حجم نمونه بیزی برای برآورد پارامتر نسبت دو جمله ای با استفاده از ناحیه‌های $p-$ تحمل با کمترین زیان پسین

حسین بیورانی[۱] و نرگس نجفی[۲]

چکیده:

این مقاله با استفاده از ناحیه‌های $p-$ تحمل با کمترین زیان پسین، تابع زیان درجه دو و استفاده از سه روش متوسط طول، متوسط پوشش و بدترین برآمد به محاسبه‌ی اندازه‌ی نمونه برای برآورد نسبت در تابع احتمال دو جمله‌ای با توزیع پیشین بتا می‌پردازد.

واژه‌های کلیدی: استنباط بیزی، تابع زیان درجه دو، توزیع بتا–دوجمله‌ای، ناحیه‌های با کمترین زیان پسین[۳] (LPL).

۱ مقدمه

برآورد حجم نمونه از سؤالاتی است که بسیاری از مواقع و به ویژه در طرح‌های پژوهشی و کاربردی با آن روبرو هستیم. اخیراً برای برآورد حجم نمونه‌ی بیزی روش‌های مختلفی پیشنهاد شده است. اشپیگل هالتر[۴] و فریدمن [۵][۹] و اشپیگل هالتر و همکاران [۱۰]، نگرش بیزی را برای پیش‌بینی توان آزمون فرض، آدکوک[۶][۱] و فام–گیا[۷] و ترکان [۸][۸] برای برآورد فاصله‌ای بر اساس تقریب‌های نرمال برای چگالی‌های پسین بر اساس میانگین‌ها و واریانس‌های پسین به کار برده‌اند.

یکی از اهداف برآورد حجم نمونه انجام استنباط یا اخذ تصمیم‌گیری در مورد پارامتر نامعلوم θ است. در نگرش کلاسیک که توسط دسو[۹] و راگاوارو[۱۰][۴] ارائه شده، مسئله‌ی اصلی یافتن برآورد نقطه‌ای $\hat{\theta}$ برای θ مجهول است. با توجه به این که در این روش از رفتار و پارامتر θ اطلاع نداریم و به عبارتی آن را ثابت فرض می‌کنیم، اندازه‌ی نمونه‌ی برآوردشده با خطای بیشتری مواجه خواهد بود در حالی که نگرش بیزی این امکان را فراهم می‌کند که در برآورد نقطه‌ای θ از توزیع پیشین آن استفاده کنیم. هدف از این مقاله برآورد حجم نمونه‌ی بیزی بر اساس ناحیه‌های با کمترین زیان پسین است. ناحیه‌های با کمترین زیان پسین فاصله‌هایی هستند که ریسک پسین نقاط داخل آن نواحی کمتر از ریسک

[۱] عضو هیئت علمی گروه آمار دانشگاه تبریز bevrani@tabrizu.ac.ir
[۲] عضو هیئت علمی دانشگاه آزاد اسلامی واحد ماکو narges_najafi63@yahoo.com
[۳] Lowest posterior loss
[۴] Spiegelhalter
[۵] Freedman
[۶] Adcock
[۷] Pham Gia
[۸] Turkkan
[۹] Desu
[۱۰] Raghavarao

[Downloaded from andisheyeamari.irstat.ir on 2024-04-25]

