیم مجموعه‌ای شامل N_L سطح داریم. قابلیت بازتاب j اُمین سطح را با یک بردارستونی S_j نشان می‌دهیم. مؤلفه های این بردار ستونی نشان دهندهٔ شکست نور مشاهده شدهٔ، منعکس شده در N_λ نوار طیف نوری است. به طور مشابه قدرت توریع پرتو تابش $S.P.D$ را با بردار ستونی e نمایش می‌دهیم. $S.P.D$ نور رسیده از دستگاه c_j است که

$$c_j = diag(e) \times s_j$$

درسیستم شهودی ازپرتو بازتابش c_j با N_r کلاس ازفوتو سنسورهای خطی نمونه گیری می شود. حساسیت های فوتوسنسور را باماتریس R مشخص می کنیم. pq اُمین درایهٔ R نشان دهندهٔ حساسیت p اُمین سنسوربه نوردر q اُمین طول موج نواراست. بردار r_j نشان دهندهٔ پاسخ هریک از کلاس های سنسوربه پرتوبازتابیده c_j است. صرفنظرازخطای سنسور معادلهٔ تحویلی r_j به j اُمین سطح رنگی به عنوان تابعی ازپارامترهای منظری، e و s_j بصورت

$$r_j = f(e_j, s_j) = R\ c_j = R\ diag(e) s_j \qquad (٦)$$

بااین نمادها، هدف یافتن پرتوتابش e و سطح بازتابش s_j ازپاسخ r_j برای هرسطح است.

۱.۴ تعاریف و نماد

تعریف ۱
گوییم یک پرتو تابش دریک مدل خطی N_e بعدی B_e صدق می کند اگربتوان نوشت: $e = B_e W_e$ که درآن B_e یک ماتریس با N_λ سطر و N_e ستون و W_e یک بردار ستونی N_e بعدی است. مؤلفه های W_e راوزن های مدل خطی برای پرتو e گوییم.
تعریف ۲
گوییم یک سطح بازتابش دریک مدل خطی B_s صدق می کند اگربتوان نوشت: $s_j = B_s W_{sj}$ که درآن B_s یک ماتریس با $N - \lambda$ سطرو N_s ستون و W_{sj} یک بردار ستونی N_s بعدی است. CIE یک مدل خطی ۳ بعدی رابرای نور خورشید استاندارد کرده است. $L.T.Maloney$ نشان داده است که درمدل های خطی بابعدهای کم (۳بعدی) درصد زیادی ازخطای اندازه‌گیری سطح بازتابش کاهش می یابد. دراین نمایش معادلهٔ تحویلی (٦) به صورت:

$$r_j = f(W_e, W_{sj}) = R\ diag(B_e W_e) B_s W_{sj} = R\ diag(B_s W_{sj}) B_e We \qquad (۷)$$

۲.۴ الگوریتم ها

پیشرفته‌ترین الگوریتم‌ها در فیزیک نور تحت شرط خطی بودن مدل هستند. $Maloney$ و $Wandell$ نشان دادند که وقتی N_r کلاس ازاشعه بازتابش داریم ممکن است معادله تحویلی معکوس گردد اگراشعه تابش در یک مدل خطی N_r بعدی وسطح بازتابش دریک مدل خطی $N_{r-۱}$ بعدی قرارگیرند. از دیدگاه شهودی $r = ۳$ است. پس نتایج حاصل شده از آنالیز $Wandell$ و $Maloney$ معتبرند اگر سطح بازتابش در یک مدل دوبعدی قرارگیرند که از نظر طبیعی چنین حالتی رخ نمی‌دهد. در الگوریتم $Buchsbaum$ اشعه تابش و سطح بازتابش دریک مدل خطی N_r بعدی قرار می‌گیرند. در این الگوریتم فرض می‌شود که میانگین فاصله‌ای سطح بازتابش بازای همه مقادیرثابت است. این فرض را $gray\ world$ گوییم. که نتیجه می‌دهد که $color\ constancy$ برای تصاویری که میانگین بازتابش آنها از میانگین فرض شده متفاوت باشد بایستی تفکیک شود. در اینجا $color\ constancy$ را به عنوان یک مسأله برآورد آماری مطالعه می‌کنیم. برای این منظور $Vhrel$ و $Trussell$ از روش ماکزیمم درستنمایی که ارتباط نزدیکی با برآورد MAP دارد استفاده کرده‌اند. $Freeman$ و $Brainnad$ محک $MMSE$ رابرای برآورد اشعه تابش به خدمت گرفته‌اند. در اینجا روش MLM رابه کار گرفته و مزیت آن را به روشهای

فوق بیان می‌کنیم. برای تعیین الگوریتم بایستی ابتدا تابع احتمال پیشین را انتخاب کرده و سپس امیدریاضی تابع زیان را می‌نیمم کنیم.

۳.۴ تعیین تابع احتمال پیشین

یک مدل خطی ۳ بعدی برای سطح بازتابش و پرتو تابش به دست می آوریم. برای به دست آوردن چنین مدلی بر روی داده های $K.L.Kelly$ تکنیک تجزیه وتحلیل مؤلفه‌های اصلی را به کار می گیریم. سپس بهترین مدل برازش وزنی‌رابرای‌هرسطح می یابیم. فرض می کنیم مدل خطی ۳ بعدی برای پرتوی تابش همان مدل CIE برای نور خورشیدباشد. برای تولید توزیع پیشین پرتوهای خورشید به رنگهای به هم مرتبط که دمایشان مطابق باتوزیع نرمال بامیانگین $۶۵۰۰^{\circ}K$ وانحراف استاندارد $۴۰۰۰^{\circ}K$ است تابیده می‌شود. W_e در هر تابع محاسبه کرده و نتیجه را با توزیع نرمال مقایسه می‌کنیم. توجه داریم که برای صحیح بودن نتایج نیازی به نرمال بودن توزیع پیشین نیست.

۴.۴ تقریبی برای جرم احتمال موضعی

می‌خواهیم برآوردگر MLM را برای مسألهٔ $color\ constancy$ بکارگیریم. ریسک بیزی که در جستجوی آن برای مینیمم کردن هستیم با انتگرال گیری روی همهٔ فضای پارامتر منظری معادله(۴) بدست می آید.با ترکیب روابط(۱) ،(۲) و (۴) داریم

$$R(\hat{\theta}|y) = -C \int exp\{-\frac{\tau}{۲}|K_n^{-\frac{۱}{۲}}(r - f(x))|^۲\} P_\theta(\theta) exp\{-\frac{\mu}{۲}(|K_L^{-\frac{۱}{۲}}(\theta - \hat{\theta})|^۲)\} d\theta \quad (۸)$$

که در این رابطه $\frac{۱}{\tau}K_n$ کواریانس اغتشاش مشاهدات است که K_n ثابت فرض شده و τ زیاد می‌شود و بطور مشابه $\frac{۱}{\mu}K_L$ کواریانس تابع زیان موضعی است که K_L به شکل تابع زیان بستگی دارد و μ بزرگ می‌شود. در اینجا بااین فرض مطالعه انجام می‌شود که $\frac{\mu}{\tau}$ کوچک باشد.

قضیه

انتگرال

$$I(\tau) = \int exp\{-\tau\phi(\theta)\} g(\theta) d\theta \quad (۹)$$

بطور تقریبی و با افزایش τ برابر است با

$$I(\tau) \approx \frac{e^{-\tau\phi(\theta_۰}}{\sqrt{|det(\phi_{\theta_i\theta_j}(\theta_۰))|}} = (\frac{۲\pi}{\tau})^{\frac{n}{۲}} g(\theta_۰) \quad (۱۰)$$

که در آن θ_0 مینیمم $\phi(\theta)$ است.
با جایگذاری $g(\theta) = P(\theta)$ و

$$\phi(\theta) = \frac{1}{2}|K_n^{-\frac{1}{2}}(y - f(\theta))|^2 + \frac{\mu}{2\tau}|K_L^{-\frac{1}{2}}(\theta - \hat{\theta})|^2 \qquad (11)$$

در قضیه فوق داریم

$$R(\hat{\theta}|y) \approx -Ce^{\{-\tau(\frac{1}{2}|K_n^{-\frac{1}{2}}(y-f(\theta_0))|^2+\frac{\mu}{2\tau}|K_L^{-\frac{1}{2}}(\theta_0-\hat{\theta})|^2)\}} \times$$

$$\frac{P_\theta(\theta_0)}{\sqrt{|det(\phi_{\theta_i\theta_j}(\theta_0))|}} \qquad (12)$$

با دوبار مشتق گیری از طرفین رابطه (۱۱) داریم

$$\phi_{\theta_i\theta_j}(\theta_0) = f_i^{'T} K_n^{-1} f_j' - (y - f(\theta_0))^T K_n^{-1} f_{ij}'' + \frac{\mu}{\tau}\{K_L^{-1}\}_{ij} \qquad (13)$$

که در آن

$$f_i' = \frac{\partial f(\theta)}{\partial \theta_i}|_{\theta=\theta_0}$$

$$f_{ij}'' = \frac{\partial^2 f(\theta)}{\partial \theta_i \theta_j}|_{\theta=\theta_0} \qquad (14)$$

برای بکارگیری رابطه (۱۲) نیازمند به یافتن یک نقطه θ_0 هستیم که $\phi(\theta)$ معادله (۱۱)را می نیمم کند.اگر تنها توجه خود را به محاسبه ریسک درنقطه ماکزیمم موضعی معطوف کنیم در این صورت هردو جمله $\phi(\theta)$ با انتخاب $\hat{\theta} = \theta_0$ بطور موضعی مینیمم شده‌اند. بنابراین با جایگذاری $\hat{\theta} = \theta_0$ در معادله (۱۲) برای محاسبه ریسک در نقطه ماکزیمم درستنمایی برآورد تقریباً خوبی خواهیم داشت.
در $\hat{\theta} = \theta_0$ تفاوت بین ریسک $R(\hat{\theta}|y)$ در معادله (۱۲) و توزیع پسین معادله (۱) عبارت

$$\frac{1}{\sqrt{|det(\phi_{\theta_i\theta_j}(\hat{\theta})|}}$$

است و این نشان می‌دهد که ریسک MLM به‌همان اندازه که به ارتفاع احتمال بستگی دارد به همان اندازه به عرض هم بستگی دارد.

۵ مقایسه الگوریتمها

چهار نمودار شکل ۵ عملکرد ۴ الگوریتم بازای یک پرتوی تابش را نشان می‌دهد. در هر نمودار پرتوی تابش واقعی باخط تیره و برآورد حاصل با خطهای نقطه چین بطور جداگانه رسم شده‌اند. برای رسم این برآوردها ابتدا مجموعه سطوحی از توزیع پیشین رسم شده و بازای آن پاسخ سنسور محاسبه گردیده است سپس به اندازه یک درصد، اغتشاش به پاسخ سنسور افزوده شده و هر الگوریتم به کار گرفته شده است. همانگونه که در شکل (۵ الف) دیده می‌شود برآوردهای حاصل ازالگوریتم MLM بسیار نزدیک مقدار واقعی (خط تیره) قرار دارند. برآوردهای حاصل ازالگوریتم $gray\ world$ نسبتاً صحیح است اما دارای انحرافات زیادی است(شکل ۵ ب). برآورد MAP همچنان که انتظار می‌رفت از اطلاعات مرتبط با توزیع پیشین صرف نظر کرده است (شکل ۵ ج). یکی دیگر از الگوریتمهایی که مورد استفاده قرار می‌گیرد، الگوریتم $subspace$ است. همانطور که در شکل (۵ د) می‌بینیم هیچ تضمینی برای صحیح بودن جواب حاصل از این الگوریتم وجود ندارد چون برآورد بسیار مغشوش است.

به علاوه عملکرد الگوریتمها تحت برخی خطاهای فرض شده در توزیع پیشین نیز بررسی شده‌اند. برای این منظور پرتوی تابش را در حالیکه سطح باز تابش، پرتوی باز تابش را منعکس می‌کند تغییر می‌دهیم در اینجا از پرتو تابش ، نور خورشید $۴۰۰۰^{\circ}K$ و $۱۰۰۰۰^{\circ}K$ همراه با نور خورشید $۶۵۰۰^{\circ}K$ استفاده کرده‌ایم. نتایج با محاسبه ریشهٔ میانگین مربع خطا $(RMSE)$ بین پرتوی تابش و برآورد حاصل به دست آمده اند. شکل (۶ الف) این خطا را برای چهار الگوریتم فوق و سه پرتوی تابش نشان می دهد. برای هر سه نوع پرتو الگوریتم MLM ازسایر الگوریتم ها بهترعمل می کنند و برآورد حاصل ازآن مستقل ازنوع پرتواست.

