محاسبهٔ فاصله‌های تحمل برای متغیر تصادفی دوجمله‌ای

مهران نقی‌زاده قمی[۱]، آزیتا نوروزی فیروز[۲]

تاریخ دریافت: ۱۳۹۳/۱۲/۱۷
تاریخ پذیرش: ۱۳۹۵/۷/۱۷

چکیده:
فاصلهٔ تحمل، یک فاصلهٔ تصادفی است که با یک سطح اطمینان مشخص، نسبتی از جامعهٔ مورد بررسی را در برمی‌گیرد و در بسیاری از زمینه‌های کاربردی از جمله قابلیت اعتماد و کنترل کیفیت، مورد استفاده قرار می‌گیرد. در این مقالهٔ آموزشی، روش‌های مختلف محاسبهٔ فاصلهٔ تحمل برای متغیر تصادفی دوجمله‌ای را با استفاده از بسته $tolerance$ در نرم‌افزار آماری R مورد بررسی قرار می‌دهیم.

واژه‌های کلیدی: توزیع دوجمله‌ای، فاصلهٔ اطمینان، فاصلهٔ تحمل.

۱ مقدمه

معمولاً برای برآورد پارامتر نامعلوم جامعه از یک فاصلهٔ اطمینان و برای پیشگویی یک مشاهدهٔ آینده، از فاصلهٔ پیشگویی بر اساس نمونهٔ تصادفی استفاده می‌شود. نوع سوم از این فواصل، فاصلهٔ تحمل[۳] است. فاصلهٔ تحمل، نسبتی از جامعهٔ مورد بررسی (p) را با یک سطح اطمینان مشخص $(1-\alpha)$ در بر می‌گیرد که معمولاً برای سادگی، آن را فاصلهٔ تحمل $(p, 1-\alpha)$ گویند. فاصله‌های تحمل به‌طور گسترده‌ای در کنترل کیفیت، مهندسی و صنایع دارویی مورد استفاده قرار می‌گیرند. به‌عنوان مثال فرض کنید یک نمونهٔ تصادفی از n=۱۰۰۰ لامپ که در بسته‌های ۵۰ تایی هستند برای بازرسی انتخاب شود که از این میان، تعداد x=۱۰۰ لامپ معیوب است. فرض کنید مهندس کارخانهٔ تولیدکنندهٔ لامپ علاقه‌مند است مشخص کند که با ۹۵ درصد اطمینان، حداقل ۹۰ درصد از بسته‌های بعدی به اندازهٔ m=۵۰ لامپ، دارای لامپ‌های معیوب بین L و U هستند. فاصلهٔ $[L, U]$ یک فاصلهٔ تحمل (۰.۹۵ و ۰.۹) نامیده می‌شود. در این‌جا، m اندازهٔ نمونهٔ بعدی یا آینده[۴] نامیده می‌شود. یا در مثالی دیگر فرض کنید مهندسی نیاز دارد تا یک حد بالای تحمل (U) برای تعداد خاموش شدن‌های ناگهانی یک سیستم پیچیده به دست آورد. به‌عنوان مثال با ۹۵ درصد اطمینان، حداقل ۸۰ درصد از سیستم‌هایی که در طول یک سال کار می‌کنند دارای تعداد خاموش شدن ناگهانی حداکثر U هستند. در بحث پزشکی و آزمایشگاهی، مفهوم فاصلهٔ تحمل، انتظاری است که پزشکان، محققان آزمایشگاهی و پژوهشگران مسائل زیستی در آزمایش‌های کلینیکی دارند. آن‌ها می‌خواهند بدانند که بیشترین مشاهده‌های آن‌ها در یک موضوع تحقیق از جامعهٔ مورد مطالعه، با چه احتمالی و در چه فاصله‌ای قرار دارد و یا از نظر آماری، احتمال این‌که نسبت معینی از افراد جامعه که دارای صفت معینی هستند و در این فاصله قرار می‌گیرند چه‌قدر است؟ (نائینی [۱]). برای مثال‌های مرتبط، به هان و میکر [۸] و کریشنامورتی و متیو [۹] مراجعه کنید. فاصله‌های تحمل مانند فاصله‌های اطمینان و فاصله‌های پیشگویی، برای متغیرهای تصادفی پیوسته و گسسته، داده‌های یک‌متغیره و چندمتغیره، پارامتری و ناپارامتری، و برای مدل‌های خطی عمومی مورد استفاده قرار می‌گیرند (کریشنامورتی و متیو [۹]). در این مقاله متغیر تصادفی گسستهٔ دوجمله‌ای را در نظر گرفته، نحوهٔ محاسبهٔ فاصلهٔ تحمل را بیان می‌کنیم. در حالت کلی، فرض کنید X یک متغیر تصادفی گسسته با تابع توزیع تجمعی F باشد. فاصلهٔ $[L(X), U(X)]$ را یک فاصلهٔ تحمل دوطرفه با