پسین نقاط خارج از آن است. برناردو[۱۱][۲] ناحیه‌های با کمترین زیان پسین را برای برآورد پارامتر نسبت در توزیع دوجمله‌ای با استفاده از تابع زیان حقیقی محاسبه نموده است. جوزف [۱۲] و همکاران [۶] این سه ملاک را بر اساس فاصله‌های با بیشترین چگالی پسین[۱۳] (HPD) برای تعیین اندازه‌ی نمونه‌ی بیزی برای پارامتر دوجمله‌ای به کار برده‌اند. کائو[۱۴] و همکاران [۳] تفاضل و نسبت بین اندازه‌های نمونه بر اساس سه ملاک متوسط پوشش، متوسط طول و بدترین برآمد را برای میانگین نرمال، تفاضل بین میانگین‌های نرمال و پارامتر دوجمله‌ای مورد مقایسه قرار می‌دهد. اولین مقاله بر اساس نظریه تصمیم برای تعیین اندازه‌ی نمونه توسط گروندی[۱۵] و همکاران [۵] بر پایه‌ی تابع مطلوبیت ارائه شد و توسط لیندلی[۱۶][۷] بسط داده شد. در این مقاله با به کار بردن نظریه‌ی تصمیم و استفاده از تابع زیان درجه دو به کمک سه ملاک متوسط طول، متوسط پوشش و بدترین برآمد، حجم نمونه‌ی بهینه را برای پارامتر θ در تابع احتمال دوجمله‌ای با پارامترهای n وθ و با توزیع پیشین بتا با پارامترهای α و β به دست می‌آوریم. بخش ۲ و زیربخش‌های آن به محاسبه اندازه نمونه بیزی برای برآورد نسبت دوجمله‌ای با استفاده از ناحیه‌های $p-$ تحمل با کمترین زیان پسین و به کمک سه ملاک مذکور اختصاص دارد. بخش ۳ نتایج بدست آمده را با مثال‌های عددی محاسبه و نشان می‌دهد. در نهایت در بخش ۴ خلاصه‌ای از نتایج بدست آمده مطرح و مورد بحث قرار گرفته‌اند.

۲ روش‌های اندازه‌ی نمونه‌ی بیزی برای نسبت دوجمله‌ای

فرض کنید متغیر تصادفی X دارای توزیع دوجمله‌ای با پارامترهای n و θ باشد، که در آن n اندازه‌ی نمونه است. همچنین فرض کنید θ دارای توزیع پیشین بتا با پارامترهای α و β، $\pi(\theta) = Be(\theta \mid \alpha, \beta)$، است. با استفاده از قضیه‌ی بیز، توزیع پسین θ، توزیع بتا با پارامترهای $x+\alpha$ و $n-x+\beta$، $\pi(\theta \mid x, n, \alpha, \beta) = Be(\theta \mid x+\alpha, n-x+\beta)$، و تابع چگالی پیشگو برای X، توزیع بتا–دوجمله‌ای خواهد بود که به صورت زیر تعریف می‌شود:

$$P_X(x \mid n, \alpha, \beta) = \binom{n}{x} \frac{B(x+\alpha, n-x+\beta)}{B(\alpha, \beta)}; \quad (۱)$$
$$x = ۰, ۱, ۲, \ldots, n,$$

که در آن $B(\alpha, \beta)$، تابع بتا و برابر با $\frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ است.

۱.۲ ناحیه‌های $p-$تحمل با کمترین زیان پسین

برای هر تابع زیان $L(\theta, \delta(x))$ برآوردهای ناحیه‌ای $p-$تحمل[۱۷] با کمترین زیان پسین ناحیه‌ای با احتمال p

[۱۱] Bernardo
[۱۲] Joseph
[۱۳] Highest posterior density
[۱۴] Cao
[۱۵] Grundy
[۱۶] lindley
[۱۷] p-tolerance

[Downloaded from andisheyeamari.irstat.ir on 2024-04-25]

است که شامل مقادیری از $\delta(x)$ می‌باشد که زیان مورد انتظار $E[L(\theta,\delta(x)) \mid x]$، برای آن‌ها کمتر از زیان مورد انتظار برای مقادیر $\delta(x)$ درخارج از این ناحیه است. فرض کنید $L(\theta,\delta(x)) = (\delta(x)-\theta)^2$ تابع زیان توأم دوم خطا باشد، ریسک پسین$\delta(x)$ به صورت زیر تعریف می‌شود:

$$R(\delta(x)) = \int_{\Theta} L(\theta,\delta(x))\pi(\theta \mid x)\, d\theta. \qquad (۲)$$