درقسمت بعد،ازنورخورشید $۶۵۰۰^{\circ}K$ استفاده شده ولی درانتخاب سطوح اریبی وجود دارد،این اریبی ممکن استفاده به دو صورت باشد: یکی، میانگین سطوح راطوری اریب کردیم که میانگین پاسخ های سنسور تحت اشعهٔ $۶۵۰۰^{\circ}K$ همان پاسخ های سنسور تحت اشعهٔ $۴۰۰۰^{\circ}K$ وبدون اریبی سطح باشد.دیگری، میانگین سطوح را طوری اریب کردیم که میانگین پاسخ های سنسورتحت اشعهٔ $۶۵۰۰^{\circ}K$ همان پاسخ های سنسورتحت اشعهٔ $۱۰۰۰۰^{\circ}K$ وبدون اریبی سطح باشد. بااین فرآیند قصد داریم تا پایداری الگوریتم ها را تحت تخطی از شرایط $gray\ world$ آزمون کنیم.شکل (۶ ب) RMSE را برای حالتی که سطوح نااریبند با دوگونه اریبی نشان می دهد.اریبی سطوح در عملکرد الگوریتم ها تأثیر دارد.چنانکه از شکل پیداست الگوریتم MLM به مراتب نسبت به سایرالگوریتم ها نسبت به تخطی از فرضیات $gray\ world$ پایدارتر است.

٦ نتایج

مااین مسأله را از دیدگاه تئوری تصمیم بیزی بررسی کردیم. ثابت کردیم هیچ کدام از دو برآوردگرMAP و $MMSE$ که معمولاً مورد استفاده قرار می گیرند مناسب نیست. برای از بین بردن نقایص این دو برآوردگر یک برآوردگر جدید به نام MLM معرفی کردیم. این برآوردگر یک انتگرال را روی چگالی احتمال پسین موضعی بیشینه می‌کند. تابع زیان این برآوردگر متناسب با نوع مسأله‌ای که با آن روبرو هستیم در نظر گرفته می‌شود و محتمل‌ترین برآورد تقریباً درست را ارائه می‌دهد. با یک مثال نشان دادیم که برآوردگر MLM یک برآورد مناسب (بهینه) به دست می‌آورد در حالیکه برآوردگرهای MAP و $MMSE$ این گونه نبودند. همچنین الگوریتم MLM رابا برآورد MAP و دو الگوریتم استانداردبرای *color constancy* یعنی *subspace* و *gray world* مقایسه کرده ونشان دادیم که MLM به طورمعنی داری حتی تحت تخطی از شرایط مفروض، بهتر از سایرین عمل می‌کند. از اینرو معتقدیم که برآوردگر معرفی شده باعث بهبودی عملکرد الگوریتم‌های رایج می‌گردد.

مراجع

[1] D.H.Brainard and W.T.Freeman (1994), *Bayesian method for recovering surface and illuminant properties from photosensor reponses*, In proc. SPIE,volume 2179.

[2] D.H.Brainard and B.A.Wandelland E.J.Chichilnisky (1993), *color constancy* ,Current dir.in psychl. science.

[3] R.A.Fisher. (1959), *statistical methods and scientific inference*, Hafner.

[4] K.L.Kelly,K.S.Gibson,and D.Nickerson. (1943), *Tristimulus specification of the munsel of color from spectrophotometric measurements*, J.Op.Soc.Am.

[5] L.T.Maloney (1986), *Evaluation of linear models of surface spectrul reflectance with small numbers of parametrs*, J.Op.Soc.Am.

تعیین مدل خوشه‌بندی احتمالاتی براساس معیار اطلاع بیزی

محمد قاسم وحیدی‌اصل[۱]، محسن محمدزاده[۲]، محمد قربانی[۳]

$P۱۱۱۹۲$

۱ گروه آمار، دانشگاه شهید بهشتی
۲ گروه آمار، دانشگاه تربیت مدرس
۳ گروه آمار، دانشگاه تبریز

چکیده: یکی از مسائل مهم در تحلیل داده‌های چند متغیره، پیدا کردن ارتباط بین آنهاست. ساده‌ترین روش کشف رابطه موجود بین داده‌ها، رسم نمودار پراکنش است. در بسیاری از مفاهیم پزشکی، ژنتیکی و غیره، یکی از مسائل مهم، خوشه‌بندی داده‌ها به گروههای همگن است. روشهای ابتکاری مختلفی از جمله، روشهای سلسله مراتبی با ماکسیمم کردن تشابهات درون‌گروهی، داده‌ها را به خوشه‌هایی افراز می‌کنند. بدلیل وابسته بودن این روشها به تعریف فاصله بین دو خوشه و همچنین انتخاب یک مقدار آستانه برای تعیین تعداد خوشه‌ها، محققان در انتخاب بهترین معیاری که تشابهات درون گروهی را ماکسیمم نماید با مشکل مواجه می‌باشند. در این مقاله روش « انتخاب مدل برای خوشه‌بندی احتمالاتی با استفاده از معیار اطلاع بیزی[۱] (BIC) » مورد بررسی قرار می‌گیرد که در آن، فرضهای مختلف برای معیار تشابه نقشی ندارند و با تجزیه طیفی ماتریس کوواریانس می‌توان معیارهای ساده‌ای برای مشخص کردن حجم، شکل و جهت خوشه‌ها به دست آورد. بعلاوه با استفاده از معیار BIC می‌توان بهترین مدل خوشه‌بندی را انتخاب نمود.

کلید واژه: خوشه‌بندی کردن، مدل‌های آمیخته، الگوریتم EM، تجزیه طیفی، معیار BIC.

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 بطوریکه 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» باهم تفاوت دارند. در روش «خوشه‌بندی با اتصال منفرد»، فاصله بین نزدیکترین عضوهای دو گروه

[۱] Baysian Information Criterion

بعنوان فاصله بین دو خوشه تعریف می‌شود، در روش «خوشه بندی با اتصال کامل»، فاصله بین دورترین زوجها و در روش اتصال میانگین، متوسط فاصله بین تمام زوجهای موجود در داخل دو خوشه به عنوان معیار خوشه‌بندی مورد استفاده واقع می‌شود (هارتیگان، ۱۹۷۵). چون این روشها براساس مدل خاصی نمی‌باشند، استنباط از نمونه به جامعه امکان‌پذیر نیست و تعداد خوشه‌ها نیز به صورت ابتکاری با تعریف آستانه‌ای دلخواه تعیین می‌گردد. بنابراین لازم است روش خوشه‌بندی حتی الامکان مبتنی بر سلیقه محقق نبوده و براساس مدل یا توزیع احتمالی باشد تا بتوان در مورد آن استنباط آماری انجام داد. در این صورت شرایطی فراهم خواهد شد که بتوان با تجزیه طیفی ماتریس کوواریانس عناصر داخل خوشه‌ها، شکل، حجم و جهت خوشه‌ها را مشخص نمود. عمدتاً مجموعه مشاهدات تحت بررسی، همگی از یک جامعهٔ خاص نمی‌باشند، برای تشخیص این که هر مشاهده از کدام جامعه آمده است، منطقی است فرض شود که هر مشاهده براساس ویژگی‌ها و خصوصیاتش دارای توزیع احتمال خاصی است. بنابراین جامعه‌ای مرکب از چند زیرجامعه، دارای توزیع احتمالی آمیخته از توزیع‌های احتمال زیر جامعه‌ها خواهد بود، که در حالت کلی بشکل

$$f(x|\psi) = \lambda_1 f_1(x|\theta_1) + \ldots + \lambda_g f(x|\theta_g)$$

است، که در آن برای $j = 1, \ldots, g$، $f_j(.)$ تابع چگالی مولفه‌ها، $0 < \lambda_j \leq 1$ و $\sum_{j=1}^{g} \lambda_j = 1$، $\lambda = (\lambda_1, \ldots, \lambda_g)$، $\theta = (\theta_1, \ldots, \theta_g)$ و $\psi = (\lambda, \theta)$. از توابع چگالی آمیخته معروف که خیلی کاربرد دارد می‌توان به

$$f(x|\psi) = \lambda\phi(x|\mu_1, \sigma_1^2) + (1-\lambda)\phi(x|\mu_2, \sigma_2^2)$$
$$\phi(x|\mu_j, \sigma_j^2) = N(\mu_j, \sigma_j^2) \quad j = 1, 2$$

اشاره نمود. بنابراین خوشه‌بندی را می‌توان معادل با تفکیک توزیع آمیخته به مولفه‌های ساده دانست. قبل از بیان خوشه‌بندی براساس مدل می‌بایست یک سری فرضیات اساسی روی توزیع مولفه‌ها داشته باشیم. به عنوان مثال، یکی از این فرضها این است که توزیع هر مولفه تک مدی است. بنابراین هدف از خوشه‌بندی، تجزیه مولفه‌های چند بعدی مبهم و آمیخته به مولفه‌های ساده تک مدی است. به همین منظور در این مقاله برای خوشه‌بندی مشاهدات از مدل‌های آمیخته استفاده می‌شود تا بتوان تمام خصوصیات مشاهدات را در خوشه‌بندی کردن بکار گرفت.

۱.۱ تعیین معیارهای خوشه‌بندی براساس مدل

فرض کنید $X_1, \ldots, X_n$ یک نمونه تصادفی از جامعه Π با زیر جامعه‌های $\Pi_1, \ldots, \Pi_g$، 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،

$$f(x;\psi) = \sum_{j=1}^{g} \lambda_j \phi(x|\mu_j, \Sigma_j) \qquad (1)$$

که در آن $\theta_j = (\mu_j, \Sigma_j)$، $\psi = (\lambda_1, \ldots, \lambda_g, \theta_1, \ldots, \theta_g)$، $\sum_{j=1}^{g} \lambda_j = 1$ و $\lambda_j > 0$ است و تابع درستنمایی نمونه تصادفی به صورت

$$L(\psi|x) = \prod_{i=1}^{n} \sum_{j=1}^{g} \lambda_j f_j(x|\theta_j). \qquad (2)$$

خواهد بود. برای $C_j = \{i, X_i \in \Pi_j\}$ معیار خوشه‌بندی حاصل از ماکسیمم کردن تابع درستنمایی

$$L(x|C) = \prod_{j=1}^{g} \lambda_j^{n_j} \prod_{X_i \in C_j} f(x_i|\theta_j) \qquad (3)$$

معادل با معیار حاصل از ماکسیمم کردن (۲) خواهد بود (مک‌لن، ۱۹۸۲ و فرالی و رافتری، ۱۹۹۸). معمولاً مولفه اصلی X_iها معلوم نیستند و برای مشخص کردن مولفه اصلی X_i، متغیرهای گروه‌بندی Z_{ij} به صورت زیر تعریف می‌شود

$$Z_{ij} = \begin{cases} 1 & X_i \in \Pi_j \\ 0 & X_i \notin \Pi_j \end{cases}$$

براساس مشخصه‌های گروه‌بندی لگاریتم تابع درستنمایی را می‌توان به صورت

$$\log L(\psi|x, Z) = \sum_{j=1}^{g} \sum_{i=1}^{n} z_{ij} \log \{\lambda_j f(x_i/\theta_j)\}. \qquad (4)$$

نوشت. اگر z_{ij}ها مشخص باشند، دراین صورت عناصر خوشه jام 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 z_{ij} تلقی نمود. چون

$$E(Z_{ij}) = P(X_i \in \Pi_j) = \frac{\lambda_j f(x_i|\theta_j)}{\sum_{j=1}^{g} \lambda_j f(x_i|\theta_j)}$$

لازم است برای هر j پارامترهای نامعلوم λ_j و θ_j برآورد شوند. برآورد این پارامترها بروش ماکسیمم درستنمایی مستلزم استفاده از روشهای عددی است.

۱.۱.۱ برآورد پارامترها با استفاده از الگوریتم EM

الگوریتم امید ریاضی و ماکسیمم‌سازی[۱] (EM) روشی کاربردی در محاسبات تکراری، برای به دست آوردن برآورد ماکسیمم درستنمایی پارامترها در توزیع‌های آمیخته است که در مسائل گوناگونی مانند داده‌های ناکامل، داده‌های گم شده و غیره کاربرد دارد. این الگوریتم در شرایطی به ویژه در هنگام مواجه‌شدن با داده‌های ناکامل بهتر از الگوریتم نیوتن—رافسون کار می‌کند (مک‌لن و کریشنان، ۱۹۹۷). فرض کنید $\psi^{(0)}$ مقدار اولیه ψ باشد، در این صورت مراحل الگوریتم EM به صورت زیر خواهد بود:
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 می‌شوند که شرط همگرایی $\forall \epsilon > 0 : |L(\psi^{(k+1)}) - L(\psi^{(k)})| < \epsilon$ برقرار شود. (جی اف وو، ۱۹۸۳).
فرض کنید $X_1, \ldots, X_n$ 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.

$$\lambda_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} z_{ij}^{(k)}$$

$$\mu_i^{(k+1)} = \sum_{i=1}^{n} \frac{z_{ij}^{(k)} X_i}{\sum_{i=1}^{n} z_{ij}^{(k)}}$$

$$(\Sigma)^{(k+1)} = \frac{1}{n} \sum_{j=1}^{g} \sum_{i=1}^{n} z_{ij}^{(k)} (X_i - \mu_j^{(k+1)})(X_i - \mu_j^{(k+1)})'$$

$$z_{ij}^{(k)} = \frac{\lambda_j^{(k)} f(x_i|\theta_j^{(k)})}{\sum_{j=1}^{g} \lambda_j^{(k)} f(x_i|\theta_j^{(k)})}$$

با z_{ij} برآورد شده به وسیله الگوریتم EM می توان تحلیل خوشه‌ای را متناسب با بیشترین مقدار $z_{ij}^{(k)}$ انجام داد (مک‌لن و کریشنان، ۱۹۹۷). اما همان طوری که بدان اشاره شد در علومی چون پزشکی و فتوگرافی یکی از اهداف خوشه‌بندی تعیین حجم، شکل و جهت

[۱] Expectation-Maximization Algorithm

خوشه‌ها است. تعیین حجم، شکل و جهت خوشه‌ها با استفاده از تجزیه طیفی ماتریس کوواریانس نخستین بار توسط بانفیلد و رافتری (۱۹۹۳) بیان شد و سرانجام گیلز سیلوکس و گوورت (۱۹۹۵) آنرا تعمیم دادند. معیارهای تحلیل خوشه‌ای توسط فرالی (۱۹۹۹) در نرم‌افزار $MCLUST$ گنجانده شد.

