تعیین حجم نمونه بیزی برای میانگین نرمال

مهدی اغماضی، انوشیروان غفاری‌پور
گروه ریاضی، دانشگاه یاسوج

چکیده

تعیین حجم نمونه یک مسئله مهم در هنگام برآورد هر پارامتر است. گروهی از محققان تعیین حجم نمونه بیزی را مورد مطالعه قرار داده‌اند. دو گرایش اصلی در تعیین حجم نمونه بیزی وجود دارند. گرایش اول از ایده‌هایی از نظریه تصمیم برای ترکیب هزینه گرفتن یک نمونه با زیان مورد انتظار پسین و انتخاب حجم نمونه‌ای که هزینه مورد انتظار را مینیمم کند استفاده می‌کند. در چنین حالاتی بیشتر مسائل تعیین حجم نمونه بر مبنای توابع زیان مربع خطا و نمایی خطی می‌باشند. در این مسائل بیشتر از تابع زیان متقارن مربع خطا استفاده می‌شود. از سوی دیگر، در مواقعی که کم برآوردی یا بیش برآوردی مهم‌تر از یکدیگر باشند، آنگاه یک تابع زیان نامتقارن باید استفاده شود. در تحقیقات بالینی، معمولا پارامترهایی که برای تعیین حجم نمونه مورد نیاز هستند مجهول می‌باشند. یک شیوه معمول برای حل این مشکل استفاده از برآوردها به جای مقدار واقعی پارامتر در محاسبات است. ولی اگر خطای نمونه‌گیری بزرگ باشد، اندازه نمونه به دست آمده ممکن است گمراه‌کننده باشد. به عنوان یک راه حل، توصیه شده که از شیوه بیز با یک توزیع پیشین ناآگاهی‌بخش برای تصحیح خطای نمونه‌گیری استفاده شود. بر اساس یک توزیع پیشین ناآگاهی‌بخش و داده‌های نمونه‌گیری شده، برآوردگرهای بیز بر اساس یک تابع زیان مناسب به دست می‌آیند، سپس روش تعیین حجم نمونه سنتی با استفاده از برآوردگرهای بیز به جای برآوردگرهای معمول قابل انجام است. نتایج نشان می‌دهند که حجم نمونه به دست آمده با استفاده از شیوه بیز با حجم نمونه تعیین شده از روش سنتی کاملا متفاوت است. در این مقاله، حجم نمونه برای میانگین نرمال را به روش‌های مختلف بیزی به دست می‌آوریم. همچنین شرایطی را در نظر می‌گیریم که نمونه‌گیری به دلیل هزینه زیاد یا اطلاعات پیشین زیاد، به صرفه نیست و ارزش ندارد. گرایش دوم این است که به شاخص‌هایی توجه کنیم که مشابه توان هستند. این شاخص‌ها شامل معیارهای بیزی و آمیخته بیزی - درستنمایی از قبیل معیار متوسط پوشش، معیار متوسط طول و معیار بدترین پیشامد و استفاده از نواحی رفیع‌ترین چگالی پسین می‌شوند. در این مقاله، این معیارها برای پیدا کردن حجم‌های نمونه برای یک میانگین نرمال زمانی که واریانس‌ها معلوم یا مجهول هستند به کار گرفته می‌شوند. شیوه‌های کاملا بیزی و شیوه‌های آمیخته بیزی-درستنمایی مورد توجه قرار می‌گیرند.

واژه‌های کلیدی: تعیین حجم نمونه بیزی، تابع زیان نمایی خطی، تابع هزینه خطی، تابع مخاطره پسین، تابع زیان مربع خطا، بازه باورمند، توزیع نرمال.

۱ مقدمه

در سال‌های اخیر، مسئله تعیین حجم نمونه مورد توجه زیادی در علم آمار قرار گرفته است. دگروت[1][۴] حجم نمونه بیزی را با توجه به مربع خطا[2] و تابع زیان[3] خطای مطلق و یک تابع زیان خطی به دست آورد. لیندلی[4][۸] اندازه نمونه را با توجه به تابع زیان مربع خطا برای توزیع نرمال با واریانس معلوم به دست آورد. ادکاک[5][۲] راه حلی با فرم بسته برای تعیین حجم نمونه برای توزیع نرمال با واریانس‌های معلوم و مجهول با استفاده از معیار پوشش متوسط[6] (ACC) ارائه کرده است. ژوسف و همکاران[7][۷] حجم نمونه را برای نسبت دوجمله‌ای با توجه به بازه رفیع‌ترین چگالی پسین[8](HPD) به دست آوردند. انها یک مدل مرتبط با ACC به نام معیار متوسط طول[9] (ALC) تعریف کردند که در آن احتمال همگرایی $(1-\alpha)$ ثابت است و طول بازه HPD برابر l است. آنها همچنین یک روش بیزی دیگر برای تعیین حجم نمونه به نام معیار بدترین پیشامد مورد توجه قرار دادند که l,α هر دو ثابت فرض می‌شوند. ویس[10][۱۲] حجم نمونه را برای توزیع نرمال با واریانس‌های معلوم با استفاده از فاکتور بیز به دست آورد. ادکاک[۳] مسئله تعیین حجم نمونه را با استفاده از روش‌های رایج و روش‌های بیزی مرور کرد. ژوسف و بلیسلی[11][۵] حجم نمونه بیزی را برای یک میانگین نرمال و تفاوت بین میانگین‌های نرمال با استفاده از هر سه معیار ACC ، ALC و معیار بدترین پیشامد[12] (WOC) به دست آوردند. لیندلی[۹] بحث واضحی را برای تعیین حجم نمونه از طریق ماکسیمم‌سازی سود مورد انتظار[13] (MEU) ارائه و نتیجه را با روش ALC مقایسه کرد. سانتیس[14][۱۰] تعیین حجم نمونه را برای تجزیه و تحلیل توانمند بیزی مورد نقد و بررسی قرار داد. سانتیس[۱۱] از داده‌های تاریخی برای تعیین حجم نمونه بیزی استفاده کرد.