[۱] استادیار گروه آمار، دانشگاه مازندران
[۲] دانش‌آموختهٔ کارشناسی ارشد آمار، دانشگاه مازندران

[۳] tolerance interval
[۴] future sample size

میزان پوشش p و میزان اطمینان $1-\alpha$ گویند هرگاه

$$Pr\{F(L(X)-F(U(X))\geq p\}\geq 1-\alpha. \qquad (1)$$

حدود تحمل یک‌طرفه به‌طور مشابه تعریف می‌شوند. برای محاسبهٔ حد پایین تحمل $(p,1-\alpha)$، بزرگ‌ترین مقدار صحیح $L_1(X)$ را به‌گونه‌ای می‌یابیم که در رابطهٔ زیر صدق کند:

$$Pr\{1-F(L_1(X))\geq p\}\geq 1-\alpha. \qquad (2)$$

مفهوم $L_1(X)$ این است که با اطمینان $1-\alpha$ حداقل نسبت p از جامعه از $L_1(X)$ بزرگ‌تر هستند. همچنین حد بالای تحمل $(p,1-\alpha)$ نیازمند یافتن کوچک‌ترین مقدار صحیح $U_1(X)$ است به‌طوری که

$$Pr\{F(U_1(X))\geq p\}\geq 1-\alpha.$$

در ادبیات تحقیق، به مسئلهٔ یافتن فاصله‌های تحمل برای متغیرهای تصادفی گسسته در مقایسه با متغیرهای تصادفی پیوسته کمتر پرداخته شده است. زاکس [۱۷] به‌طور یکنواخت دقیق‌ترین حدود بالای تحمل[۵] را برای توزیع‌های گسستهٔ دارای خاصیت نسبت درست‌نمایی یکنوا[۶] به دست آورد. هان و چاندرا [۷] با استفاده از روشی که بسیار فراگیر شد، فاصله‌های تحمل را برای توزیع‌های دوجمله‌ای و پوآسون مورد مطالعه قرار دادند. وانگ و تسونگ [۱۲] با استفاده از مینیمم و میانگین احتمال‌های پوشش، فاصله‌های تحمل در توزیع‌های دوجمله‌ای و پوآسون را بهبود بخشیدند. کریشنامورتی و همکاران [۱۰] یک روش تقریبی برای به دست آوردن فاصله‌های تحمل در این توزیع‌ها ارائه کردند. یانگ [۱۵و۱۶] و نقی‌زاده قمی و همکاران [۱۱] به‌ترتیب به ارائه فاصلهٔ تحمل برای توزیع‌های دوجمله‌ای منفی، فوق‌هندسی (فوق‌هندسی منفی) و پوآسون-لیندلی پرداختند. در این مقاله، نخست روش کلی به دست آوردن فاصله‌های تحمل برای متغیرهای تصادفی گسسته بیان می‌شود. سپس فاصله‌های اطمینان مطرح شده برای پارامتر نسبت توزیع دوجمله‌ای ارائه می‌شود. در پایان با ذکر مثالی و با استفاده از بستهٔ *tolerance* در نرم‌افزار آماری R نسخهٔ ۳.۱.۲ فاصله‌های تحمل را به دست می‌آوریم.

۲ روش عمومی برای ساختن فاصله‌های تحمل با دم‌های برابر

فرض کنید متغیر تصادفی X دارای تابع توزیع $F_X(x|\theta,n)$ با پارامتر نامعلوم θ و اندازهٔ نمونهٔ معلوم n باشد. همچنین فرض کنید دارای خاصیت «به‌طور تصادفی صعودی»[۷] باشد؛ یعنی به‌ازای هر t و هر $\theta_1>\theta_2$ داشته باشیم:

$$Pr(X>t|\theta_1)\geq Pr(X>t|\theta_2)$$

هان و چاندرا [۷] روش دومرحله‌ای زیر را برای به دست آوردن فاصله‌های تحمل با دم‌های برابر معرفی کردند:

۱) بر اساس نمونهٔ مشاهده‌شدهٔ x ،یک فاصلهٔ اطمینان $100(1-\alpha)$ درصدی دوطرفهٔ $(\theta_L(x;n),\theta_U(x;n))$ برای θ به دست می‌آوریم.