۲.۱ تعیین حجم، شکل و جهت خوشه‌ها

تابع درستنمایی (۳) را می‌توان به صورت

$$L(\psi|Z,X)=\prod_{j=1}^{g}\lambda_j^{n_j}|\Sigma_j|^{-\frac{n_j}{2}}\exp\left\{-\frac{1}{2}\sum_{X_i\in C_j}(X_i-\mu_j)'\Sigma_j^{-1}(X_i-\mu_j)\right\} \quad (5)$$

نوشت، که در آن C_j تعداد مشاهدات X_{ij}ای است که توسط Z_i به زیرجامعه Π_j تخصیص یافته است و n_j تعداد مشاهدات در C_j است. فرض کنید Σ_j، $j=1,\ldots,g$، ماتریس کوواریانس مولفه jام باشد. در این صورت تجزیه طیفی ماتریس متقارن Σ_j بصورت

$$\Sigma_j=D_jB_jD_j'=\sum_{k=1}^{j}\xi_je_je_j'$$

است، که در آن $D_j=(e_1,\ldots,e_j)$ ماتریس متعامد از بردارهای ویژه یکه و B_j ماتریس قطری از مقادیر ویژه می‌باشد. تجزیه طیفی فوق را می‌توان به صورت $\Sigma_j=\xi_jD_jA_jD_j'$ نوشت که در آن $\xi_j=|\Sigma_j|^{\frac{1}{d}}$ و A_j ماتریس قطری از مقادیر ویژه نرمال شده می‌باشد. پارامتر ξ_j، مشخص کننده حجم مولفه jام، D_j جهت آن و A_j شکل آن می‌باشد. فرض کنید برخی از این کمیت‌ها (حجم، شکل و جهت) بین خوشه‌ها متفاوت باشند. در این صورت می‌توان معیارهای ساده‌ای برای خوشه‌بندی به دست آورد که در اکثر علوم کاربرد دارند (بن اسماعیل و سیلوکس، ۱۹۹۶). در مدل‌های زیر علامت اختصاری $E(Equal)$، بیان کننده تساوی، $F(Fixed)$ نشان دهنده ثابت بودن کمیت و $V(Vary)$ نشان دهنده متغیر بودن کمیت است.

در مدل EEE که با $\xi DAD'$ نیز نشان داده می‌شود حجم، شکل و جهت کلیه خوشه‌ها یکسان است و به راحتی ثابت می‌شود تخصیص بهینه $\hat{Z}$، معادل با مینیمم کردن $|W|$ است.

در مدل VEE (ξ_jDAD') شکل و جهت خوشه‌ها ثابت ولی حجم خوشه‌ها متغیر است. فرض کنید $\Sigma_j=\xi_jC$ که $C=DAD'$، با جایگذاری $\Sigma_j=\xi_jC$ در (۵) داریم،

$$\log L(X|\psi)\propto-\frac{d}{2}\sum_{j=1}^{g}n_j\log(\xi_j)-\frac{1}{2}\sum_{j=1}^{g}\frac{1}{\xi_j}tr(W_jC^{-1})$$

که در آن $\hat{\lambda}_j = \frac{n_j}{n}$، $\hat{\mu}_j$ و $\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{g} W_j$، به ترتیب برآوردهای ماکسیمم درستنمایی λ_j، μ_j و Σ هستند و $W_j = \sum_i (X_{ij} - \bar{X}_j)(X_{ij} - \bar{X}_j)'$ است. ماکسیمم کردن تابع درستنمایی (۵) با در نظر گرفتن مدل "VEE"، هم‌ارز مینیمم کردن

$$F(\xi_1, \ldots, \xi_g, C) = \sum_{j=1}^{g} \frac{1}{\xi_j}(tr(W_j C^{-1}) + d\sum_{j=1}^{g} n_j \log(\xi_j)$$

است و نشان داده می‌شود که مینیمم کردن آن باید به روش تکراری (الگوریتم EM) صورت گیرد. با فرض ثابت بودن ماتریس C مقداری از ξ_j که $F(\xi_1, \ldots, \xi_g, C)$ را مینیمم می‌کند عبارت است از،

$$\xi_j = \frac{tr(W_j C^{-1})}{d n_j} \qquad (۶)$$

فرض کنید ξ_jها ثابت باشند در این صورت مینیمم کردن $F(\xi_1, \ldots, \xi_g, C)$، معادل مینیمم کردن $tr(\sum_{j=1}^{g}(\frac{1}{\xi_j}W_j)C^{-1})$ است. سیلوکس و گورت (۱۹۹۵) نشان دادند ماتریس متقارن d بعدی $M = \frac{Q}{|Q|^{\frac{1}{d}}}$، که در آن Q ماتریس متقارن و معین مثبت است و $|M| = 1$، $tr(QM^{-1})$ را مینیمم می‌کند بنابراین ماتریسی که $tr(\sum_{j=1}^{g}(\frac{1}{\xi_j}W_j)C^{-1})$ را مینیمم کند بصورت

$$C = \frac{\sum_{j=1}^{g}(\frac{1}{\xi_j}W_j)}{|\sum_{j=1}^{g}(\frac{1}{\xi_j}W_j)|^{\frac{1}{d}}} \qquad (۷)$$

خواهد بود با توجه به معادلات (۶) و (۷) نمی‌توان فرم بسته‌ای از یک معیار خوشه‌بندی ارائه نمود و بایستی از الگوریتم EM استفاده شود.

مدل EVV (مدل $\xi D_j A_j D_j'$): با قرار دادن $C_j = D_j A_j D_j'$، ماکسیمم کردن (۵) معادل با مینیمم کردن

$$F(\xi, C_1, \ldots, C_g) = \sum_{j=1}^{g} \frac{1}{\xi}(tr(W_j C_j^{-1}) + nd\log(\xi)$$

است. بنابراین $C_j = \frac{W_j}{|W_j|^{\frac{1}{d}}}$ و از طرفی ξای که تابع F را مینیمم می‌کند عبارت است از

$$\xi = \frac{\sum_{j=1}^{g} tr(W_j C_j^{-1})}{nd}$$

با جایگذاری C_j، در معادله فوق داریم $\xi = \frac{\sum_{j=1}^{g}(|W_j|^{\frac{1}{d}})}{nd}$. بنابراین برای مدل "$EVV$"، افرازی که $\sum_{j=1}^{g}(|W_j|^{\frac{1}{d}})$ را مینیمم می‌کند، معادل با ماکسیمم درستنمایی است.

در مدل‌های قبلی شکل خوشه‌ها چه ثابت و چه متغیر، بیضوی هستند. اکنون دو مدل که شکل خوشه‌های آنها کروی است، معرفی می‌شوند. در مدل $\Sigma_j = \xi I$ شکل خوشه‌ها کروی و حجم آنها با هم مساوی است. با فرض این مدل، ماکسیمم کردن (۵) هم‌ارز مینیمم کردن

$$F(\xi) = \frac{1}{\xi} tr(W) + nd \log \xi$$

می‌باشد و $\xi = \frac{tr(W)}{nd}$ تابع $F(\xi)$ را مینیمم می‌کند. با جایگذاری ξ در F داریم، $F(\xi) = nd + nd \log(\frac{tr(W)}{nd})$. پس به جای مینیمم کردن F می‌توان $nd \log(tr(W))$ را مینیمم کرد، که کاربردی‌ترین و قدیمی‌ترین معیار خوشه‌بندی است.
در مدل $\Sigma_j = \xi_j I$ شکل خوشه‌ها کروی ولی حجم آنها با هم متفاوت است. در این صورت ماکسیمم کردن (۵) هم‌ارز مینیمم کردن

$$F(\xi_1, \ldots, \xi_g) = \sum_{j=1}^{g} \frac{1}{\xi_j} tr(W_j) + d \sum_{j=1}^{g} n_j \log \xi_j$$

است. با مشتق گیری از F نسبت به ξ_j و مساوی صفر قرار دادن آن داریم

$$\xi_j = \frac{tr(W_j)}{dn_j}$$

با جایگذاری ξ_j در F خواهیم داشت،

$$F(\xi_1, \ldots, \xi_g) = \sum_{j=1}^{g} dn_j + d \sum_{j=1}^{g} n_j \log(\frac{tr(W_j)}{dn_j}).$$

بنابراین مینیمم کردن F هم‌ارز مینیمم کردن $\sum_{j=1}^{g} n_j \log(\frac{tr(W_j)}{n_j})$ می‌باشد. معیارهای خوشه‌بندی براساس مدل را می‌توان به طور خلاصه در جدول ۱ ملاحظه نمود.

۲ انتخاب بهترین مدل با استفاده از معیار BIC

فرض کنید X_1، ...، X_n یک نمونه تصادفی از توزیع آمیخته با تابع چگالی احتمال (۱) باشد. در استنباط بیزی برای مدل های آمیخته گاوسی، یک روش ساده برای انتخاب بهترین مدل و تعیین تعداد مولفه‌ها محاسبه عامل بیزی[۱] است. عامل بیزی B_{10}، برای مدل M_1 در مقابل مدل M_0 به صورت

$$B_{10} = \frac{P(\underline{X}|M_1)}{P(\underline{X}|M_0)}$$

[۱] Bayese Factor

معیار خوشه‌بندی	جهت	شکل	حجم	شکل خوشه‌ها	مدل	علامت اختصاری
$tr(W)$	موجود نیست	مساوی	مساوی	کروی	ξI	EI
$\sum_{j=1}^{g} n_j \log(\frac{tr(W_j)}{dn_j})$	موجود نیست	مساوی	متغیر	کروی	$\xi_j I$	VI
$\lvert W \rvert$	مساوی	مساوی	مساوی	بیضوی	$\xi D A D'$	EEE
$\sum_{j=1}^{g} n_j \log(\frac{\lvert W_j \rvert}{n_j})$	متغیر	متغیر	متغیر	بیضوی	$\xi_j D_j A_j D_j'$	VVV
از الگوریتم EM	مساوی	مساوی	متغیر	بیضوی	$\xi_j D A D'$	VEE
از الگوریتم EM	متغیر	ثابت	متغیر	بیضوی	$\xi_j D_j A D_j'$	VFV
$\sum_{j=1}^{g} \lvert W \rvert^{\frac{1}{d}}$	متغیر	متغیر	مساوی	بیضوی	$\xi D_j A_j D_j'$	EVV

جدول ۱: معیارهای خوشه بندی براساس مدلهای آمیخته با مولفه‌های نرمال چند متغیره

تعریف می‌شود، که در آن

$$P(\underline{X}|M_j) = \int P(\underline{X}|\psi_j, M_j) P(\psi_j|M_j) d\psi_j. \qquad (۸)$$

و ψ_j بردار پارامتری (λ_j, θ_j) در مدل M_j و $P(\psi_j|M_j)$ تابع چگالی پیشین است. بنابراین عامل بیزی شانس پسین یک مدل در برابر مدل دیگر می‌باشد (کاس و رافتری، ۱۹۹۵). هنگامی که از الگوریتم EM برای برآورد ماکسیمم درستنمایی پارامترهای مدل آمیخته استفاده شود یک روش کلاسیک برای تقریب (۸) استفاده از معیار اطلاع بیزی (BIC) است. این تقریب عبارت است از

$$۲ \log P(\underline{X}|M_j) = ۲ L_M(\underline{X}|\hat{\psi}) - m_M \ln(n) = BIC(M, g).$$

که در آن $L_M(\underline{X}|\hat{\psi})$، لگاریتم تابع درستنمایی و m_M تعداد پارامترهایی است که با در نظر گرفتن مدل M می‌بایست برآورد شوند. بر این اساس مدلی که دارای بیشترین مقدار BIC باشد، بهترین خواهد بود. درحالیکه براساس عامل بیزی اگر $B_{۱۰} > ۱$ باشد آنگاه مدل $M_۱$ بهتر از $M_۰$ است

س‌اساربی‌دنبه‌شوخی‌امهشورح‌یرشتی‌ارب :قبنزلگی‌اهه‌داد ن‌درکی‌دنبه‌شوخ :لاثم
)۹۷۹۱(ایدرام رد اهه‌داد ن‌یا .دوشی‌م ه‌دافتسا)۶۳۹۱(رشیف قبنزلگ ی‌اهه‌داد زا لدم
ه‌س زا قبنزلگ ۵۱۰ ی‌ارب اهگ‌ربلگ و اهگ‌ربساک ضرع و لوط لماش ه‌ک دناه‌دش ه‌داد
.دنشابی‌م ی‌یاینیجرو و گ‌نراگنر ،رادراخ عون

۱.۲ کاربرد الگوریتم EM در خوشه‌بندی براساس مدل

همانطوری که بیان شد در مرحله E الگوریتم EM ماتریس مشخصه گروهبندی، $\{Z_{ij}\}$، محاسبه می‌شود که در آن $\hat{Z}_{ij}$، برآورد احتمال شرطی تعلق مشاهده iام به گروه k با انتخاب مقادیر اولیه برای پارامترها یا برآورد آنها است. سپس در «مرحله M» برآورد ماکسیمم درستنمایی پارامترها مشخص می‌شود.