تعیین حجم نمونه یک مولفه مهم در طراحی مطالعات در خیلی از زمینه‌هاست. با توجه به اهمیت توزیع نرمال در علم آمار، این توزیع نقشی مرکزی در برآورد ملزمات حجم نمونه بازی می‌کند. فرمول استاندارد کلاسیک برای توزیع نرمال مانند:

$$n \geq \frac{4\sigma^2 Z^2_{1-\frac{\alpha}{2}}}{l^2}, \qquad (1)$$

تضمین می‌کند که یک فاصله اطمینان $100(1-\alpha)$ درصد دارای طول کل l خواهد بود، به شرط آن‌که بدانیم واریانس به صورت قیاسی دقیقا برابر با σ^2 است که $Z_{1-\frac{\alpha}{2}}$ صدک $1-\frac{\alpha}{2}$ توزیع نرمال استاندارد می‌باشد. سه محدودیت جدی برای این فرمول وجود دارد. اولی این است که قبل از آن که آزمایش انجام شود σ تقریبا هرگز با دقت خیلی بالا معلوم نیست، اما حجم نمونه توصیه شده با

[1] Degroot
[2] Error Squared
[3] Function Loss
[4] Lindley
[5] Adcock
[6] Criterion Coverage Average
[7] Josef
[8] Density Posterior Highest
[9] Criterion Length Average
[10] Weiss
[11] Belislie
[12] Criterion Outcome Worst
[13] Utility Expected the of Maximization
[14] Santis

نامساوی (۱) به صورت مستقیم با مربع σ متناسب است. دومی این است که بدون توجه به مقدار σ استفاده شده در نامساوی (۱)، استنباط‌های نهایی بر اساس داده مشاهده شده محاسبه می‌شوند، که البته در مرحله طراحی آزمایش معلوم نخواهند بود. سومی این است که اطلاعات پیشین درباره میانگین ممکن است موجود باشد و نادیده گرفتن این اطلاعات می‌تواند منجر به حجم نمونه بزرگ‌تر از حد مورد نیاز شود[۵].

در این مقاله، شیوه‌های بیزی برای تعیین حجم نمونه بحث می‌شوند که به هر سه این مشکلات پاسخ می‌دهند. به جای تنها یک مقدار، یک توزیع پیشین به σ نسبت داده می‌شود که شک و تردید درباره مقدار آن قبل از انجام آزمایش را منعکس می‌کند. این توزیع پیشین سپس برای تولید یک توزیع ماقبل پسین[۱۵] برای داده‌ها استفاده می‌شود، که یک وزن روی هر مجموعه داده که ممکن است اتفاق افتد قرار می‌دهد. اطلاعات پیشین درباره σ با اطلاعات موجود در داده‌ها به روزرسانی می‌شود تا استنباط‌های پسین از طریق قضیه بیز اجرا شوند که اغلب منجر به حجم‌های نمونه متفاوت با حجم‌های به دست آمده از روش‌های غیربیزی می‌شوند.

هر بردار داده بالقوه با طول n منجر به یک طول بازه متفاوت برای پوشش ثابت یا برعکس منجر به یک پوشش متفاوت برای یک طول بازه ثابت می‌شود. متوسط‌گیری از این طول‌ها یا پوشش‌ها روی توزیع ماقبل پسین برای داده‌ها منجر به دو معیار متفاوت برای تعیین حجم نمونه بیزی می‌شود. یک معیار سوم محافظه‌کارانه با توجه به مجموعه Φ که مثلا شامل ۹۵ یا ۹۹ درصد از محتمل‌ترین مقادیر داده‌های متناظر با توزیع ماقبل پسین می‌شود و به طور همزمان یک طول بازه کوچک و یک پوشش بزرگ باکفایت برای همه بردارهای داده در Φ را حتمی می‌کند به دست می‌آید.

بسیاری از محققان قبلا شیوه‌های بیزی برای تعیین حجم نمونه را مورد توجه قرار داده‌اند. برای یک میانگین نرمال، ادکاک[۲] در سال ۱۹۸۸ یک فرمول بسته برای حالات واریانس معلوم و مجهول با میانگین‌گیری مجموعه‌های باورمند پسین با طول برابر روی توزیع ماقبل پسین داده‌ها را ارائه کرد. فرمول ادکاک را مرور می‌کنیم و از تکنیک‌های مشابه برای رسیدن به فرم بسته فرمول برای حالت طول‌های متوسط مجموعه‌های باورمند پسین با پوشش ثابت، در کنار معیار ”بدترین حالت“ استفاده می‌کنیم[۵].

در این مقاله، حجم نمونه را برای برآورد میانگین نرمال با استفاده از توابع زیان مربع خطا و نمایی خطی و همچنین معیارهای بیزی و آمیخته بیزی-درستنمایی به دست می‌آوریم.