۲) برای یک اندازهٔ نمونهٔ آیندهٔ m، بزرگ‌ترین مقدار صحیح $L(x;m)$ و کوچک‌ترین مقدار صحیح $U(x;m)$ را طوری می‌یابیم که:

$$1-F(L(x;m)|\theta_L(x;n))\geq \frac{1+p}{2}, \qquad (3)$$

و

$$F(U(x;m)|\theta_U(x;n))\geq \frac{1+p}{2}. \qquad (4)$$

به‌طور مشابه می‌توان حدود تحمل یک‌طرفه را به دست آورد. برای این کار، در ابتدا فاصله‌های اطمینان $100(1-\alpha)$ درصدی یک‌طرفه برای θ، یعنی $\theta_{L_1}(x;n)$ و $\theta_{U_1}(x;n)$ را یافته، به جای $\frac{1+p}{2}$ مقدار p را قرار می‌دهیم. توزیع دوجمله‌ای که در این مقاله مورد بحث قرار می‌گیرد، نسبت به پارامتر نسبت p به‌طور تصادفی صعودی است (نوروزی فیروز [۲]). بنا بر این اگر $(\theta_L(x;n),\theta_U(x;n))$ یک فاصلهٔ اطمینان $100(1-\alpha)$ درصدی دوطرفه برای θ باشد، می‌توان با توجه به دومرحلهٔ گفته‌شده، فاصلهٔ تحمل را به دست آورد.

[۵] uniformly most accurate upper tolerance limits

[۶] monotone likelihood ratio

[۷] stochastically increasing

و

$$\theta_u = B\Big(1-\frac{\alpha}{2}; x+a_1+1, n-x+a_2\Big)$$

که در آن a_1 و a_2 پارامترهای توزیع پیشین بتا هستند. در حالت $a_1 = a_2 = 0/5$ فاصلهٔ اطمینان جفریز به دست می‌آید (براون و همکاران [۴]).

دورایی-راج [۶] نسخه جدید بستهٔ *binom* در نرم‌افزار R شامل محاسبهٔ فواصل اطمینان مختلف برای توزیع دوجمله‌ای و موارد مرتبط با آن را ارائه کرد. دستور محاسبهٔ فاصلهٔ اطمینان به‌صورت زیر است:

```
binom.confint(x,n,conf.level=0.95,
method="exact",...)
```

که در آن x و n به‌ترتیب بردار تعداد موفقیت‌ها و بردار تعداد آزمایش‌های مستقل در آزمایش دوجمله‌ای، *conf.level* سطح اطمینان، و *method* روش یافتن فاصلهٔ اطمینان است. در حالت بیزی، پارامترهای *prior.shape*1 و *prior.shape*2 نیز به دستور فوق اضافه می‌شوند.

مثال ۱.۳. وانگ و تسونگ [۱۰] داده‌های تراشه‌های کامپیوتری را مورد بررسی قرار دادند که از لینک زیر قابل دسترسی است:

http:..www.itl.nist.gov.div898.handbook
.pmc.section3.pmc332.htm

برای ۳۰ قطعهٔ سیلیکونی که بر روی آن مدارات مجتمع برای ایجاد یک تراشه قرار دارند، موقعیت ۵۰ تراشه اندازه‌گیری و نسبت تراشه‌های معیوب ثبت شد. وانگ و تسونگ برای محاسبهٔ فاصلهٔ تحمل، تعداد ۲۱ قطعهٔ سیلیکونی را در نظر گرفتند، که از بین آن‌ها تعداد تراشه‌های معیوب برابر با ۱۹۶ است. اندازهٔ نمونه هم برابر با n=۵۰×۲۱=۱۰۵۰ است. فاصلهٔ اطمینان ۹۵ درصدی دقیق (کلاپر-پیرسون) برای نسبت تراشه‌های معیوب عبارت‌اند از:

$$\theta_l = B(0/025, 196, 855) = 0/1635$$

و

$$\theta_u = B(0/975, 197, 854) = 0/2115.$$

۳ ساختن فاصله‌های تحمل در توزیع دوجمله‌ای

فرض کنید $X \sim Bin(n,\theta)$ مستقل از $Y \sim Bin(m,\theta)$ باشد. همان‌طور که در بخش قبل بیان شد، فاصله‌های تحمل با توجه به فاصله‌های اطمینان به دست می‌آیند.