برای این منظور از نرم‌افزار $Mclust$ استفاده شده است. براساس یک سری محاسبات با $MCLUST$ داریم $z_{۱,۱} = ۱$، $z_{۱,۲} = ۵.۸e^{-۰۲۲}$ و $z_{۱,۳} = ۳.۱۷e^{-۰۴۲}$ همچنین $z_{۱۱۹,۳} = ۱$ $z_{۱۱۹,۱} = ۲.۱۷e^{-۰۵۸}$ و $z_{۱۱۹,۲} = ۱.۹۶e^{-۰۰۹}$. چون $z_{۱,۱}$ از $z_{۱,۲}$ و $z_{۱,۳}$ بزرگتر است مشاهده ۱ام در خوشه اول قرار می‌گیرد، در حالی که مشاهده ۱۱۹ام به خوشه سوم تعلق می‌یابد. به همین ترتیب براساس z برآورد شده، در مورد خوشه اصلی سایر مشاهدات نیز می‌توان تصمیم گرفت. بر این اساس مشاهدات به گروهی تعلق می‌گیرند که دارای $\hat{z}$ بیشتری باشند. بعد از خوشه‌بندی کردن داده‌ها، می‌توان برآورد ماکسیمم درستنمایی پارامترها را به دست آورد که به صورت زیر می‌باشند.

$$\hat{\mu} = \begin{pmatrix} ۵.۰۰۶ & ۵.۹۴۲۳۳۱ & ۶.۵۷۴۶۲۳ \\ ۳.۴۲۸ & ۲.۷۶۰۷۵۷ & ۲.۹۸۰۷۹۲ \\ ۱.۴۶۲ & ۴.۲۵۸۷۱۸ & ۵.۵۳۹۰۱۶ \\ ۰.۲۶۴ & ۱.۳۱۹۲۰۳ & ۲.۰۲۴۹۳۳ \end{pmatrix}$$

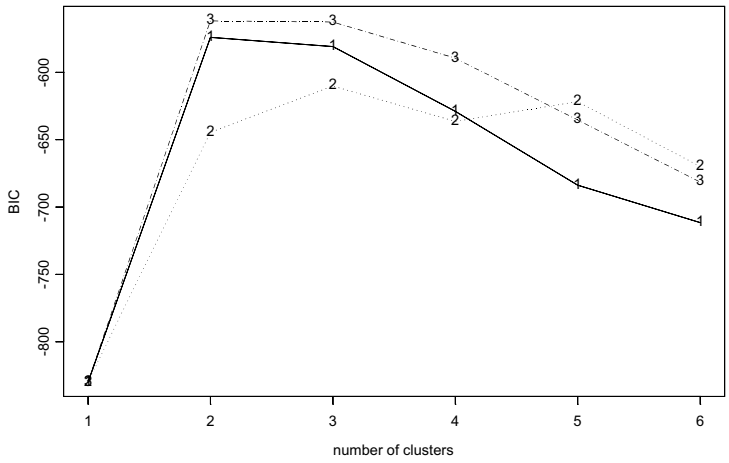
$$\hat{\Sigma} = \begin{pmatrix} ۰.۲۶۳۹۳۴۸۰ & ۰.۰۸۹۸۴۹۸۲ & ۰.۱۶۹۶۵۷۷۸ & ۰.۰۳۹۳۳۸۰۸ \\ ۰.۰۸۹۸۴۹۸۲ & ۰.۱۱۱۹۴۷۷۶ & ۰.۰۵۱۱۲۰۷۶ & ۰.۰۲۹۹۷۸۳۴ \\ ۰.۱۶۹۶۵۷۷۸ & ۰.۰۵۱۱۲۰۷۶ & ۰.۱۸۶۵۳۴۷۶ & ۰.۰۴۱۹۷۳۳۵ \\ ۰.۰۳۹۳۳۸۰۸ & ۰.۰۲۹۹۷۸۳۴ & ۰.۰۴۱۹۷۳۳۵ & ۰.۰۳۹۷۱۱۹۵ \end{pmatrix}$$

$$\hat{\lambda} = (\ ۰.۳۳۳۳۳۳۳ \quad ۰.۳۲۹۶۱۹۱ \quad ۰.۳۳۷۰۴۷۵)$$

پس از خوشه بندی کردن داده‌ها توسط الگوریتم EM و برآورد پارامترها، به انتخاب بهترین مدل با استفاده از معیار BIC می‌پردازیم.

با توجه به شکل ۱ و جدول ۱ مقدار BIC برای مدل خوشه‌بندی "VEV" با تعداد دو خوشه بیشتر از سایر حالت‌هاست. لذا این مدل با فرض دو خوشه بهترین مدل براساس معیار BIC در بین سه مدل فوق با کلیه حالات درنظر گرفته شده برای تعداد خوشه‌ها می‌باشد.

مدل	۱	۲	۳	۴	۵	۶
VVV	−۸۲۹.۹۷۸۲	−۵۷۴.۰۱۷۸	−۵۸۰.۸۳۸۹	−۶۲۸.۹۵۶۴	−۶۸۳.۸۱۱۴	−۷۱۱.۵۶۵۷
EEV	−۸۲۹.۹۷۸۲	−۶۴۴.۵۹۹۷	−۶۱۰.۰۸۳۶	−۶۴۵.۹۹۵۰	−۶۲۱.۶۹۰۱	−۶۶۹.۷۰۶۹
VEV	−۸۹۲.۹۷۸۲	−۵۶۱.۷۲۸۵	−۵۶۲.۵۵۰۷	−۵۸۹.۳۵۱۰	−۶۳۵.۲۰۵۱	−۶۸۱.۲۹۷۶



شکل ۱: نمودار برای سه مدل (۱): ، (۲): و (۳):

۳ بحث و نتیجه‌گیری

در این مقاله روش خوشه‌بندی براساس مدل با مولفه‌های نرمال چند متغیره مورد بررسی قرار گرفت و نشان داده شد که این روش معادل با برآورد پارامترهای مولفه‌ها می‌باشد. سپس از الگوریتم EM برای برآورد پارامترها استفاده شد. همچنین بیان شد که در این روش فرضیات دلخواه اشخاص در مورد معیار تشابه نقشی ندارد و با تجزیه طیفی ماتریس کوواریانس می‌توان حجم، شکل و جهت خوشه‌ها را به دست آورد. بعلاوه می‌توان با استفاده از معیار BIC بهترین مدل را تعیین نمود که هم‌ارز بهترین مقدار ممکن برای تعداد خوشه‌ها نیز می‌باشد.

مراجع

Banfield, J. D. and Raftery, A. E. (1993), Model-Based Gaussian and Non-Gaussin Clustering. *Biometrics*, 49, 803-821.

Bensmail, H. and Celeux, G. (1996), Regularized Gaussian Discriminant Analysis Through Eiegnvalue Decomposition. *J. Amer. Stat. Assoc.*, 91, 1743-1748.

Celeux, G. and Govert, G. (1995), Gaussian Parsimonious Clustring Models. *Pattern Recognition*, 28, 781-793.

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.

Fraley, C. and Raftery, A. E. (1999), MCLUST: Software for Model-Based Cluster Analysis, *J. Classification*, 16, 297-306.

Hartigan, J. A. (1975), *Clustering Algorithms.* Wiley, New York.

Jeff C. F., (1983), On the Convergence of the EM Algorithm, *Annals of Statistcs*, 11, 95-103.

Kass, R. E. and Raftery, A. E. (1995), Bayes Factors. *J. Amer. Stat. Assoc.*, 90, 773-775.

Mardia, K. V. and Kent, J. M. (1979), *Multivariate Analysis*, Bibby, acadimic Press.

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.

McLachlan, G. J. and Krishnan, T. (1997), *The EM Algorithm and Extensions.* Wiley, New York.

کاربرد گرافهای تصادفی بازه‌ای در تحلیل خوشه‌ای

محمد قاسم وحیدی‌اصل[۱]، محسن محمدزاده[۲]، فرانک گودرزی[۲]

$P۱۳۰۳۹$

[۱] گروه آمار، دانشگاه شهید بهشتی
[۲] گروه آمار، دانشگاه تربیت مدرس

چکیده: در این مقاله، مدلی برای گرافهای تصادفی بازه‌ای معرفی شده و خواصی در ارتباط با این مدل عنوان می‌شود. سپس نتایج دقیقی برای توزیع متغیرهای تصادفی مربوط به همبندی این گراف تصادفی بازه‌ای مورد بحث قرار می‌گیرد.

درادامه خواص مجانبی تعدادی آزمون ترکیبیاتی بالقوه سودمند برای تحلیل خوشه‌ای بر اساس نظریه گرافهای تصادفی بازه‌ای برای داده‌هایی در فضای k بعدی اقلیدسی توصیف شده، سپس مقایسه‌هایی از اثر مجانبی یک رده از این آزمونها ارائه می‌شوند. به عنوان یک مورد از کاربردهای ممکن، آشکارسازی آمیزه‌های توزیعهای احتمال مورد بحث قرارگرفته و مثالی عددی ارائه می‌شود.

واژه‌های کلیدی: گرافهای تصادفی بازه‌ای، تحلیل خوشه‌ای، مدلهای آمیخته، آشکارسازی.

۱ بخش اول

مدلهای نظری گراف وقتی مفید هستند که قصد ما اکتشاف ساختارهای مجموعه‌های داده‌ها باشد. در تحلیل خوشه‌ای، ماتریس تشابه یک ساختار گراف را روی n شئ که باید خوشه‌بندی شوند (که به عنوان رأسهای ۱، ...، n تعبیر می شود) به وجود می‌آورد؛ دو رأس به وسیلهٔ یک یال به هم متصل می شوند اگر و تنها اگر شیء های متناظر به قدرکافی شباهت داشته باشند. با به کاربردن مفهوم گرافهای تصادفی، می‌توان آماره‌های آزمونی برای آزمون تصادفی بودن خوشه‌های حاصل به دست آورد.

موضوع گرافهای بازه‌ای درتقاطع مسیرهای ترکیبیات و احتمال قرار می‌گیرد. شاینرمن این دو حوزه را به امید آشکارسازی قضیه‌های جدید و کاربردها در هم آمیخته است. نه تنها این حوزه یک شاخه سحرآمیز و زیبای ریاضیات است، بلکه همچنین، کاربردهای زیادی از جمله در مطالعهٔ پلیمریزاسیون، محاسبه کردن خطای تحمل و تحلیل احتمالاتی الگوریتم‌ها دارد.
گرافهای بازه‌ای، ردهٔ خاصی از گرافها با چندین خاصیت نظری جالب توجه را تشکیل

می‌دهند. به علاوه، آنها به عنوان مدلهای مفیدی در خیلی از کاربردها به انضمام تخصیص فرکانس رادیو، زمانبندی پردازشگر، توالی زمانی باستان شناسی و زمان بندی چراغهای راهنمایی به کار می‌روند.
در این مقاله مدلی برای گرافهای تصادفی بازه‌ای معرفی می‌کنیم، این مدل "ایستا" است: برای هر عدد صحیح مثبت n، تنها یک فضای احتمال برای گرافهای بازه‌ای روی n رأس برچسب گذاری شده معرفی می‌کنیم.
به عنوان بازه، یک بازهٔ بستهٔ حقیقی را در نظر می‌گیریم. گراف (ساده، متناهی) G یک گراف بازه‌ای نامیده می‌شود اگر بتوان یک بازه به هر یک از رأسهای آن منتسب کرد، به طوری که دو رأس مجاور باشند اگر و تنها اگر بازه‌های متناظر با آنها اشتراک داشته باشند. به طور رسمی‌تر، فرض کنید $\mathcal{I}$ مجموعهٔ همهٔ بازه‌های حقیقی را نشان دهد. نمایشی برای گراف بازه‌ای G تابع $f : V(G) \to \mathcal{I}$ است، به طوری که برای هر زوج از رأسهای متمایز v و w

$$v \sim w \Leftrightarrow f(v) \cap f(w) \neq \emptyset.$$

نماد $v \sim w$ به معنی آن است که v با w مجاور است.
گرافهای بازه‌ای، گرافهایی هستند که نمایش بازه‌ای را بپذیرند (گیلمور و هافمن (۱۹۶۴)، گُلومبیک (۱۹۸۰) را ببینید).