۲ تحلیل ماقبل پسین

فرض کنید $\underline{x} = (x_1, ..., x_n)$ یک نمونه تصادفی از توزیعی با تابع چگالی احتمال $f(x_i|n,\theta)$ باشد و توزیع پیشین برای θ به صورت $\pi(\theta)$ باشد. اگر درستنمایی نمونه به صورت $f(x|n,\theta)$ باشد آنگاه توزیع ماقبل پسین برای $x|n$ به صورت زیر خواهد بود[۱]:

$$f(x) = \int f(x|\theta, n)\pi(\theta)d\theta \qquad (۲)$$

[۱۵] pre-posterior

۳ توابع زیان

اگر d برآوردی برای θ باشد، آنگاه تابع زیان مربع خطا به صورت زیر تعریف می‌شود[۱]:

$$L_1(d,\theta) = a(d-\theta)^2 \tag{۳}$$

که a مقیاس است. برآورد بیز[۱۶]، $\widehat{d}_B$ تحت این تابع زیان برابر با میانگین پسین است. برای سادگی، معمولا a را برابر ۱ فرض می‌کنیم. فرض کنید $\pi(\theta|x)$ یک چگالی پسین باشد، آنگاه مخاطره[۱۷] پسین تحت این تابع زیان به صورت زیر است[۱]:

$$PR_1 = \int (\widehat{d}_B - \theta)^2 \pi(\theta|x) d\theta \tag{۴}$$

که به وضوح برابر با واریانس پسین است، زیرا برآوردگر بیز یک برآوردگر نااریب است. زلنر[۱۸][۱۳] یک تابع زیان نامتقارن را به صورت زیر در نظر گرفت:

$$L_2(d,\theta) = a[exp(b(d-\theta)) - b(d-\theta) - 1], a > 0, b \neq 0 \tag{۵}$$

که a مقیاس است و b شکل تابع زیان را تعیین می‌کند. این تابع زیان، تابع زیان نمایی خطی[۱۹] (Linex) نامیده می‌شود. برآوردگر بیز تحت این تابع زیان به صورت زیر خواهد بود[۱]:

$$\widehat{d}_B = \frac{-1}{b} lnE(exp(-b\theta)|x) \tag{۶}$$

[۱۶]Estimate Bayes
[۱۷]Risk
[۱۸]Zellner
[۱۹]Exponential Linear

برای سادگی، فرض می‌کنیم که پارامتر مقیاس یعنی a برابر ۱ باشد و تابع مخاطره تحت این تابع زیان عبارت است از:

$$PR_2 = \int [exp(b(\widehat{d}_B - \theta)) - b(\widehat{d}_B - \theta) - 1]\pi(\theta|x) d\theta$$

$$= exp(b\widehat{d}_B)E[exp(-b\theta|x)] - b\widehat{d}_B + bE(\theta|x) - 1$$

$$= exp(b\widehat{d}_B)exp(-b\widehat{d}_B) - b\widehat{d}_B + bm - 1$$

$$= b(m - \widehat{d}_B) \tag{۷}$$

که m میانگین پسین است[۱].

۴ تابع هزینه

هزینه نمونه‌گیری به ازای هر واحد از عوامل خیلی مهمی است که باید برای به دست آوردن یک حجم نمونه مناسب مورد توجه قرار گیرد. در این مقاله، یک تابع هزینه[۲۰] خطی که توسط لیندلی[۸] ارائه شده است را مورد توجه قرار می‌دهیم[۱]:

$$C(n) = c_0 + cn, n > 0 \tag{۸}$$

و $C(0) = 0$.
در اینجا، c_0 هزینه آماده‌سازی نمونه‌گیری و یا هر هزینه مرتبط دیگر که در نمونه‌گیری مشمول است بوده و c هزینه نمونه‌گیری به ازای هر واحد می‌باشد. اگر هیچ برنامه‌ای برای نمونه‌گیری وجود نداشته باشد، به طور آشکار هیچ هزینه آماده‌سازی و یا هزینه مرتبط دیگری وجود ندارد.

[۲۰]Function Cost

سرانجام، برای تعیین حجم نمونه مناسب هزینه کل را مینیمم می‌کنیم تا به یک حجم مناسب نمونه n برسیم. فرض کنید a مقداری در واحدهای یکسان با $C(n)$ باشد که تمایل داریم برای کاهش مخاطره پسین به‌اندازه ۱ بپردازیم. بنابراین هزینه کل می‌شود[۱]:

$$TC(n) = c_0 + cn + a(PR)$$

می‌توان این تابع هزینه را به صورت زیر نوشت:

$$TC'(n) = c'_0 + c'n + PR$$

به طوری که:

$$TC'(n) = \frac{TC(n)}{a}, c'_0 = \frac{c_0}{a}, c' = \frac{c}{a}$$

می‌توان گفت، مینیمم کردن $C(n)$ معادل است با مینیمم کردن $C'(n)$ [۱]. برای راحتی، همان طور که در بالا اشاره شد، a را برابر ۱ می‌گیریم. اما توجه داشته باشید که مقدار a برای توابع زیان مختلف در حالت کلی یکسان نیست، پس حجم‌های نمونه مناسب برای زیان مربع خطا و زیان نمایی خطی به طور مستقیم قابل مقایسه نیستند. همچنین توجه داشته باشید که در پیدا کردن یک هزینه مینیمم معمولا n را به عنوان یک مقدار حقیقی در نظر می‌گیریم و با توجه به آن از تابع هزینه مشتق می‌گیریم. البته در عمل n یک عدد صحیح است و باید $\lfloor n \rfloor$ را بررسی کنیم، که $\lfloor n \rfloor$ جزء صحیح n می‌باشد[۱].