۱.۳ فاصله‌های اطمینان برای پارامتر نسبت توزیع دوجمله‌ای

براون و همکاران [۴] فاصله‌های اطمینان موجود در ادبیات تحقیق برای نسبت توزیع دوجمله‌ای را از لحاظ احتمال پوشش آن‌ها بررسی کردند. همچنین فاصله‌های اطمینان جایگزین را معرفی و احتمال پوشش و طول آن‌ها را مورد بررسی قرار دادند. در این بخش به تعدادی از این فواصل اطمینان اشاره می‌کنیم.

- فاصلهٔ اطمینان والد برای نمونه‌های بزرگ:

$$(\theta_l, \theta_u) = \hat\theta \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat\theta(1-\hat\theta)}{n}}$$

- فاصلهٔ اطمینان ویلسون [۱۳] که به فاصلهٔ اطمینان امتیاز[۸] معروف است:

$$(\theta_l, \theta_u) = \left(\frac{\hat\theta + \frac{z^2_{\frac{\alpha}{2}}}{2n}}{1+z^2_{\frac{\alpha}{2}}}\right) \pm \frac{\frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\hat\theta(1-\hat\theta)\frac{z_{\frac{\alpha}{2}}}{4n}}}{1+\frac{z^2_{\frac{\alpha}{2}}}{n}}.$$

- فاصلهٔ اطمینان اگرستی-کول [۳]:

$$(\theta_l, \theta_u) = \tilde\theta \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\tilde\theta(1-\tilde\theta)}{n}},$$

که در آن $\tilde n = n + z^2_{\frac{\alpha}{2}}$ و $\tilde\theta = \frac{1}{\tilde n}\big(X + \frac{1}{2}z^2_{\frac{\alpha}{2}}\big)$.

- فاصلهٔ اطمینان دقیق کلاپر-پیرسون [۵]:

$$(\theta_l, \theta_u) = \Big(B\big(\tfrac{\alpha}{2}; x, n-x+1\big), B\big(1-\tfrac{\alpha}{2}; x+1, n-x\big)\Big),$$

که در آن $B(q;a,b)$ چندک qام توزیع بتا با پارامترهای شکل a و b است.

- فاصلهٔ اطمینان بیزی:

کران‌های پایین و بالای اطمینان عبارت‌اند از:

$$\theta_l = B\Big(\frac{\alpha}{2}; x+a_1, n-x+a_2\Big)$$

[۸] score confidence interval

دستور محاسبهٔ فاصلهٔ اطمینان ۹۵ درصدی دقیق (کلاپر-پیرسون) برای نسبت تراشه‌های معیوب در نرم‌افزار R به‌صورت زیر است:

```
###CI using clopper-pearson' method
binom.confint(x=196,n=1050,conf.level=0.95,
method="exact")
method x  n   mean  lower  upper
exact 196 1050 0.1867  0.1635 0.2115
```

۴ محاسبهٔ فاصلهٔ تحمل

یانگ [۱۴] بستهٔ $tolerance$ را برای محاسبهٔ فاصله‌های تحمل معرفی کرد که شامل بسیاری از توزیع‌های پیوسته و گسسته می‌شود. اخیراً نسخه ۲۰۱۶ این بسته نیز ارائه‌شده است. دستور محاسبهٔ فاصلهٔ تحمل برای توزیع دوجمله‌ای به‌صورت زیر است:

```
bintol.int(x,n,m=NULL, alpha=0.05, P=0.99,
side=2, method=c("LS", "WS", "AC", "JF",
 "CP"),a1=0.5,a2=0.5)
```

که در آن

x: تعداد موفقیت‌ها

n: تعداد آزمایش‌ها

m: اندازهٔ نمونهٔ بعدی یا آینده

$alpha$: سطح خطا به‌طوری‌که $alpha$ − ۱ سطح اطمینان است

P: سطح پوشش

بردار $method$ شامل LS، WS، AC، JF و CP به‌ترتیب فاصله‌های تحمل والد با نمونه‌های بزرگ، ویلسون، اگرستی-کول، جفریز و کلاپر-پیرسون می‌باشند.

a_1 و a_2 به‌ترتیب پارامترهای توزیع پیشین با استفاده از روش جفریز هستند.