۲ مدل

مدل یکنواخت $\mathcal{IG}_{n,d}$. فرض کنید X_1، X_2، ...، X_n، n متغیر تصادفی مستقل باشند که به طور یکنواخت روی بازهٔ واحد توزیع شده‌اند. یک گراف تصادفی $\mathcal{IG}_{n,d}$ به وسیلهٔ مجموعهٔ رئوس $\mathcal{V} = \{1, \ldots, n\}$، متناظر با این مجموعه از n متغیر تصادفی، و مجموعهٔ یالهای $\mathcal{E} = \{(i,j) :| X_i - X_j | \leq d\}$ تعریف می‌شود، که $0 < d < 1$. برای n تحقق x_1، ...، x_n از متغیرهای تصادفی، یک گراف بازه‌ای را به عنوان تحققی از گراف تصادفی $\mathcal{IG}_{n,d}$ به دست می آوریم.

در اینجا متذکر می‌شویم که سطح فاصلهٔ d در $\mathcal{IG}_{n,d}$ معمولاً به عنوان تابعی مشخص از n در نظر گرفته می‌شود.

می‌توان به آسانی بررسی کرد که برای $\mathcal{IG}_{n,d}$،

$$P((i,j) \in \mathcal{E}) = P(|X_i - X_j| \leq d) = 2d - d^2.$$

این تعریف مستلزم آن است که علاوه بر خواص نظری گراف و مشخصه‌ها ، بتوان خواص مربوط به هندسه رأسها را مطالعه کرد ، یعنی، تحققهای x_1، x_2، ...، x_n روی بازه واحد. برای مثال، طول یک زیرگراف همبند، یا " اندازهٔ" (تعداد رأسهای) اولین مؤلفه. بدیهی است که، یالهای این گرافهای تصادفی مستقل نیستند. برای d ی کوچک (مرحله تنک از تکامل $\mathcal{IG}_{n,d}$)، این وابستگی قوی نیست. بنابراین وقتی که احتمالهای یال $P(n) \sim 2d(n)-d^2(n)$ به اندازه کافی کوچک هستند، انتظار می‌رود مشابهت‌هایی با $\mathcal{G}_{n,p}$ (گیلبرت (۱۹۵۹) را ببینید) وجود داشته باشد. اما، در حالت کلی، هر دو مدل، حتی وقتی d خیلی کوچک است، در بسیاری جهات متفاوت هستند.

۱.۲ خواص گرافهای تصادفی بازه‌ای

از مهمترین خواص گرافهای تصادفی دو قضیه زیر از شاینرمن (۱۹۸۸) هستند که آنها را بدون برهان در زیر نقل می‌کنیم.

قضیه ۱ : تقریباً همه گرافها در G_n، $\frac{n^2}{4} + o(n^2)$ یال دارند.

قضیه ۲ : فرض کنید $G \in G_n$ و $v \in V(G)$. برای مقدار ثابت x با $x \in [0, 1]$، داریم

$$\lim_{n\to\infty} P(d(v) \le xn) = \begin{cases} 1-(1-x)^{\frac{\pi}{2}} & ; x \ge \frac{1}{2} \\ 1-(1-x)\left\{\frac{\pi}{2} - 2\cos^{-1}[\frac{1}{\sqrt{2-2x}}]\right\} - \sqrt{1-2x} & ; x < \frac{1}{2} \end{cases}$$

δ و Δ مینیمم و ماکسیمم درجه‌های یک گراف را نشان می‌دهند. نتیجه فوق ایجاب می‌کند:

نتیجه ۱ : برای هر $\varepsilon > 0$، تقریباً همهٔ گرافهای بازه‌ای در $\delta < \varepsilon n$ و $\Delta > (1-\varepsilon)n$ صدق می‌کنند.

۳ نتایج دقیق برای گراف تصادفی بازه‌ای $\mathcal{IG}_{n,d}$

این بخش را با دنباله‌ای از لم‌ها آغاز می‌کنیم که بر اساس آنها گودهارت ویاوورسکی (۱۹۹۶) فرمولهای دقیقی را برای احتمالهای بعضی خواص اصلی همبندی $\mathcal{IG}_{n,d}$ به دست آوردند. در اینجا متذکر می‌شویم که لم کلیدی، یعنی لم ؟؟، در واقع هم‌ارز با نتیجه مربوط به مسائل پوشش است. احتمالی که می‌خواهیم به دست آوریم مساوی با این احتمال است

که $k-۱$ کمان انتخاب شده به تصادف روی محیط یک دایره با پیرامون y به طور کامل دایره را بپوشاند.

لم ۲ : فرض کنید $[x, x+y]$ یک زیربازه به طول y از $[۰, ۱]$ باشد. فرض کنید دو تا از k رأس مفروض در مرزهای این زیربازه قرار داده شوند و $\mathcal{A}_{k,y,d}$ عبارت از این پیشامد باشد که $k-۲$ نقطه، متناظر با باقیماندهٔ رأسها که به طور تصادفی از $[۰, ۱]$ استخراج می‌شوند، درون $[x, x+y]$ باشند و مرزها را به هم "متصل" کنند، یعنی، k رأس یک زیرگراف همبند به طول y تشکیل می‌دهند و فرض کنید $P(k,y,d) = P(\mathcal{A}_{k,y,d})$. آنگاه رابطه بازگشتی زیر برقرار است:

$$P(k,y,d) = yP(k-۱,y,d) + ((k-۱)d-y)P(k-۱,y-d,d)$$

با شرایط اولیه

$$P(i,u,d) = \begin{cases} u^{i-۲} & \text{اگر } i \geq ۲ \text{ و } u \leq d \\ ۰ & \text{اگر } i = ۲ \text{ و } u > d \end{cases}$$

یک فرمول بسته برای این احتمال عبارت است از

$$P(k,y,d) = \sum_{j=۰}^{\min\{k-۱,[y/d]\}} \binom{k-۱}{j} (-۱)^j (y-jd)^{k-۲}.$$

برهان. توجه می‌کنیم که $P(k,y,d)$ به x، $۰ \leq x \leq ۱-y$، و انتخاب k رأس در فضای احتمال بستگی ندارد. بدیهی است که،

$$\text{اگر } k = ۲,۳,\ldots \quad \text{و} \quad y > (k-۱)d \qquad P(k,y,d) = ۰$$

زیرا در این صورت زیربازه بزرگتر از آن است که مرزها را با به کار بردن $k-۲$ رأس به هم متصل کند. از این پس فرض کنید $y \leq (k-۱)d$ و $\mathcal{B}$ عبارت از این پیشامد باشد که پس از انتخاب کردن $k-۳$ نقطه در بازهٔ به طول y، از میان $k-۲$ فاصله‌گذاری به وجود آمده بین این $k-۳$ نقطه و نقاط مرزی، $k-۳$ فاصله‌گذاری با طول کمتر یا مساوی d و دقیقاً یک فاصله‌گذاری با طول بزرگتر از d اما کوچکتر یا مساوی $۲d$ وجود دارند. برای بیان عبارت زیر، فرض می‌شود که $\mathcal{A}_{k-۱,y,d}$ به وسیله برخی از $k-۳$ رأس مفروض از میان $k-۲$ رأس دیگر تولید شود. آنگاه

$$P(\mathcal{A}_{k,y,d}) = P(\mathcal{A}_{k,y,d}|\mathcal{A}_{k-۱,y,d})P(\mathcal{A}_{k-۱,y,d}) + P(\mathcal{A}_{k,y,d} \cap \mathcal{A}^c_{k-۱,y,d} \cap \mathcal{B})$$

و به وضوح

$$P(\mathcal{A}_{k,y,d}|\mathcal{A}_{k-۱,y,d}) = y \qquad \text{و} \qquad \mathcal{A}^c_{k-۱,y,d} \cap \mathcal{B} = \mathcal{B}$$

حال پیشامد $\mathcal{A}_{k,y,d} \cap \mathcal{B}$ را در نظر می‌گیریم . ملاحظه کنید که پس از حذف یک بازه به طول d از فاصله‌گذاری بزرگ یکتا، باید $k-۳$ نقطه مرزهای بازهٔ به طول $y-d$ را به یکدیگر متصل کنند. بنابراین می‌توانیم ابتدا $k-۳$ نقطه مفروض را در بازهٔ به طول $y-d$ به گونه‌ای انتخاب کنیم که پیشامد $\mathcal{A}_{k-۱,y-d,d}$ برقرار باشد. سپس برای به دست آوردن $\mathcal{B}$ می‌بایست یکی از $k-۲$ فاصله‌گذاری را با اضافه کردن بازه‌ای به طول d بسط دهیم و سرانجام k امین رأس را در فاصله‌گذاری "بزرگ" برای به دست آوردن $\mathcal{A}_{k,y,d}$ انتخاب کنیم. $k-۳$ نقطه در نظر گرفته شده بازهٔ به طول $y-d$ را به $k-۲$ بازهٔ از هم جدا به طولهای $y_۱$، $y_۲$، ...، $y_{k-۲}$ به گونه‌ای تقسیم می‌کند که برای $i = ۱, ۲, \ldots, k-۲$، $y_i \leq d$ و $y_۱ + y_۲ + \ldots + y_{k-۲} = y - d$. از این رو ما $k-۲$ امکان مجزّا برای ایجاد $\mathcal{A}_{k,y,d} \cap \mathcal{B}$ داریم. حال فرض کنید یکی از آنها را در نظر بگیریم: برای مثال، فرض کنید که i امین بازهٔ به طول y_i را به یک بازه به طول $y_i + d$ بسط دهیم و آنگاه k امین رأس را برای به دست آوردن $\mathcal{A}_{k,y,d}$ ارائه دهیم. این با احتمال $d - y_i$ انجام می‌شود. این بدان معنی است که برای یک انتخاب مفروض $y_۱$، $y_۲$، ...، $y_{k-۲}$ به گونه‌ای که برای $i = ۱, ۲, \ldots, k-۲$ و $y_i \leq d$ و $y_۱ + y_۲ + \ldots + y_{k-۲} = y - d$ ما یک بازه به طول d را اضافه می‌کنیم و k امین رأس را به گونه‌ای انتخاب می‌کنیم که $\mathcal{A}_{k,y,d}$ با احتمال زیر روی دهد.

$$\sum_{i=۱}^{k-۲} (d - y_i) = (k-۱)d - y,$$

که در واقع به انتخاب $y_۱$، $y_۲$، ...، $y_{k-۲}$ بستگی ندارد. بنابراین،

$$P(\mathcal{A}_{k,y,d} \cap \mathcal{B}) = ((k-۱)d - y)P(k-۱, y-d, d)$$

و چون شرایط اولیه بدیهی هستند، اثبات فرمول بازگشتی برای $y \leq (k-۱)d$ کامل می‌شود.

به آسانی می‌توان تحقیق کرد که فرمول برای $P(k,y,d)$ جواب فرمول بازگشتی ما با شرایط اولیه داده شده می‌باشد. به علاوه،

$$\sum_{j=۰}^{k-۱} \binom{k-۱}{j} (-۱)^j (y - jd)^{k-۲} = d^{k-۲} A_{k-۱}(-\frac{y}{d}, \frac{y}{d} - k + ۱; ۰, -۱),$$

که

$$A_N(a,b;r,s)=\sum_{i=0}^{N}\binom{N}{i}(i+a)^{i+r}(N-i+b)^{N-i+s}$$

نماد ری یوردان برای مجموع آبل[1] است (ری یوردان (۱۹۶۸) را ببینید). با به کار بردن فرمول آبل

$$A_N(a,b;0,-1)=b^{-1}(a+b+N)^N,$$

به دست می‌آوریم

اگر $k=2,3,\ldots$ و $y>(k-1)d$ $\sum_{j=0}^{k-1}\binom{k-1}{j}(-1)^j(y-jd)^{k-2}=0$

$P(k,y,$ همچنین برای $y>(k-1)d$ نیز معتبر است. □

حال این پیشامد را در نظر بگیرید که k رأس مفروض، یک زیرگراف همبند $\mathcal{IG}_{n,d}$ را با این خاصیت اضافی تشکیل دهند که بین نقاط متوالی روی خط (تحقق رأسها)، رأسهای دیگری از گراف بازه‌ای موجود نباشد. یک زیرگراف همبند با این خاصیت را یک زیرگراف همبند کامل می‌نامیم.