۵ معیار بیزی برای تعیین حجم نمونه

فرض کنید θ نشان دهنده پارامتر تحت مطالعه، Θ فضای پارامتری برای θ و $\pi(\theta)$ توزیع پیشین θ باشند. فرض می‌کنیم که آزمایش تحت توجه شامل داده‌های $\underline{x} = (x_1, x_2, ..., x_n)$ باشد که n حجم نمونه است و مولفه‌های $\underline{x}$ قابل جایگزینی هستند و به فضای داده‌های χ تعلق دارند.
توزیع پسین θ به شرط $\underline{x}$ می‌شود:

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{\int_\Theta f(\underline{x}|\theta)\pi(\theta)d\theta} \qquad (۹)$$

که $f(x|\theta)$ درستنمایی داده‌های x است. اگر $\pi(\theta)$ یک چگالی یکنواخت روی Θ باشد، آنگاه (۹) درستنمایی نرمال شده است. در معیار بعدی، می‌توان فرض کرد که $\pi(\theta)$ نمایانگر توزیع واقعی پیشین در (۲) و (۹) باشد که می‌شود یک شیوه کاملا بیزی[۲۱] (FB) و یا می‌توان فرض کرد که $\pi(\theta)$ اطلاعات پیشین درست در (۲) باشد اما آن را با یک چگالی یکنواخت در (۹) عوض کنیم که ما این روش را شیوه آمیخته بیزی - درستنمایی[۲۲] (MBL) می‌نامیم.
معمولا یک بازه رفیع‌ترین چگالی پسین یا بازه باورمند پسین دیگری به طول l که θ را با احتمال $1-\alpha$ می‌پوشاند مورد نظر ماست. اما توزیع پسین θ به داده‌های $\underline{x}$ وابسته است، که البته در مرحله طراحی آزمایش مجهول است. می‌توانیم این تردید را با چندین شیوه متفاوت از بین ببریم که منجر به معیارهای بیان شده در بخش‌های بعدی

[۲۱] Bayesian Fully
[۲۲] Likelihood Bayesian Mixed

می‌شود[۶].

۱.۵ معیار متوسط پوشش (ACC)

اگر طول فاصله HPD را ثابت در نظر بگیریم، و اجازه دهیم که پوشش احتمال $\alpha - 1$ بسته به داده‌ها تغییر کند، آنگاه حجم نمونه برابر است با مینیمم n ای که در رابطه زیر صدق کند

$$\int_{\chi}\int_{a(x,n)}^{a(x,n)+l}\pi(\theta|x)d\theta f(x)dx \geq 1-\alpha \qquad (10)$$

که $f(x)$ توسط (۲) و $\pi(\theta|x)$ توسط (۹) داده شده‌اند و $a(x,n)$ حد پایین بازه HPD به طول l برای چگالی پسین $\pi(\theta|x)$ که در کل به هر دوی x,n وابسته است می‌باشند. می‌توان به سمت چپ رابطه (۱۰) به عنوان متوسط احتمالات پوشش پسین به طول ثابت l که با توزیع ماقبل پسین $f(x)$ وزن داده شده‌اند نگاه کرد[۶].

۲.۵ معیار متوسط طول(ALC)

اگر پوشش احتمال را ثابت در نظر بگیریم و به طول بازه HPD اجازه دهیم که بسته به داده‌ها تغییر کند، برای هر x در χ ابتدا باید طول بازه HPD یعنی $l'(x,n)$ را طوری پیدا کنیم که:

$$\int_{a(x,n)}^{a(x,n)+l'(x,n)}\pi(\theta|x)d\theta = 1-\alpha \qquad (11)$$

و حجم نمونه مینیمم عدد صحیح n ای است که در رابطه زیر صدق کند.

$$\int_{\chi} l'(x,n)f(x)dx \leq l \qquad (12)$$

که l متوسط طول از پیش تعیین شده است. قسمت سمت چپ رابطه (۱۲) از طول‌های بازه‌های HPD با پوشش ثابت که توسط توزیع ماقبل پسین $f(x)$ وزن داده شده‌اند میانگین می‌گیرد. ACC و ALC اغلب منجر به حجم‌های نمونه اساسا متفاوت می‌شوند[۶].

۳.۵ معیار بدترین پیشامد (WOC)

معیارهای ACC و ALC بر اساس متوسط روی همه نمونه‌ها هستند، اما از آنجا که استنباط‌ها مشروط به نمونه مشاهده شده هستند، منجر به پوشش‌ها و طول‌هایی بزرگ‌تر از حد مطلوب برای بعضی از نمونه‌ها می‌شوند. یک شیوه محافظه‌کارانه این است که یک طول ماکسیمم l و یک پوشش احتمال مینیمم $\alpha - 1$ را بدون توجه به داده‌های x در نظر بگیریم. بنابراین، $\alpha - 1$ و l هر دو از قبل ثابت هستند و مینیمم n را طوری انتخاب می‌کنیم که:

$$inf_{x\in\chi}\int_{a(x,n)}^{a(x,n)+l}\pi(\theta|x)d\theta \geq 1-\alpha \qquad (13)$$

یک اصلاح کوچک در WOC ، که منجر به $MWOC$[۲۳] می‌شود این است که در (۱۳) اینفیمم را روی یک زیرمجموعه φ از χ بگیریم. برای مثال، ممکن است مجموعه φ را به عنوان ناحیه ۹۹ درصد HPD متناظر با توزیع ماقبل پسین (۲) انتخاب کنیم. در اینجا یک طول ماکسیمم l با یک پوشش مینیمم $\alpha - 1$ برای ۹۹ درصد همه مجموعه داده‌های x که تمایل زیادی به اتفاق افتادن بر اساس اطلاعات پیشین دارند را تضمین می‌کنیم. بنابراین می‌توانیم از شرایط مجبور بودن به انتخاب یک حجم نمونه بزرگ غیرضروری برای جلوگیری از داده‌های

[۲۳] Criterion Outcome Worst Modified

شدیدا غیرمحتمل اجتناب کنیم[۶].