در این صورت ناحیه‌ی $p-$ تحمل با کمترین زیان پسین برای پارامتر دوجمله‌ای، زیرمجموعه‌ای به صورت $R_p^l = R_p^l(x,\Theta)$ از فضای پارامتر است هرگاه داشته باشیم:

$$i) \quad \int_{R_p^l} Be(\theta \mid x+\alpha, n-x+\beta)\, d\theta = p,$$
$$ii) \quad R(\theta) \le R(\theta'); \ \forall \theta \in R_p^l, \quad \forall \theta' \notin R_p^l,$$

که در آن

$$R(\delta(x)) = \int_{\circ}^{1} Be(\theta \mid x+\alpha, n-x+\beta)(\delta(x)-\theta)^2 \ d\theta,$$

است. حال فرض کنید$LPL_L(x,n,\alpha,\beta,l) = (h,g)$ که $h < g$، مطابق با ناحیه‌ی با کمترین زیان پسین برایθ با در نظر گرفتن طول l، برای فاصله و $LPL_C(x,n,\alpha,\beta,p)$ ناحیه‌ی با کمترین زیان پسین برایθ با در نظر گرفتن احتمال پوشش p باشد. تعریف می‌کنیم:

$$\ell'_p(x,n) = \int_{LPL_C(x,n,\alpha,\beta,p)} d\theta,$$

$$p'_\ell(x,n) = \int_{LPL_L(x,n,\alpha,\beta,\ell)} \pi(\theta \mid x,n,p,\alpha,\beta)\, d\theta,$$

که درآن $\ell'_p(x,n)$ طول دقیق ناحیه‌ی با کمترین زیان پسین با پوشش پسین p و $p'_\ell(x,n)$ پوشش پسین دقیق ناحیه‌ی با کمترین زیان پسین با طول معلوم ℓ برای مقادیر مشخص x و n است.

۲.۲ ملاک طول متوسط[۱۸] (ALC)

برای یک فاصله‌ی معتبر با کمترین زیان پسین با احتمال پوشش ثابت، مینیمم اندازه‌ی نمونه‌ی n را چنان تعیین می‌کنیم که طول مورد انتظار حداکثر برابر ℓ شود، یعنی ملاک طول متوسط، مینیمم n را به گونه‌ای جستجو می‌کند که داشته باشیم:

$$\sum_{x=\circ}^{n} \left[\int_{LPL_C(x,n,\alpha,\beta,p)} d\theta\right] \binom{n}{x} \cdot \frac{B(x+\alpha,n-x+\beta)}{B(\alpha,\beta)} \le \ell, \qquad (۳)$$

که درآن ℓ طول معین است. سمت چپ نامعادله‌ی (۳) متوسط طول فاصله با کمترین زیان پسین برای $X-$های مختلف است [۶].

۳.۲ ملاک پوشش متوسط[۱۹] (ACC)

همانندALC، ACC مینیمم اندازه‌ی نمونه‌ی n را چنان جستجو می‌کند که داشته باشیم:

$$\sum_{x=\circ}^{n} \left[\int_{LPL_L(x,n,\alpha,\beta,\ell)} Be(\theta \mid x+\alpha,n-x+\beta)\, d\theta\right] \cdot \binom{n}{x} \frac{B(x+\alpha,n-x+\beta)}{B(\alpha,\beta)} \ge p.$$

[۱۸] Average length criterion
[۱۹] Average coverage criterion

[Downloaded from andisheyeamari.irstat.ir on 2024-04-25]

به عبارت دیگر این ملاک با ثابت نگاه داشتن طول فاصله‌ی باکمترین زیان پسین، احتمال پوشش پسین را برای x–های مختلف به دست آورده و مینیمم n را چنان پیدا می‌کند که متوسط این پوشش پسین حداقل p شود، که درآن p مقداری معین است [٦].