لم ۳ : فرض کنید $\mathcal{CS}_k(u)$ عبارت از این پیشامد باشد که k رأس مفروض یک زیرگراف همبند کامل با طول کمتر از $u(u>0)$ را تشکیل دهند. آنگاه

$$P(\mathcal{CS}_k(u))=\frac{n-k+1}{\binom{n}{k}}\sum_{j=0}^{\min\{k-1,[u/d]\}}\binom{k-1}{j}(-1)^j\sum_{t=k-1}^{n}\binom{n}{t}(1-u)^{n-t}\times(u-jd)^t.$$

به ویژه، فرض کنید $\mathcal{CS}_k$ پیشامد آن باشد که k رأس مفروض یک زیرگراف همبند کامل را تشکیل دهند، یعنی، پیشامد $\mathcal{CS}_k(u)$، که $u\geq(k-1)d$. آنگاه

$$P(\mathcal{CS}_k)=\frac{n-k+1}{\binom{n}{k}}\sum_{j=0}^{\min\{k-1,[1/d]\}}\binom{k-1}{j}(-1)^j(1-jd)^n.$$

برهان. با استفاده از تعریف، $P(\mathcal{CS}_1(u))=P(\mathcal{CS}_1)=1$، و به آسانی می‌توان تحقیق کرد که فرمولهای مفروض برای $k=1$ نیز مساوی ۱ است. از این رو، می‌توان فرض

[1] Abel

کرد که $k \geq ۲$. لم فوق را می‌توان توسط انتگرالگیری با به کار بردن لم ؟؟ ثابت کرد. داریم

$$P(\mathcal{CS}_k(u)) = k(k-۱)\int_۰^u \int_۰^{۱-y} dx P(k,y,d)(۱-y)^{n-k} dy$$
$$= k(k-۱)\int_۰^u P(k,y,d)(۱-y)^{n-k+۱} dy.$$

ابتدا قسمت دوم لم را اثبات می‌کنیم، زیرا با به کار بردن کران بالای ۱ در انتگرال فوق روابط به سادگی محاسبه می‌شوند. (پایان اثبات لم ؟؟ را ببینید). با تقسیم کردن بازه $[۰,۱]$ به زیربازه‌ها، به دست می‌آوریم،

$$P(\mathcal{CS}_k) = k(k-۱)\sum_{t=۰}^{[۱/d]-۱}\int_{td}^{(t+۱)d}\sum_{j=۰}^{M}\binom{k-۱}{j}(-۱)^j(y-jd)^{k-۲}(۱-y)^{n-k+۱}dy$$
$$+ k(k-۱)\int_{[۱/d]d}^{۱}\sum_{j=۰}^{M}\binom{k-۱}{j}(-۱)^j(y-jd)^{k-۲}(۱-y)^{n-k+۱}dy$$

با $M = \min\{t, k-۱\}$. حال می‌توان مجموع بیرون انتگرال را به داخل برد. آنگاه پس از تغییر دادن ترتیب مجموعیابی به دست می‌آوریم:

$$P(\mathcal{CS}_k) = k(k-۱)\sum_{j=۰}^{M^*}\binom{k-۱}{j}(-۱)^j\int_{jd}^{۱}(y-jd)^{k-۲}(۱-y)^{n-k+۱}dy$$
$$= k(k-۱)\sum_{j=۰}^{M^*}\binom{k-۱}{j}(-۱)^j\int_۰^{۱} z^{k-۲}(۱-z)^{n-k+۱}(۱-jd)^n dz$$
$$= k(k-۱)\sum_{j=۰}^{M^*}\binom{k-۱}{j}(-۱)^j(۱-jd)^n\frac{\Gamma(k-۱)\Gamma(n-k+۲)}{\Gamma(n+۱)},$$

که در آن $M^* = \min\{k-۱, [۱/d]\}$. چون برای هر عدد صحیح نامنفی m، $\Gamma(m+۱) = m!$، اثبات برای $u > (k-۱)d$ کامل می‌شود.

به همین شیوه برای حالت کلی به دست می‌آوریم:

$$P(\mathcal{CS}_k(u)) = k(k-۱)\sum_{j=۰}^{\min\{k-۱,[u/d]\}}\binom{k-۱}{j}(-۱)^j$$
$$\times\sum_{i=۰}^{n-k+۱}\binom{n-k+۱}{i}(۱-u)^{n-k+۱-i}(u-jd)^{k+i-۱}\frac{\Gamma(k-۱)\Gamma(i+۱)}{\Gamma(k+i)},$$

که برهان را کامل می‌کند. □

حال فرض کنید که k رأس مفروض، یک زیرگراف همبند کامل تشکیل دهند، به گونه‌ای که چپ ترین رأسش درون فاصله d از ۰ قرار گرفته و رأسهای دیگری قبل از آنها روی خط قرار نمی‌گیرند. چنین زیرگرافی، یک زیرگراف مرزی همبند کامل نامیده می‌شود.

لم ۴ : فرض کنید $\mathcal{CBS}_k$ عبارت از این پیشامد باشد که k رأس مفروض یک زیرگراف مرزی همبند کامل را تشکیل دهند. آنگاه

$$P(\mathcal{CBS}_k) = \frac{۱}{\binom{n}{k}} \sum_{j=۰}^{\min\{k,[۱/d]\}} \binom{k}{j} (-۱)^j (۱ - jd)^n.$$

برهان. فرض کنید یک رأس اضافی را در نقطه صفر قرار دهیم. آنگاه پیشامد $\mathcal{CBS}_k$ بر آن اشاره دارد که این رأس همراه با k رأس مفروض، یک زیرگراف همبند را تشکیل دهند. بنابراین مجدداً با به کار بردن لم ؟؟، شبیه اثبات لم ؟؟، داریم

$$\begin{aligned} P(\mathcal{CBS}_k) &= k \int_۰^۱ P(k+۱, y, d)(۱-y)^{n-k} dy \\ &= k \sum_{j=۰}^{\min\{k,[۱/d]\}} \binom{k}{j} (-۱)^j \int_۰^{۱-jd} y^{k-۱} (۱ - jd - y)^{n-k} dy \\ &= k \sum_{j=۰}^{\min\{k,[۱/d]\}} \binom{k}{j} (-۱)^j (۱-jd)^n \frac{\Gamma(k)\Gamma(n-k+۱)}{\Gamma(n+۱)} \square \end{aligned}$$

لم ۵ : فرض کنید $\mathcal{CC}_k$ پیشامد آن باشد که k رأس مفروض، یک مؤلفهٔ همبند را تشکیل دهند. آنگاه، برای $k < n$، احتمال این پیشامد به صورت زیر داده می‌شود

$$\begin{aligned} P(\mathcal{CC}_k) = \frac{۲}{\binom{n}{k}} &\sum_{j=۰}^{\min\{k,[۱/d]-۱\}} \binom{k}{j} (-۱)^j (۱ - (j+۱)d)^n \\ &+ \frac{n-k+۱}{\binom{n}{k}} \sum_{j=۰}^{\min\{k-۱,[۱/d]-۲\}} \binom{k-۱}{j} (-۱)^j (۱ - (j+۲)d)^n \end{aligned}$$

و

$$P(\mathcal{CC}_n) = P(\mathcal{CS}_n).$$

برهان. با استفاده از لم ؟؟، احتمال آنکه تمام n رأس از یک زیربازه به طول $۲d - ۱$ انتخاب شوند و در مجموع k رأس مفروض از آنها یک زیرگراف همبند کامل تشکیل دهند، به صورت زیر است

$$\frac{n-k+۱}{\binom{n}{k}} \sum_{j=۰}^{\min\{k-۱,[۱/d]-۲\}} \binom{k-۱}{j} (-۱)^j (۱-۲d-jd)^n$$

زیرا تحت فرض فوق، عملاً با یک زیرگراف بازه‌ای $\mathcal{IG}_{n,d^*}$ سر و کار داریم، که $d^* = d/(۱-۲d)$. حال فرض کنید این بازه به طول $۲d - ۱$ را با افزودن دو بازه به طول d به بازه اولیه $[۰, ۱]$ بسط دهیم، یکی قبل از نقطه منتهی‌الیه سمت چپ زیرگراف و دیگری بعد از نقطه منتهی‌الیه سمت راست. آنگاه زیرگراف همبند، یک مؤلفهٔ $\mathcal{IG}_{n,d}$ خواهد بود. بنابراین فرمول فوق احتمال مطلوب را برای حالت خاصی نشان می‌دهد که هر دو نقاط انتهایی مؤلفه، درون فاصله d از مرزهای بازه واحد واقع نیستند. سه حالت دیگری که اینجا در نظر گرفته می‌شوند عبارتند از: دو، هم ارزی بالفعل به واسطهٔ تقارن، که درست یکی از نقاط انتهایی درون فاصله d از حاشیهٔ بازهٔ واحد است؛ و آخرین حالت موردی است که هر دو نقاط انتهایی دارای این خاصیت می‌باشند. اما، آخرین حالت تنها برای $k = n$ امکان پذیر است، که بدان معنی است که، درست یک مؤلفه داریم؛ احتمال برای $k = n$ با استفاده از لم ؟؟ داده می‌شود. برای دو حالت متقارن، لم ؟؟ به شیوه‌ای همانند با حالت نخست به کار برده می‌شود؛ یعنی هم اکنون n نقطه را روی یک زیربازه به طول d قرار می‌دهیم. با استفاده از لم ؟؟، احتمال اینکه تمام n رأس از یک زیربازه به طول $d - ۱$ انتخاب شوند و k رأس مفروض آنها یک زیرگراف مرزی همبند کامل تشکیل دهند عبارت است از

$$\frac{۱}{\binom{n}{k}} \sum_{j=۰}^{\min\{k,[۱/d-۱]\}} \binom{k}{j} (-۱)^j (۱-d-jd)^n.$$

حال، زیربازهٔ به طول $d - ۱$ را به بازهٔ $[۰, ۱]$ بسط می‌دهیم، با افزودن یک بازه به طول d، یا قبل از نقطه منتهی‌الیه سمت چپ زیرگراف یا بعد از نقطه منتهی‌الیه سمت راست که و اثبات قضیه فوراً نتیجه می‌شود. □

لم ٦ : فرض کنید $\mathcal{P}_۲$ عبارت از این پیشامد باشد که دو رأس مفروض به وسیله یک راه به هم متصل شوند. آنگاه

$$P(\mathcal{P}_۲) = ۱ - \frac{۱}{\binom{n}{۲}} \sum_{j=۱}^{\min\{n-۱,[۱/d]\}} \binom{n+۱}{j+۲} (-۱)^{j+۱} (۱-jd)^n.$$

برهان. بدیهی است که، اگر دو رأس توسط یک راه به هم متصل شوند، آنگاه آنها در مؤلفه یکسانی می‌باشند و بالعکس. آنگاه

$$P(\mathcal{P}_۲) = ۲\sum_{k=۰}^{n-۲}\binom{n-۲}{k}\frac{۱}{(k+۱)(k+۲)}P(\mathcal{CS}_{k+۲}).$$

از این رو با استفاده از لم ؟؟، بعد از تغییر دادن ترتیب مجموعیابی، به دست می آوریم

$$P(\mathcal{P}_۲) = \frac{۱}{\binom{n}{۲}}\sum_{j=۱}^{\min\{n-۱,\lfloor ۱/d\rfloor\}}(-۱)^j(۱-jd)^n\sum_{k=j-۱}^{n-۲}\binom{k+۱}{j}(n-k-۱)$$
$$+\frac{۱}{\binom{n}{۲}}\sum_{k=۰}^{n-۲}(n-k-۱).$$

آخرین مجموع $\binom{n}{۲}$ است و چون

$$\sum_{k=j-۱}^{n-۲}\binom{k+۱}{j}(n-k-۱) = \sum_{k=۰}^{n-j-۱}\binom{n-k}{j+۱} = \binom{n+۱}{j+۲},$$

د (برای اتحادها برای مثال گولد (۱۹۷۲) را ببینید). □

حال گودهارت و یاوورسکی (۱۹۹۶) نتایج اصلی‌اشان را در رابطه با توزیعهای متغیرهای تصادفی بیان می‌کنند: تعداد C_n مؤلفه، تعداد C_n^k مؤلفه با اندازه k، و تعداد $C_n^۱$ رأس منزوی.