۶ تعیین حجم نمونه با استفاده از توابع زیان و هزینه

در این بخش تعیین حجم نمونه را برای توزیع نرمال برای برآورد میانگین زمانی که دقت معلوم است مورد توجه قرار می‌دهیم. فرض کنید $X_1, X_2, ..., X_n$ یک نمونه تصادفی به حجم n از یک توزیع نرمال با میانگین θ و واریانس معلوم σ^2 باشد که یعنی:

$$X_i \sim N(\theta, \sigma^2), i = 1, 2, ..., n$$

به دقت یعنی $\xi = \frac{1}{\sigma^2}$ توجه داشته باشید. حال یک توزیع نرمال با میانگین θ و دقت ξ را در نظر بگیرید.
با در نظر گرفتن یک توزیع پیشین مزدوج[۲۴] یعنی $N_0(\mu_0, n_0\xi)$ ، توزیع پسین یک توزیع نرمال با میانگین $\frac{n\overline{x}+n_0\mu_0}{n+n_0}$ و دقت $(n+n_0)\xi$ می‌باشد. فرض کنید میانگین پسین را با m و واریانس پسین را با ν^2 نشان دهیم، یعنی: $\pi(\theta|x) \sim N(m, \nu^2)$.
حال با توجه به تابع زیان مربع خطا، مخاطره پسین برابر با واریانس خواهد بود، ν^2 . یعنی:

$$PR_1 = \int (\widehat{d}_B - \theta)^2 \pi(\theta|x) d\theta$$

$$= \int (m - \theta)^2 \pi(\theta|x) d\theta$$

$$= \frac{1}{(n + n_0)\xi} \qquad (14)$$

اکنون با توجه به تابع هزینه (۸) و اضافه کردن این هزینه به تابع مخاطره بالا، هزینه کل به صورت زیر خواهد بود:

$$TC_1(n) = c_0 + cn + PR_1$$

لذا

$$TC_1(n) = c_0 + cn + \frac{1}{(n + n_0)\xi} \qquad (15)$$

که به وضوح مستقل از بردار داده‌های $\underline{X}$ می‌باشد. برای داشتن یک حداقل حجم نمونه n ، از رابطه بالا نسبت به n مشتق گرفته و آن را برابر صفر قرار می‌دهیم، یعنی $\frac{\partial TC_1(n)}{\partial n} = 0$ نتیجه می‌دهد:

$$n_{se} = \frac{1}{\sqrt{c\xi}} - n_0 \qquad (16)$$

اگر که اطلاعات پیشین کمی در اختیار داشته باشیم، یعنی $n_0 \approx 0$ ، آنگاه حجم نمونه تحت تابع زیان مربع خطا به صورت زیر خواهد بود:

$$n_{se} = \frac{1}{\sqrt{c\xi}} = \frac{\sigma}{\sqrt{c}}$$

واضح است که حجم نمونه به صورت مستقیم با انحراف استاندارد داده‌ها و به صورت معکوس با هزینه نمونه‌گیری به ازای هر واحد متناسب است. یعنی اگر تغییرپذیری داده‌ها افزایش پیدا کند، حجم نمونه هم افزایش می‌یابد و اما اگر هزینه نمونه‌گیری افزایش یابد، حجم نمونه مناسب کاهش پیدا می‌کند.
تحت تابع زیان نمایی خطی یعنی رابطه (۵)، از (۶)

[۲۴] Conjugate

برآورد بیز عبارت است از:

$$\frac{-۱}{b}\ ln\ E(exp(-b\theta)|x)$$

حال در اینجا داریم:

$$E(exp(-b\theta)|x) = exp(-mb + \frac{\nu^۲ b^۲}{۲})$$

با استفاده از این نتیجه برآورد بیز تحت تابع زیان نمایی خطی برابر خواهد بود با:

$$\widehat{d}_B = \frac{-۱}{b}\ ln\ exp(-mb + \frac{\nu^۲ b^۲}{۲})$$

$$= \frac{-۱}{b}(-mb + \frac{\nu^۲ b^۲}{۲})$$

$$= m - \frac{\nu^۲ b}{۲}$$

که m میانگین پسین و $\nu^۲$ واریانس پسین است. حال از رابطه (۷) مخاطره پسین تحت تابع زیان نمایی خطی به صورت زیر است:

$$PR_۲ = b(m - \widehat{d}_B)$$

$$= b(m - m + \frac{\nu^۲ b}{۲})$$

$$= b(\frac{۱}{(n+n_۰)\xi}\frac{b}{۲})$$

پس داریم

$$PR_۲ = \frac{b^۲}{۲\xi(n+n_۰)} \qquad (۱۷)$$

حال تابع هزینه خطی (۸) را در نظر بگیرید و آن را به تابع مخاطره بالا اضافه کنید. بنابراین هزینه کل عبارت است از:

$$TC_۲(n) = c_۰ + cn + PR_۲ \qquad (۱۸)$$

پس داریم:

$$TC_۲(n) = c_۰ + cn + \frac{b^۲}{۲\xi(n+n_۰)} \qquad (۱۹)$$

از آنجا که $C(n)$ از بردار داده $\underline{X}$ مستقل است، بنابراین برای به دست آوردن حجم نمونه مناسب از عبارت بالا نسبت به n مشتق گرفته و آن را برابر با صفر قرار می‌دهیم. پس داریم:

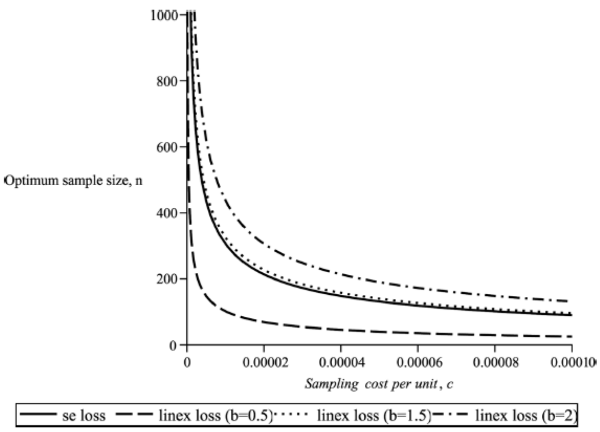
$$\frac{\partial TC_۲(n)}{\partial n} = ۰ \Rightarrow n_{lin} = \frac{b}{\sqrt{۲\xi c}} - n_۰ \qquad (۲۰)$$