۴.۲ ملاک بدترین برآمد[۲۰](WOC)

دو روش ALC و ACC، طول و پوشش را به طور متوسط بررسی می‌کنند. این روش‌ها هیچ‌گونه ضمانتی روی تک تک مشاهدات ندارند. روش محافظه‌کار دیگری که به ما این اطمینان را می‌دهد که طول و احتمال پوشش مورد نظر و مطلوب روی تک تک مشاهدات برقرار باشد، روش بدترین برآمد است. این روش کمترین مقدار را برای n به گونه‌ای می‌یابد که داشته باشیم:

$$\inf_{۰\le x\le n}\left[\int_{LPL_L(x,n,\alpha,\beta,\ell)}\pi(\theta\mid x,n,p,\alpha,\beta)\,d\theta\right]\ge p,$$

که درآن l و p مقادیر ثابتی هستند [٦].

۳ شبیه‌سازی

برای توزیع پیشین بتا با $\alpha=$ ۱، $\beta=$ ۱، $n=$ ۱۰ و $x=$ ۲، ریسک پسین در نمودار (۱)، ناحیه‌ی با کمترین زیان پسین در نمودار (۲) نشان داده شده است. همان‌گونه که در این نمودارها مشاهده می‌شود الگوریتم کلی برای بدست آوردن ناحیه‌های با کمترین زیان پسین بدین صورت است که مقدار ثابتی از ریسک پسین را به‌گونه‌ای می‌یابد تا ریسک پسین نقاط داخل این ناحیه مینیمم بوده و سطح زیر نمودار برای این ناحیه روی تابع چگالی پسین ۰/۹۵ شود. مقادیر حجم نمونه برآوردشده برای پارامترهای معلوم $\alpha=$ ۱ و $\beta=$ ۱ با استفاده از سه ملاک متوسط پوشش، متوسط طول و بدترین برآمد بر حسب مقادیر مختلف پوشش p و طول l به ترتیب در جداول (۱)، (۲) و (۳)آمده است. ردیف اول هر یک از این جداول مقادیر مختلف طول و ستون اول از سمت چپ مقادیر مختلف پوشش است. برای مثال با احتمال پوشش ۰/۹ و طول برابر با ۰/۳ حجم نمونه بدست آمده با روش ACC برابر ۱۹ و با ALC برابر با ۱۵ و با WOC برابر با ۳۰ است. نمودار (۳) تغییرات اندازه‌ی نمونه با سه روش ذکر شده برای l مساوی ۰/۳ و مقادیر مختلف p و نمودار (۴) تغییرات اندازه‌ی نمونه با این سه ملاک را برای p مساوی ۰/۹ و مقادیر l مختلف نشان می‌دهد. همان‌گونه که مشاهده می‌شود اندازه‌ی نمونه‌ی بدست آمده با ملاک WOC بزرگتر از دو ملاک دیگر است. در نمودار (۳) برای طول ثابت(l)، با افزایش پوشش (p)، حجم نمونه افزایش و در نمودار(۴) برای p ثابت با افزایش l، حجم نمونه کاهش می‌یابد.

۴ نتیجه‌گیری

در این مقاله چند روش محاسباتی برای تعیین اندازه‌ی نمونه‌ی بیزی برای برآورد پارامتر نسبت در توزیع دوجمله‌ای ارائه و بررسی شده است. همان‌گونه که قبلاً اشاره شد جوزف و همکاران [٦] این روش‌ها را برای

[۲۰] Worst outcome criterion

[Downloaded from andisheyeamari.irstat.ir on 2024-04-25]