قضیه ۳ : فرض کنید C_n تعداد مؤلفه‌ها در یک گراف تصادفی بازه‌ای $\mathcal{IG}_{n,d}$ باشد، توزیع احتمال گسستهٔ متغیر تصادفی C_n با رابطه زیر داده می‌شود.

$$P(C_n = r) = \sum_{j=r-۱}^{\min\{n-۱,\lfloor ۱/d\rfloor\}}\binom{n-۱}{j}\binom{j}{r-۱}(-۱)^{j+r-۱}(۱-jd)^n$$

برای $r = ۱, ۲, \ldots, \min\{n-۱, \lfloor ۱/d\rfloor\} + ۱$. t امین گشتاور فاکتوریل برای تعداد مؤلفه‌های کاهش یافته به اندازهٔ ۱ عبارت است از

$$E_t(C_n - ۱) = (n-۱)_t(۱-td)^n$$

هرگاه $۱ - td > ۰$، و در غیر این صورت صفر می‌شود.
برهان. این قضیه با استفاده از روش فاصله‌گذاری اثبات می‌شود. n نقطه، بازه را به $n-$

۱ زیربازهٔ متوالی بین کوچکترین نقطه و بزرگترین نقطه افراز می‌کنند که طولهای آن زیربازه‌ها فاصله گذاری ها نامیده می‌شوند (باربور، هولست و جانسون (۱۹۹۲) را ببینید). $n-۱$ متغیرهای $۰-۱$ (نشانگرها) $I_۱$، $I_۲$، ...، $I_{n-۱}$ را به گونه‌ای تعریف کنید که برای $i = ۱, ۲, \ldots, n-۱$،

$$I_i = \begin{cases} ۱ & \text{اگر } i \text{ امین فاصله‌گذاری از } d \text{ تجاوز کند} \\ ۰ & \text{در سایر جاها} \end{cases}$$

به آسانی می‌توان دید که

$$C_n - ۱ = I_۱ + I_۲ + \cdots + I_{n-۱}.$$

فرض کنید $j < [۱/d]$. پیشامدی را در نظر بگیرید که j فاصله‌گذاری مفروض، مثلاً فاصله‌گذاری $i_۱$، $i_۲$، ...، i_j، از d بزرگتر باشند. بدیهی است که، در این مدل احتمال این پیشامد به انتخاب $i_۱$، $i_۲$، ...، i_j بستگی ندارد. می‌توان n نقطه به طور یکنواخت و مستقل از زیربازهٔ به طول $۱ - jd$ با احتمال $(۱ - jd)^n$ استخراج کرد. سپس فاصله گذاریهای $i_۱$، $i_۲$، ...، i_j را، هر کدام به اندازهٔ d بسط می‌دهیم که تمامی توزیعهای ممکن n نقطه در بازهٔ واحد را که برای آن فاصله‌گذاریهای $i_۱$، $i_۲$، ...، i_j، بزرگتر از d هستند را خواهد داد. بنابراین، احتمال پیشامدی که در نظر گرفته می‌شود، عبارت است از

$$P(I_{i_۱} = ۱, I_{i_۲} = ۱, \ldots, I_{i_j} = ۱) = (۱ - jd)^n.$$

حال برهان قضیه ؟؟ با استفاده از کاربرد متداول فرمول رد و قبول کامل می‌شود (بلاباش(۱۹۸۵) را ببینید).

نتیجه ۲ :احتمال آنکه یک گراف تصادفی بازه‌ای $\mathcal{IG}_{n,d}$ همبند باشد عبارت است از

$$P(\text{همبند باشد } \mathcal{IG}_{n,d}) = \sum_{j=۰}^{\min\{n-۱,[۱/d]\}} \binom{n-۱}{j} (-۱)^j (۱ - jd)^n,$$

که برای $d < ۱/(n-۱)$ به صورت زیر کاهش می‌یابد

$$P(\text{همبند باشد } \mathcal{IG}_{n,d}) = n! d^{n-۱} (۱ - \frac{۱}{۲}(n-۱)d).$$

برهان. چون

$$P(\text{همبند باشد } \mathcal{IG}_{n,d}) = P(C_n = ۱),$$

اولین قسمت نتیجه یک پیامد مستقیمی از قضیه ؟؟ است و نیز می‌توان آن را با استفاده از لم ؟؟ نیز اثبات کرد ($P(\mathcal{CS}_n)$).

برای $d < ۱/(n-۱)$، کران بالا در مجموع به $n-۱$ کاهش می‌یابد. در نتیجه

$$P(C_n = ۱) = d^n \sum_{j=۰}^{n-۱} \binom{n-۱}{j} (-۱)^j (\frac{۱}{d} - j)^n$$
$$= -d^n \sum_{j=۰}^{n-۱} \binom{n-۱}{j} (j - \frac{۱}{d})^{j+۱} (\frac{۱}{d} - n + ۱ + n - ۱ - j)^{n-۱-j}$$

با استفاده مجدد از نماد ری یوردان برای جمعهای آبل (اثبات لم ؟؟ را ببینید) و اتحادهای دیگر آبل، می‌توان احتمال همبندی را به صورت زیر دوباره نویسی کرد.

$$P(C_n = ۱) = -d^n A_{n-۱}(-۱/d, ۱/d - n + ۱; ۱, ۰) = -d^n(\alpha + \beta(-۱/d))^{n-۱}$$
$$= -d^n \sum_{i=۰}^{n-۱} \binom{n-۱}{i} i!(n-۱-i)!(n-۱-i-۱/d)$$
$$= -d^n \sum_{i=۰}^{n-۱} (n-۱)!(n-۱-i-۱/d) = n! d^{n-۱} \left(۱ - \frac{۱}{۲}(n-۱)d\right),$$

که α و $\beta(x)$ متغیرهای ظاهری هستند به طوری که

$$(\beta(x))^i = \beta_i(x) = i!(x+i) \qquad \text{و} \qquad \alpha^i = \alpha_i = i!$$

(ری یوردان (۱۹۶۸) را ببینید). □

یکی از نتایج مستقیم لم ؟؟ فرمولی برای مقدار مورد انتظار تعداد C_n^k مؤلفه به اندازه k است. با شیوه‌ای نظیر لم ؟؟ و با به کار بردن همان اتحادهای دو جمله‌ای قضیه بعدی به دست می‌آید.

قضیه ۴ : فرض کنید C_n^k تعداد مؤلفه‌های به اندازه k در گراف تصادفی بازه‌ای باشد. آنگاه، تعداد مورد انتظار مؤلفه‌های به اندازه بزرگتر از m به صورت زیر داده می‌شود

$$\sum_{k=m+۱}^{n} E(C_n^k) = \sum_{j=۰}^{\min\{m+۱,[۱/d]\}} \binom{m+۱}{j} (-۱)^j (۱ - jd)^n$$
$$+(n-m) \sum_{j=۰}^{\min\{m,[۱/d]-۱\}} \binom{m}{j} (-۱)^j (۱ - (j+۱)d)^n.$$

برای به دست آوردن توزیع احتمال دقیق تعداد مؤلفه‌هایی با یک اندازهٔ مفروض ابتدا حالت خاص رأسهای منزوی را در نظر می‌گیریم.

قضیه ۵ : توزیع تعداد $C_n^{(1)}$ رأسهای منزوی در یک گراف تصادفی بازه‌ای $\mathcal{IG}_{n,d}$ به صورت زیر داده می‌شود

$$P(C_n^{(1)}=r)=\sum_{k=r}^{\min\{n-1,[1/d]\}}\binom{k}{r}(-1)^{k-r}\sum_{j=j_0}^{k}\binom{n-k-1}{k-j}\binom{k+1}{j}\times(1-(2k-j)d)^n+\binom{n}{r}(-1)^{n-r}(1-(n-1)d)_+^n$$

با $j_0=\max\{0,2k-[1/d],2k-n+1\}$، که برای a مثبت، $a_+=a$ و در غیر این صورت $a_+=0$. به علاوه، t امین گشتاور فاکتوریل برای تعداد رأسهای منزوی به صورت زیر داده می‌شود

$$E_t(C_n^{(1)})=t!\sum_{j=j_0}^{t}\binom{n-t-1}{t-j}\binom{t+1}{j}(1-(2t-j)d)^n$$

با $j_0=\max\{0,2t-[1/d],2t-n+1\}$ هرگاه $1-td>0$، و در غیر این صورت صفر می‌شود.

قضیه ۶ : t امین گشتاور فاکتوریل برای تعداد مؤلفه‌های همبند به اندازه k به صورت زیر داده می‌شود

$$E_t(C_n^k)=t!\sum_{j=j_0}^{t}\binom{n-kt-1}{t-j}\binom{t+1}{j}\sum_{i=0}^{(k-1)t}\binom{(k-1)t}{i}(-1)^i\times(1-(2t-j+i)d)_+^n$$

با $j_0=\max\{0,t(k+1)-n+1\}$ و برای a مثبت، $a_+=a$ و در غیر این صورت $a_+=0$.

۴ مفاهیم نظریه گراف در فضای k بعدی

فرض کنید $F_{\underline{X}}$ یک تابع توزیع تجمعی روی فضای اقلیدسی k بعدی، E_k، و نسبت به اندازه لبگ k بعدی مطلقاً پیوسته باشد. تابع چگالی احتمال متناظر را با $f_{\underline{X}}$ نشان می‌دهیم. فرض کنید یک نمونه تصادفی به اندازه n از $F_{\underline{X}}$ به دست آمده است و یافته‌ها را با $\underline{x}_1$، ...، $\underline{x}_n$ نشان دهید.

حال گراف $G_n = (V, E)$ چنین تعریف می‌شود: V یک مجموعه با $|V| = n$ و E یک مجموعه از جفتهای (بی‌ترتیب) عناصر V است. عناصر V رأسهای گراف G_n و جفتها در E به عنوان یالهای G_n نامیده می‌شوند. بدون از دست دادن کلیت، می‌توانیم فرض کنیم $V = \{1, 2, \ldots, n\}$. برای اهداف مد نظر، یک فاصله ρ روی E_k و یک آستانه $d > 0$ انتخاب می‌کنیم. بنابراین برای $i \neq j$، (i, j) را در E قرار دهید اگر و تنها اگر $\rho(\underline{x}_i, \underline{x}_j) \leq d$. چون $\underline{x}_1$، $\underline{x}_2$، ...، $\underline{x}_n$ تحققهای متغیرهای تصادفی هستند، مجموعه E یک مجموعه تصادفی و گراف G_n یک گراف تصادفی است. به ویژه، این گرافها تعمیمهایی از گرافهای بازه‌ای هستند. مشخصاً اگر I_1، I_2، ...، I_n بازه‌هایی روی خط حقیقی باشند، آنگاه گراف بازه‌ای $G(I_n)$ به وسیله $V = \{1, 2, \cdots, n\}$ و $(i, j) \in E$ اگر $1 \leq i < j \leq n$ و $I_i \cap I_j \neq \emptyset$، تعریف می‌شوند. بنابراین برای مدل مورد بررسی، اگر $k = 1$، بازه‌های I_i، بازه‌های $[x_i - \frac{d}{2}, x_i + \frac{d}{2}]$، $i = 1, 2, \cdots, n$ هستند.

تعریف ۱: فرض کنید $V_m \subset V$ با $|V_m| = m < n$. $K_{m,d}$ یک زیرگراف کامل از مرتبه m است، اگر همهٔ $\binom{m}{2}$ جفت عناصر V_m در E باشند. اگر $m = 1$، آنگاه $K_{1,d}$ یک رأس است، اگر $m = 2$، آنگاه $K_{2,d}$ یک یال است و اگر $m = 3$، آنگاه $K_{3,d}$ یک مثلث نامیده می‌شود.

تعریف ۲: یک زیرگراف کامل از مرتبه m یک زیرگراف کامل بیشینه نامیده می‌شود، و با $K^*_{m,d}$ نشان داده می‌شود، هرگاه رأسی در $V \setminus V_m$ موجود نباشد به طوری که از متصل کردن آن رأس به V_m یک زیرگراف کامل از مرتبه $m + 1$ منتج شود.

درجهٔ رأس v، $v = 0, 1, 2, \cdots, n - 1$، در G_n تعداد یالهای G_n است که v بر آنها واقع است. اگر $v = 0$، آنگاه آن رأس منزوی نامیده می‌شود.

۵ توزیعهای احتمال برای مشخصه‌های حقیقی گرافهای بازه‌ای

حال احتمال آن را که یک مجموعه معین از m رأس، یک $K_{m,d}$ تشکیل دهند توصیف می‌کنیم. بدون از دست دادن کلیت، فرض می‌کنیم که این رأسها با ۱، ۲، ...، m برچسب

گذاری می‌شوند.
احتمال آنکه یک مجموعه معین از m رأس، یک $K_{m,d}$ تشکیل دهند به صورت زیر است.

$$P\{\max_{1\le i\le m} X_i - \min_{1\le i\le m} X_i \le d\} = m\int_{-\infty}^{\infty}\{F(x+d)-F(x)\}^{m-1}f(x)dx. \quad (1)$$

احتمال آنکه یک رأس معین دارای درجهٔ v، $v = 0, 1, \cdots, n-1$ باشد عبارت است از

$$P\{\text{درجهٔ } v \text{ رأس } 1 \text{ دارد}\} = \binom{n-1}{v}\int_{-\infty}^{\infty}\{F(x+d)-F(x-d)\}^{v} \quad (2)$$
$$\times\{1-F(x+d)+F(x-d)\}^{n-v-1}f(x)dx.$$

نتیجه ۳ : احتمال آنکه یک رأس معین یک $K^*_{1,d}$ تشکیل دهد (یعنی منزوی باشد) عبارت است از

$$P\{\text{رأس } 1 \text{ منزوی است}\} = \int_{-\infty}^{\infty}\{1-F(x+d)+F(x-d)\}^{n-1}f(x)dx. \quad (3)$$

با توجه به تعریف رأس منزوی اگر در فرمول (۲)، $v = 0$ را قرار دهیم رابطه (۳) به دست خواهد آمد.