اگر $b = \sqrt{۲}$ ، آن گاه با مقایسه عبارات (۱۶) و (۲۰)، می‌بینیم که حجم نمونه مناسب تحت توابع زیان نمایی خطی و مربع خطا با هم برابر هستند. به وضوح، n به دقت داده‌ها، پارامتر شکل تابع زیان و هزینه نمونه‌گیری به ازای هر واحد بستگی دارد. اگر اطلاعات پیشین با داشتن $n_۰ \approx ۰$ آگاهی‌بخش[۲۵] نباشد، در رابطه (۲۰) حجم نمونه تحت تابع زیان نمایی خطی عبارت است از:

$$n_{lin} = \frac{\sigma b}{\sqrt{۲c}}$$

یعنی، حجم نمونه تحت تابع زیان نمایی خطی به طور مستقیم با پارامتر شکل تابع زیان و تغییرپذیری داده‌ها در ارتباط است اما با هزینه نمونه‌گیری به ازای هر واحد زمانی که هیچ اطلاع پیشینی در دسترس نباشد، رابطه عکس دارد. در شکل (۱)، حجم نمونه مناسب تحت

[۲۵] Informative



شکل ۱: حجم نمونه با $\sigma = 1$ و $n_0 = 10$ تحت توابع زیان مربع خطا و نمایی خطی

توابع زیان نمایی خطی و مربع خطا برای مقادیر مختلف b به ازای مقادیر مشخص σ و n_0 و با تغییر c رسم شده است[۱].

حال شرایطی را در نظر می‌گیریم که نمونه‌گیری ارزش انجام ندارد. تابع هزینه را تحت تابع زیان مربع خطا در نظر بگیرید. حال هزینه کل بدون داشتن نمونه را با گرفتن یک مشاهده مقایسه کنید. با گرفتن $n = 0$ و $n = 1$ در $TC_1(n) = c_0 + cn + \frac{1}{(n+n_0)\xi}$ ، داریم:

$$TC_1(0) = \frac{1}{n_0\xi}, TC_1(1) = c_0 + c + \frac{1}{(1+n_0)\xi}$$

زمانی که $n = 0$ ، آنگاه c_0 حذف خواهد شد چرا که هیچ هزینه آماده سازی‌ای وجود ندارد. به وضوح:

$$c > \frac{\sigma^2}{n_0(n_0+1)} - c_0 \Rightarrow TC_1(0) < TC_1(1)$$

یعنی، با دانستن مقادیر σ^2 و n_0 ، اگر هزینه نمونه‌گیری از مقدار $\frac{\sigma^2}{n_0(n_0+1)} - c_0$ بیشتر شود، اصلا نمونه‌گیری ارزش انجام ندارد. به طور معادل، اگر اطلاعات پیشین کافی در اختیار داشته باشیم، یعنی:

$$n_0^2 + n_0 > \frac{\sigma^2}{c + c_0}$$

برای σ معلوم و هزینه نمونه‌گیری به ازای هر واحد c ، آنگاه نمونه‌گیری ارزش انجام ندارد. به وضوح، اگر حجم نمونه پیشین n_0 افزایش پیدا کند، آنگاه حجم نمونه مناسب یعنی n کاهش می‌یابد. در بعضی از مواقع، اگر n_0 به‌اندازه کافی بزرگ باشد، اصلا به نمونه احتیاج نداریم.
زمانی که تابع زیان نمایی خطی استفاده شود، ما حالت داشتن یک مشاهده را با هیچ نمونه یعنی $n = 1$ ، $n = 0$ را در عبارت $TC_2(n) = c_0 + cn + \frac{b^2}{2\xi(n+n_0)}$ مقایسه می‌کنیم. نمونه‌گیری ارزش انجام ندارد اگر:

$$c > \frac{\sigma^2 b^2}{2n_0(n_0+1)} - c_0$$

برای c_0، b، n_0 و σ^2 معلوم.
به طور معادل، اگر برای حجم نمونه پیشین داشته باشیم:

$$n_0^2 + n_0 > \frac{\sigma^2 b^2}{2(c + c_0)}$$

آن گاه نیز نمونه‌گیری اصلا ارزش انجام ندارد[۱].

۷ تعیین حجم نمونه با استفاده از معیارهای بیزی و آمیخته بیزی-درستنمایی

در این بخش یک توزیع پیشین مزدوج نرمال-گاما را برای (μ, ξ) فرض می‌کنیم، تا داشته باشیم:

$$\xi \sim gamma(\nu, \beta),$$

و

$$\mu|\xi \sim N(\mu_\circ, n_\circ\xi).$$

۱.۷ حجم‌های نمونه برای یک میانگین نرمال با دقت معلوم

اگر دقت با اطلاعات پیشین برابر با ξ باشد، آنگاه به سادگی می‌توان نشان داد که با اطلاعات پسین:

$$\mu|x \sim N(\mu_x, \xi_n),$$

که

$$\xi_n = (n + n_\circ)\xi,$$

$$\mu_n = \frac{n_\circ\mu_\circ + n\overline{x}}{n_\circ + n},$$

$$\overline{x} = \frac{۱}{n}\sum x_i.$$

از آنجا که دقت پسین فقط به n بستگی دارد و با بردار داده‌های مشاهده شده x تغییر نمی‌کند، هر سه معیار (ACC, ALC, WOC) منجر به یک راه حل یکسان می‌شوند، که همچنین معادل با فرمول ارائه شده توسط ادکاک در سال ۱۹۸۸ می‌باشد[۵]:

$$n \geq \frac{۴Z^۲_{۱-\frac{\alpha}{۲}}}{\xi l^۲} - n_\circ \qquad (۲۱)$$

۲.۷ حجم‌های نمونه برای یک میانگین نرمال با دقت مجهول

اگر دقت مجهول باشد، آنگاه توزیع پسین حاشیه‌ای μ به صورت زیر تعریف می‌شود:

$$\mu|x \sim \sqrt{\frac{\beta_n}{(n+n_\circ)(\nu + \frac{n}{۲})}}t_{۲\nu+n} + \mu_n,$$