فاصله‌های HPD برای پارامتر نسبت در توزیع دوجمله‌ای به کار برده‌اند. در این مقاله ناحیه‌های با کمترین زیان پسین را تعریف کرده، سپس به جای فاصله‌های HPD از ناحیه‌های با کمترین زیان پسین استفاده نموده‌ایم. با وجود این‌که فاصله‌های HPD کوتاهتر از فاصله‌های با کمترین زیان پسین هستند، ولی مزیتی که فاصله‌های با کمترین زیان پسین دارند این است که در محاسبه‌ی این فاصله‌ها از تابع زیان استفاده می‌شود. در جداول (۱)، (۲) و (۳) اندازه‌ی نمونه برای مقادیر مختلف l و p و در نظر گرفتن توزیع پیشین یکنواخت به ترتیب با استفاده از ملاک‌های ALC، ACC و WOC به دست آمده است. همان‌گونه که مشاهده می‌کنیم اندازه‌ی نمونه‌ای که از روش WOC به دست آمده بزرگتر از سایر روش‌ها است. برای احتمال پوشش ثابت، با کاهش طول، اندازه‌ی نمونه افزایش می‌یابد و برای طول ثابت با افزایش احتمال پوشش، اندازه‌ی نمونه افزایش پیدا می‌کند. در این مقاله برای تعیین اندازه نمونه بیزی برای برآورد نسبت دوجمله‌ای از تابع زیان درجه دو استفاده شد. در تحقیقات بعدی می‌توان توابع زیان دیگر را نیز به کاربرد و نتایج را مقایسه نمود.

۵ سپاسگزاری

نویسندگان از داوران محترم که پیشنهادات ارزنده ایشان موجب بهبود این مقاله گردید، تشکر و قدردانی می‌نمایند.

مراجع

[1] Adcock, C. J. (1988). A Bayesian approach to calculating sample sizes. *The Statistician*, 37: 433-439.

[2] Bernardo, J. M. (2005). Intrinsic credible regions: An objective Bayesian approach to interval estimation. *Test*, 14: 317- 384.

[3] Cao, J., Lee, J. J. and Alber, S. (2009). Comparison of Bayesian sample size criterion: ACC, ALC, WOC. *Journal of Statistical Planning and Inference*, 139:4111-4122.

[4] Desu, M. M. And Raghavarao, D. (1990). Sample Size Methodology. Boston: Academic Press.

[5] Grundy, P. M., Healy, M. J. R. and Rees, D. H. (1956) . Economic choice of the amount of experimentation. *Journal of the Royal Statistical Society: Series A* 18, 32-48 .

[Downloaded from andisheyeamari.irstat.ir on 2024-04-25]

[6] Joseph, L., and Wolfson, D. B., and Berger , R. D.(1995). Sample size calculations for Binomial proportions via highest posterior density intervals. *Statistician*, 44: 143-154.

[7] Lindley, D. V., (1997). The choice of sample size . Statistician, 46, 129-138 .

[8] Pham-Gia, T. and Turkkan, N. (1992). Sample size determination in Bayesian analysis (Disc: P399-404). *Statistician*, 41:389-397.

[9] Spiegelhalter, D. J. and Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, 5, 1-13.

[10] Spiegelhalter, D. J. and Freedman, L. S. and Parmar, M. K. B. (1994). Bayesian approaches to randomized trials (with discussion). *Journal of the Royal Statistical Society: Series A*, 157, 357-416.

[Downloaded from andisheyeamari.irstat.ir on 2024-04-25]

٦ پیوست ها

جدول ۱. اندازه‌ی نمونه‌ی بیزی به روش ACC برحسب طول و احتمالات پوشش مختلف با $\alpha = ۱$، $\beta = ۱$.

l / p	۰/۲	۰/۲۵	۰/۳	۰/۴	۰/۵
۰/۵	٦	۴	۳	۱	۱
۰/۸	۲٦	۱٦	۱۰	۵	۲
۰/۸۵	۳۴	۲۱	۱۴	۷	۴
۰/۹۰	۴٦	۲۹	۱۹	۱۰	۵
۰/۹۵	۷۰	۴۴	۲۹	۱۵	۹
۰/۹۹	۱۳۲	۸۳	۵۵	۳۰	۱۸