در مثال ۱.۳ فاصلهٔ اطمینان ۹۵ درصدی برای نسبت تراشه‌های معیوب با استفاده از روش کلاپر-پیرسون را به‌صورت

$$(\theta_l, \theta_u) = (۰/۱۶۳۵, ۰/۲۱۱۵)$$

به دست آوردیم. برای محاسبه یک فاصلهٔ تحمل با میزان پوشش ۹۰ درصد و میزان اطمینان ۹۵ درصد برای تعداد تراشه‌های معیوب از بین ۵۰ تراشه، با استفاده از روابط (۳) و (۴) داریم

$$\sum_{i=۴}^{۵۰} \binom{۵۰}{i} \theta_l^i (۱-\theta_l)^{۵۰-i} = ۰/۹۷۲۹, \qquad (۵)$$

و

$$\sum_{i=۰}^{۱۵} \binom{۵۰}{i} \theta_u^i (۱-\theta_u)^{۵۰-i} = ۰/۹۵۱۱. \qquad (۶)$$

بنا بر این ۴ بزرگ‌ترین عدد صحیح است به‌طوری که عبارت (۵) حداقل برابر با $\frac{۱+p}{۲}$=۰.۹۵ و ۱۵ کوچک‌ترین عدد صحیح است به‌طوری که عبارت (۶) حداقل برابر با ۰.۹۵ است. در نتیجه (۱۵و۴) یک فاصلهٔ تحمل با میزان پوشش ۹۰ درصد و میزان اطمینان ۹۵ درصد است. به عبارت دیگر با اطمینان ۹۵ درصد می‌توان گفت که حداقل ۹۰ درصد از ۵۰ تراشهٔ بعدی (m=۵۰) دارای تعداد تراشه‌های معیوب بین ۴ و ۱۵ هستند. دستور محاسبهٔ فاصلهٔ تحمل میزان پوشش ۹۰ درصد و میزان اطمینان ۹۵ درصد با استفاده از روش کلاپر-پیرسون در نرم‌افزار R به‌صورت زیر است:

```
###TI using clopper-pearson' method
bintol.int(x=196,n=1050,m=50,alpha=0.05,
P=0.9,side=2, method="CP")
alpha P   p.hat 2-sided.lower 2-sided.upper
0.05 0.9 0.1867  4     15
```

مراجع

[۱] نائینی، م. (۱۳۸۳). حدود تحمل در آمار زیستی و آزمایش‌های کلینیکی، مجلهٔ علوم پزشکی مدرس، دورهٔ ۷، شمارهٔ ۲،صص ۹۳-۱۰۶.

[۲] نوروزی فیروز، آ. (۱۳۹۴)، محاسبهٔ فاصله‌های تحمل برای برخی از متغیرهای تصادفی گسسته، پایان‌نامهٔ کارشناسی ارشد آمار، دانشگاه مازندران .

[3] Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportion. *American Statistician*, **52**, 119–125.

[4] Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion), *Statistical Science*, **16**(**2**), 101-133.

[5] Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404–413.

[6] Dorai-Raj, S. (2015). https:..cran.r project.org.web.packages.binom.binom.pdf.

[7] Hahn, G. J. and Chandra, H. (1981). Tolerance intervals for Poisson and binomial random variables, *Journal of Quality Technology*, **13**, 100-110.

[8] Hahn, G. J. and Meeker, W. Q. (1991). *Statistical Intervals: A Guide for Practitioners*, Wiley, New York.

[9] Krishnamoorthy, K. and Mathew, T. (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation,* Wiley, Hoboken, NJ.

[10] Krishnamoorthy, K., Xia, Y., and Xie, F. (2011). A simple approximate procedure for constructing binomial and Poisson tolerance intervals, *Communications in Statistics-Theory and Methods*, **40**, 2443-2458.

[11] Naghizadeh Qomi, M., Kiapour, A., and Young, D. S. (2016). Approximate tolerance intervals for the discrete Poisson-Lindley distribution, *Journal of Statistical Computation and Simulation*, **86**(**4**), 841-854.

[12] Wang, W. and Tsung, F. (2009). Tolerance intervals with improved coverage probabilities for binomial and Poisson variables, *Technometrics*, **51**, 25-33.

[13] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209–212.

[14] Young, D. S. (2010). Tolerance: an R package for estimating tolerance intervals, *Journal of Statistical Software,* **36**, 1-39.

[15] Young, D. S. (2014). A procedure for approximate negative binomial tolerance intervals, *Journal of Statistical Computation and Simulation,* **84**(**2**), 438-450.

[16] Young, D. S. (2015). Tolerance Intervals for Hypergeometric and Negative Hypergeometric Variables, *Sankhya: The Indian Journal of Statistics, Series B*, **77**(**1**), 114-140.

[17] Zacks, S. (1970). Uniformly most accurate upper tolerance limits for monotone likelihood ratio families of discrete distributions, *Journal of the American Statistician Association*, **65**, 307-316.