که

$$\mu_n = \frac{n_\circ\mu_\circ + n\overline{x}}{n + n_\circ},$$

$$\beta_n = \beta + \frac{۱}{۲}ns^۲ + \frac{۱}{۲}\frac{nn_\circ}{n+n_\circ}(\overline{x} - \mu_\circ)^۲,$$

$$ns^۲ = \sum (x_i - \overline{x})^۲,$$

و $t_{۲\nu+n}$ یک توزیع t با $۲\nu + n$ درجه آزادی را نمایش می‌دهد. از آنجا که دقت پسین با داده‌های x تغییر می‌کند، معیار متفاوت منجر به حجم‌های نمونه متفاوت خواهد شد.

۱.۲.۷ معیار پوشش متوسط

ادکاک[۲] ثابت کرده است که حجم نمونه ACC عبارت است از:

$$n = \frac{4\beta}{\nu l^2} t^2_{2\nu_i 1-\frac{a}{2}} - n_0 \qquad (۲۲)$$

از آنجا که $\frac{\nu}{\beta}$ میانگین پیشین برای دقت ξ می‌باشد، حجم نمونه ACC برای دقت مجهول شبیه است به حالت دقت معلوم، از این نظر که فقط نیاز داریم که دقت میانگین پیشین برای ξ در نامساوی (۲۱) را تعویض کنیم و چندک نرمال Z را با چندک یک توزیع $t_{2\nu}$ عوض کنیم. از آنجا که درجات آزادی توزیع t با حجم نمونه افزایش پیدا نمی‌کنند، رابطه (۲۲) می‌تواند منجر به حجم‌های نمونه‌ای شود که به صورت اساسی متفاوت از حجم نمونه نامساوی (۲۱) باشد، حتی زمانی که n بزرگ باشد.

۲.۲.۷ معیار طول متوسط

زمانی که یک میانگین نرمال را با دقت مجهول ξ با یک توزیع $gamma(\nu,\beta)$ روی ξ برآورد می‌کنیم، رابطه ALC زمانی برقرار است که برای n به اندازه کافی بزرگ رابطه زیر برقرار باشد.

$$2t_{n+2\nu,1-\frac{\alpha}{2}} \sqrt{\frac{2\beta}{(n+2\nu)(n+n_0)}} \frac{\Gamma(\frac{n+2\nu}{2})\Gamma(\frac{2\nu-1}{2})}{\Gamma(\frac{n+2\nu-1}{2})\Gamma(\nu)} \leq l \qquad (۲۳)$$

اگر چه به نظر ممکن نمی‌رسد که فرم بسته نامساوی (۲۳) برای n به دست آید، قسمت سمت چپ با داشتن n، α، ν و n_0 از نظر محاسباتی سرراست است. بنابراین، مینیمم n دقیق که در رابطه (۲۳) صدق کند می‌تواند توسط یک الگوریتم جستجوی دوبخشی پیدا شود.

۳.۲.۷ معیار بدترین پیشامد

فرض کنید Φ زیرمجموعه‌ای از χ (فضای داده‌ها) باشد به طوری که

$$\int_{\Phi} f(x)dx = 1 - \omega$$

که برای هر $x \in \Phi$ و $y \notin \Phi$ داشته باشیم $f(x) \geq f(y)$. بنابراین Φ یک ناحیه رفیع‌ترین چگالی پسین بر اساس توزیع ماقبل پسین می‌باشد. پس می‌توان نشان داد که زمانی که یک میانگین نرمال را با دقت مجهول ξ با یک توزیع پیشین $gamma(\nu,\beta)$ روی ξ برآورد می‌کنیم، رابطه WOC زمانی برقرار است که برای n به‌اندازه کافی بزرگ، رابطه زیر برقرار باشد.

$$\frac{l^2(n+2\nu)(n+n_0)}{8\beta(1+(\frac{n}{2\nu})F_{n,2\nu;1-\omega})} \geq t^2_{n+2\nu;1-\frac{\alpha}{2}} \qquad (۲۴)$$

که $F_{n,2\nu;1-\omega}$ صدک $100(1-\omega)$ از یک توزیع F با درجات آزادی n و 2ν را نشان می‌دهد. مانند بخش قبلی، کوچک‌ترین n که در نامساوی (۲۴) صدق کند با یک جستجوی دوبخشی قابل پیدا کردن است. اگر $\chi = \Phi$ آنگاه حجم نمونه تعریف نشده است، چرا که با میل کردن ω به بینهایت داریم: $F_{n,2\nu;1-\omega} \to \infty$ بنابراین نامساوی (۲۴) به ازای هیچ n برقرار نیست[۵].