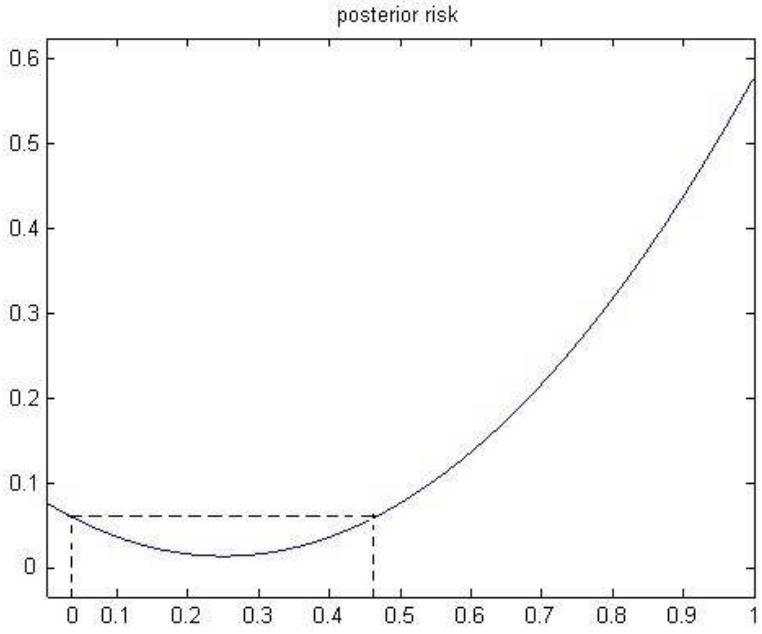
جدول ۲. اندازه‌ی نمونه‌ی بیزی به روش ALC برحسب طول و احتمالات پوشش مختلف با $\alpha = ۱$، $\beta = ۱$.

l / p	۰/۲	۰/۲۵	۰/۳	۰/۴	۰/۵
۰/۵	۷	۴	۳	۱	۱
۰/۸	۲۳	۱۴	۹	۵	۳
۰/۸۵	۲۸	۱۷	۱۲	٦	۳
۰/۹۰	۳۷	۲۳	۱۵	۸	۴
۰/۹۵	۵۲	۳۲	۲۲	۱۱	٦
۰/۹۹	۸۲	٦۰	۳۵	۱۸	۱۱

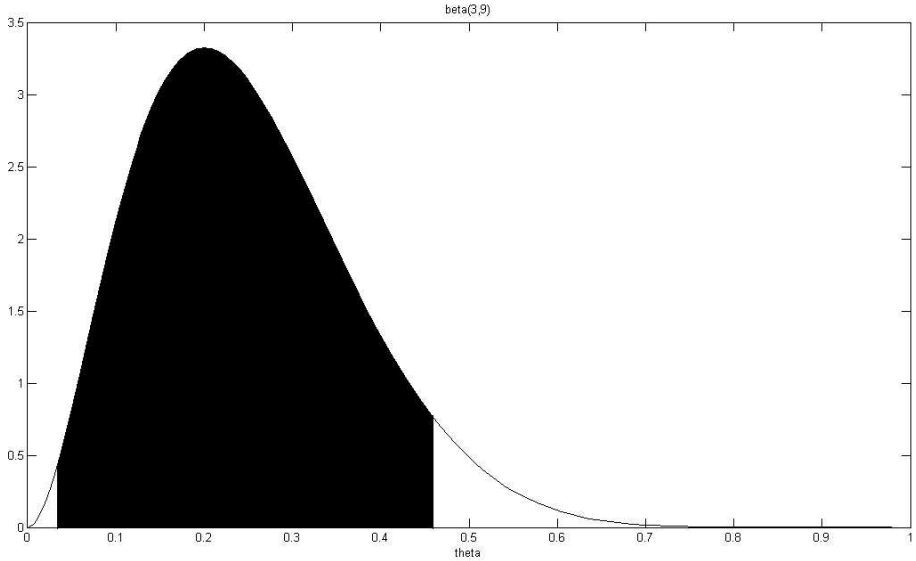
جدول ۳. اندازه‌ی نمونه‌ی بیزی به روش WOC برحسب طول و احتمالات پوشش مختلف با $\alpha = ۱$، $\beta = ۱$.