۱.۵ خواص مجانبی گرافهای تصادفی بازه‌ای

برای به دست آوردن تقریب‌های مجانبی برای توزیعهای بالا، به فرضهایی دربارهٔ رفتار تابع چگالی احتمال $f_X(x)$ نیاز داریم. از این رو فرض می‌کنیم که تابع چگالی احتمال روی هر زیر مجموعه فشرده از مجموعهٔ حامل X به طوریکنواخت پیوسته است و فرض می‌کنیم که $f'_X(x)$ موجود و روی مجموعهٔ حامل X به طوریکنواخت کراندار است. حال رفتار مجانبی توزیعهای احتمال معرفی شده در بخش قبلی را، تحت شرایط $n \to \infty$ و $d \to 0$ (معمولاً، d تابعی مفروض از n است، $d = d(n)$) آزمون می‌کنیم. همچنین شرایط نظم را برای $f_X(x)$ داده شده در بالا در نظر می‌گیریم. فرض کنید $(A_t)_{t\in T}$ خانواده‌ای از پیشامدها باشد که به پارامتر t وابسته است. می‌گوییم احتمال مجانبی A_t، برای $t \to t'$، p است هرگاه $\lim_{t\to t'} P\{A_t\}/p = 1$.

احتمال مجانبی ($d \to 0$ هرگاه $n \to \infty$) آنکه رأسهای ۱، ۲، ...، m یک $K_{m,d}$ تشکیل دهند،

$$md^{m-1}\int_{-\infty}^{\infty} f^m(x)dx \quad (4)$$

است.

میانگینها و واریانسهای مجانبی تعداد زیرگرافهای کامل از مرتبه m به صورت زیر داده می‌شوند.

$$E\{|K_{m,d}|\} \sim \binom{n}{m} P\{K_{m,d}\} = \binom{n}{m} m d^{m-1} \int_{-\infty}^{\infty} f^m(x)dx \qquad (۵)$$

و

$$Var\{|K_{m,d}|\} \sim \binom{n}{۲m-۱} (۲d)^{۲m-۲} \int_{-\infty}^{\infty} f^{۲m-۱}(x)dx. \qquad (۶)$$

توجه کنید که هرگاه $d \to ۰$، احتمال آنکه رأسهای معین یک $K_{m,d}$ تشکیل دهند به صفر میل می‌کند. فرمول (۴) همچنین به صفر میل می‌کند، اما نسبت آنها به سمت واحد میل می‌کند.

احتمال مجانبی ($d \to ۰$ هرگاه $n \to \infty$، به طوری که $nd \to ۰$) آنکه رأس ۱ دارای درجه v باشد

$$\binom{n-۱}{v} (۲d)^v \int_{-\infty}^{\infty} f^{v+۱}(x)\{۱ - ۲ndf(x)\}dx \qquad (۷)$$

است.

۲.۵ آشکارسازی آمیزه‌های توزیعهای احتمال

یک مسئله که دقیقاً به آشکارسازی خوشه‌ها مربوط می‌شود، آشکارسازی آمیزه‌های توزیعهای احتمال است. یک آمیزه از k تابع چگالی احتمال یک تابع چگالی احتمال به شکل

$$f_X(x) = \sum_{i=۱}^{k} \alpha_i f_i(x),$$

است که در آن $\alpha_i > ۰$ و $\sum_{i=۱}^{k} \alpha_i = ۱$. به منظور اجتناب از بعضی پیچیدگیهای ریاضی، آن توزیعهای احتمالی را به کار می‌بریم که تابع چگالی احتمالشان $f_i(x)$ بوده و دو به دو مطلقاً پیوسته و متمایز باشند. (یعنی یک مجموعه $A_(i,j)$ از اندازهٔ لبگ مثبت وجود دارد که برای آن $\int_{A(i,j)} f_i(x)dx \neq \int_{A(i,j)} f_j(x)dx$ برای $i \neq j$، $۱ \leq i < j \leq k$). فرض کنید $X_۱$، $X_۲$، ...، X_n یک نمونه تصادفی از $f_X(x)$ است و فرض کنید $d > ۰$ یک آستانه معین باشد. تقریبات مجانبی که می‌بایست اینجا به کار روند تحت فرض آنکه $d \to ۰$ و $n \to \infty$ معتبر هستند، به طوری که $n^m d^{m-۱} \to \infty$. بنابراین،

برای مثال، احتمال مجانبی آنکه m تحقق به طور تصادفی انتخاب شده به یک زیرگراف کامل از مرتبه m منجر شود به صورت

$$P\{K_{m,d}\} \sim md^{m-1} \sum_{j_1,\cdots,j_k} \binom{m}{j_1,\cdots,j_k} \prod_{i=1}^{k} \left\{ {\alpha_i}^{j_i} \int_{-\infty}^{\infty} {f_i}^{j_i}(x)dx \right\},$$

است، که در آن $j_i \geq 0$، $i = 1,\cdots,k$ و $\sum_{l=1}^{k} j_l = m$ می‌باشد.

هریس و گودهارت(۱۹۹۸) روشی را به شرح زیر پیشنهاد می‌کنند:

ابتدا فرض می‌شود که آمیزه‌ای وجود ندارد، یعنی، $k = 1$، $\alpha_1 = 1$. مقدار مورد انتظار و واریانس $K_{m,d}$ تحت این فرض برآورد شده و تعداد مشاهده شدهٔ زیرگرافهای کامل از مرتبه m با این میانگین مقایسه می‌شود. اگر این عدد اساساً (نسبت به واریانس) از مقدار مورد انتظار برآورد شده متفاوت باشد، آنگاه نتیجه خواهد شد که یک آمیزهٔ نابدیهی وجود دارد. می‌توان با تکرار این عمل فرض سازگاری داده‌ها را با آمیزه‌ای از دو توزیع آزمون نمود.

آزمونهای دنباله‌ای توصیف شده به طور تصادفی مستقل نیستند، هرچند که بیشتر روشهای عملی تحلیل داده‌های آماری در حقیقت به همین روش انجام می‌شوند.

۳.۵ آمیزه‌هایی از توزیعهای نمایی

کاربرد زیر از این روشها احتمالاً مفهوم ضعیفی در تحلیل خوشه‌ای دارد، اما در نظریهٔ قابلیت اعتماد و تحلیل ریسک به طور طبیعی بروز می‌کند. ما در اینجا، با به کار بردن آزمونهایی بر اساس نظریه زیرگرافهای کامل از مرتبهٔ m، آشکار کردن وجود آمیزه‌هایی از توزیعهای نمایی را بررسی می‌کنیم. یک روش طبیعی که می‌توان مطرح کرد استفاده کردن از آزمون نسبت درستنمایی و نظریه مجانبی متناظر است. متأسفانه در مورد آمیزه‌ها، شرایط نظم مورد نیاز برای نظریه مجانبی نسبت درستنمایی برقرار نیست. بنابراین روش کنونی ممکن است راه حل مناسبی باشد.

در نتیجه فرض کنید

$$f_i(x) = \lambda_i e^{-\lambda_i x}, \lambda_i > 0, x > 0, i = 1, 2, \ldots, k.$$

از این رو به دست می‌آوریم

$$P\{K_{m,d}\} \sim md^{m-1} \sum_{j_1,\cdots,j_k} \binom{m}{j_1,\cdots,j_k} \prod_{i=1}^{k} \left\{ {\alpha_i}^{j_i} {\lambda_i}^{j_i} \Big/ \sum_{l=1}^{k} j_l \lambda_l \right\}.$$

یک مورد خاص در نظر گرفته می‌شود. به منظور سادگی محاسبات در این مثال فرض می‌کنیم $k = 2$ و $m = 2$؛ یعنی یک آمیزه از دو توزیع نمایی در نظر می‌گیریم و تعداد

یالهای تحقق یافته را می‌شماریم.

مثال ۱: فرض کنید X طول عمر تصادفی یک دستگاه و دارای توزیع نمایی با شدت λ باشد. پس از مدتی گمان می‌شود که منبع تولید دومی با نرخ شدت متفاوتی معرفی شده است. بنابراین ممکن است برخی از داده‌ها از منبع تولید اول و برخی دیگر از داده‌ها از منبع تولید دوم آمده باشند، و بنابراین داده‌های مشاهده شده ممکن است از یک آمیزه از دو توزیع نمایی باشند. بدون از دست دادن کلیت مسئله، می‌توانیم فرض کنیم $\lambda_1 = ۱$. برای ارائهٔ یک مثال عددی، ۳۰۰ مشاهده شبیه‌سازی شدند، ۱۸۳ مشاهده از یک توزیع نمایی با شدت $\lambda_1 = ۱$ و ۱۱۷ مشاهده از توزیع نمایی با $\lambda_2 = ۲$، از این رو $\alpha_1 = ۰/۶۱$ و $\alpha_2 = ۰/۳۹$. داده‌های طول عمر شبیه‌سازی شده یک میانگین نمونه‌ای $\bar{x} = ۰/۷۸۰$ و یک واریانس نمونه‌ای $s^2 = ۰/۷۲۹$ دارند. آستانه d در سطح ۰/۰۱ اختیار شده است (البته این مقدار عمدتاً بستگی به نظر پژوهشگر دارد) که به تعداد ۵۶۳ یال مشاهده شده منتج می‌شود. در مجموع، گشتاور دوم نمونه‌ای طول عمرها $m_2 = ۱/۳۳۵$ و گشتاور سوم نمونه‌ای $m_3 = ۳/۶۲۸$ است. برای این آمیزه خاص، میانگین نظری طول عمر $\mu = ۰/۸۰۵$ و واریانس نظری $\sigma^2 = ۰/۷۶۷$ است. از این رو داده شبیه‌سازی شده توافق معقولی با نظریه دارد.

برای مثال ۱، فرض کردیم که $\lambda_1 = ۱$ معلوم و λ_2 نامعلوم است. فرض صفر طبیعی H_0، $\alpha = ۱$ است، یعنی، که هیچ آمیزشی موجود نیست. اگر این درست باشد، آنگاه مقدار مجانبی برای تعداد مورد انتظار یالها ۴۴۸/۵ است، به همین نحو، مقدار مجانبی برای واریانس تعداد یالها تحت فرض صفر ۵۹۴/۰۱ است که یک انحراف استاندارد ۲۴/۴ را نشان می‌دهد. بنابراین می‌توان تشخیص داد که تعداد مشاهده شده یالها به اندازه کافی بزرگ است تا فرض صفر رد شود. برای آمیزه خاص به کار رفته، تعداد مورد انتظار یالها ۵۸۷/۸ است. بنابراین واضح است که برای مثال ۱، آشکارسازی وجوه یک آمیزه، را می‌توان با به کار بردن تعداد یالها انجام داد. شخص می‌بایست α و λ_2 را از روی داده‌ها برآورد کند. چندین روش وجود دارد که شخص ممکن است به کار برد. نوعاً شخص داده طول عمر را مورد استفاده قرار می‌دهد که یکی از فنهای برآورد آماری استاندارد نظیر روش حداکثر درستنمایی یا روش گشتاورها را به کار می‌گیرد. به دلیل سادگی محاسباتی، روش گشتاورها مورد استفاده قرار گرفته، که به برآوردهای $\alpha = ۰/۵۵$ و $\lambda_2 = ۱/۹۶$ منجر می‌شود.

مراجع

Barbour, A. D., Holst, L. and Janson, S., (1992). Poisson Approximation, Clarendon Press, Oxford.

Bollobas, B., (1985). Random Graphs, Academic Press, New York.

Gilbert, E. N., (1959). Random Graphs, *Ann. Math. Stat., 30*, 1141-1144.

Gilmore, P. and Hoffman, A., (1964). A Characterization of Comparability Graphs and of Interval Graphs, *Canad. J. Math. 16*, 539-548.

Godehardt, E., (1990). Graphs as Structural Models, *Vieweg, Braunschweig.*

Godehardt, E. and Jaworski, J., (1996). On the Connectivity of a Random Interval Graph, *Random Structures and Algorithms, Vol. 9, Nos. 1 and 2*, 137-161.

Golumbic, M. C., (1980). Algorithmic Graph Theory and Perfect Graphs , (Academic Press).

Gould, H. W., (1972). Combinatorial Identities, *Morgantown Printing and Binding, Morgantown, WV.*

Harris, B. and Godehardt, E., (1998). Probability Models and Limit Theorems for Random Interval Graphs with Applications to Cluster Analysis, *Classification, Data Analysis, and Data Highways.* Springer, Berlin-Heidelberg-New York, 54-61.

Harris, B. and Godehardt, E., (1999). The Comparative Efficacy of Some Combinatorial Tests for Detection of Clusters and Mixture of Probability Distributions, *Classification in the Information Age.* Springer, 295-301.

Riordan, J., (1968). Combinatorial Identities, Wiley, New York.

Scheinerman, E. R., (1990). An Evolution of Interval Graphs, *Discrete Math., 82*, 287-302.

Scheinerman, E. R., (1988). Random Interval Graphs, *Combinatorica 8*, 357-371.

نمایه کلید واژه