l / p	۰/۲	۰/۲۵	۰/۳	۰/۴	۰/۵
۰/۵	۱۱	۷	۳	۲	۱
۰/۸	۴۴	۲۷	۱۸	۹	۵
۰/۸۵	۵۵	۳۴	۲۳	۱۱	۷
۰/۹۰	۷۳	۴۵	۳۰	۱۵	۹
۰/۹۵	۱۰۳	٦۴	۴۳	۱۷	۱۳
۰/۹۹	۱۳۲	۱۱۹	۷۴	۳۹	۲۳

[Downloaded from andisheyeamari.irstat.ir on 2024-04-25]

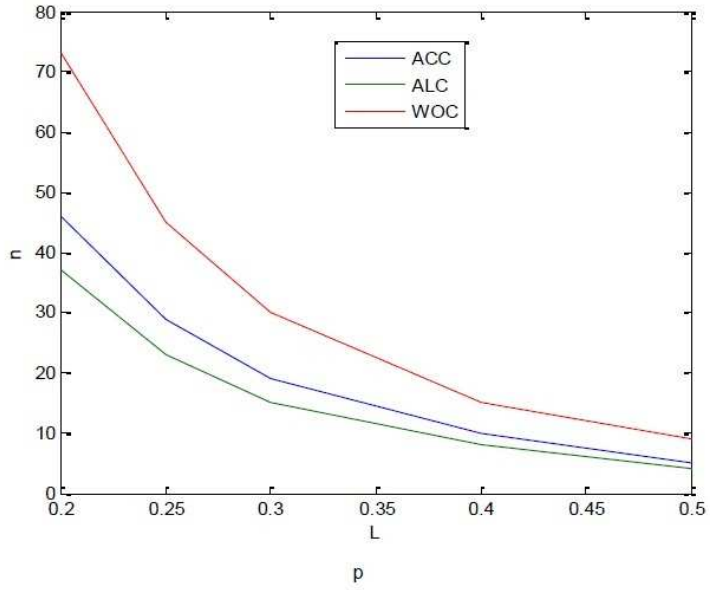


شکل ۱. ریسک پسین برای پارامتر دوجمله‌ای با $\alpha = ۱$، $\beta = ۱$.

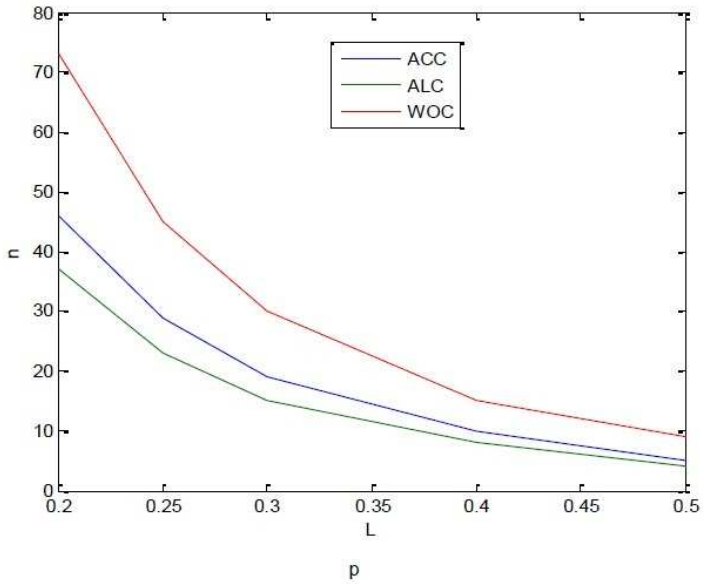


شکل ۲. فاصله‌ی تحمل ۹۵/. با کمترین زیان پسین برای پارامتر دوجمله‌ای با $\alpha = ۱$، $\beta = ۱$.

[Downloaded from andisheyeamari.irstat.ir on 2024-04-25]



شکل ۳. حجم نمونه با سه ملاک برای احتمالات پوشش مختلف و $l = ۰/۳$.



شکل ۴. حجم نمونه با سه ملاک برای طول‌های مختلف و $p = ۰/۹$.

[Downloaded from andisheyeamari.irstat.ir on 2024-04-25]