۳.۷ شیوه‌های آمیخته بیزی - درستنمایی برای یک میانگین نرمال

شیوه‌های آمیخته بیزی - درستنمایی از توزیع پیشین برای به دست آوردن توزیع ماقبل پسین داده‌ها استفاده می‌کنند، اما فرض می‌کنند که فقط از تابع درستنمایی برای استنباط‌های نهایی استفاده خواهد شد. این شیوه‌ها قصدشان راضی کردن محققانی می‌باشد که تشخیص می‌دهند اطلاعات پیشین برای طرح مهم هستند اما ترجیح می‌دهند که استنباط‌های نهایی را فقط بر اساس داده‌ها انجام دهند. برای مثال، آنها می‌توانند توسط محققانی که می‌خواهند فاصله اطمینان‌های ۹۵ درصد گزارش کنند، استفاده شوند. فرمول‌های بسته بیزی کامل در شیوه آمیخته بیزی - درستنمایی کاربرد ندارند. اما یک الگوریتم شبیه‌سازی می‌تواند برای تقریب حجم نمونه مورد نیاز به کار گرفته شود. مثلا برای ACC به دنبال مینیمم n هستیم به طوری که

$$\int_\chi \int_a^{a+l} \pi(\mu|x,n)d\mu f(x)dx \geq ۱-\alpha \qquad (۲۵)$$

برقرار باشد، که $f(x)$ توزیع ماقبل پسین داده‌های x به شرط یک توزیع پیشین $gamma(\nu,\beta)$ روی ξ و $\pi(\mu|x,n)$ توزیع پسین حاصل از توزیع پیشین ناآگاهی بخش $\pi(\mu,\xi) \propto \frac{1}{\xi}$ و $(a,a+l)$ یک فاصله باورمند رفیع‌ترین چگالی پسین برای μ می‌باشد. بنابراین،

$$(a,a+l) = (\overline{x}-\frac{l}{۲}, \overline{x}+\frac{l}{۲}),$$

$$\mu|x \sim \overline{x} + \sqrt{\frac{ns^۲}{n(n-۱)}}t_{n-۱},$$

$$\int \pi(\mu|x)d\mu = ۲p(\frac{l}{۲}\sqrt{\frac{n(n-۱)}{ns^۲}}; n-۱)$$

که $p(c,d)$ محل بین ۰ و c زیر توزیع t با d درجه آزادی است. از آنجا که این محل فقط از طریق $nS^۲$ به داده‌ها بستگی دارد، الگوریتم زیر حجم نمونه تقریبی را پیدا می‌کند.

(a) یک برآورد ابتدایی از حجم نمونه n انتخاب کنید.

(b) به‌اندازه m مقدار از متغیر تصادفی $nS^۲$ تولید کنید. از آنجا که

$$nS^۲ \sim gamma(\frac{n-۱}{۲}, \frac{\xi}{۲}),$$

$$\xi \sim gamma(\nu,\beta)$$

داریم:

$$nS^۲ \sim gamma-gamma(\nu, ۲\beta, \frac{n-۱}{۲})$$

یعنی $nS^۲$ از یک توزیع گاما-گاما پیروی می‌کند. یک متغیر تصادفی X از یک توزیع گاما-گاما پیروی می‌کند اگر

$$f(x|\nu,\beta,n) = \frac{\Gamma(\nu+n)\beta^\nu}{\Gamma(\nu)\Gamma(n)} \frac{x^{n-۱}}{(\beta+x)^{\nu+n}}$$

برای $n>۰$، $\beta>۰$، $\nu>۰$ و $x>۰$.

(c) برای هر یک از m مقدار $ns^۲$ و $i=۱,۲,...,m$ در مرحله (b) رابطه زیر را محاسبه کنید:

$$coverage(ns^{۲}) = ۲p(\frac{l}{۲}\sqrt{\frac{n(n-۱)}{ns_i^{۲}}}; n-۱)$$

(d) آنگاه

$$\frac{۱}{m}\sum coverage(ns_i^{۲})$$

پوشش متوسط را برآورد می‌کند که سمت چپ نامساوی (۲۵) را تشکیل می‌دهد.

با تکرار گام‌های (b)-(d) برای مقادیری از n که توسط یک فرآیند جستجوی دوبخشی انتخاب شده‌اند به یک حجم نمونه درست تقریبی می‌رسیم. دقت این برآورد با افزایش m افزایش می‌یابد.
به طور مشابه، حجم نمونه WOC می‌تواند با برقراری رابطه

$$۲p(\frac{l}{۲}\sqrt{\frac{n(n-۱)}{ns_i^{۲}}}) \geq ۱-\alpha$$

برای ۹۵ درصد یا ۹۹ درصد از مقادیر $ns_i^{۲}$ برای یک مقدار داده شده n تقریب زده شود.
برای ALC ساده‌سازی بیشتر نیاز به شبیه‌سازی داده‌ها را رفع می‌کند. این به این دلیل است که می‌توان نشان داد که طول متوسط ناحیه باورمند پسین بر پایه درستنمایی برای μ در سطح $۱-\alpha$ عبارت است از:

$$\frac{۲}{\sqrt{n(n-۱)}}E(\sqrt{ns^{۲}})t_{n-۱;۱-\frac{\alpha}{۲}}$$

به طوری که

$$E(\sqrt{ns^{۲}}) = \sqrt{۲\beta}\frac{\Gamma(\frac{n}{۲})}{\Gamma(\frac{n-۱}{۲})}\frac{\Gamma(\nu-\frac{l}{۲})}{\Gamma(\nu)}$$

بنابراین طول متوسط که توسط توزیع پیشین $gamma(\nu,\beta)$ روی دقت ξ برای یک حجم نمونه n به دست می‌آید عبارت است از:

$$۲t_{n-۱;۱-\frac{\alpha}{۲}}\sqrt{\frac{۲\beta}{n(n-۱)}}\frac{\Gamma(\frac{n}{۲})}{\Gamma(\frac{n-۱}{۲})}\frac{\Gamma(\nu-\frac{l}{۲})}{\Gamma(\nu)} \quad (۲۶)$$

پیدا کردن حداقل n به طوری که رابطه (۲۶) کمتر از طول مورد انتظار l باشد حجم نمونه دقیق ALC تحت یک شیوه آمیخته بیزی - درستنمایی را حاصل می‌کند[۵].

۸ بحث

در این مقاله، نشان دادیم که چگونه یک شیوه نظری تصمیم بیزی برای تعیین حجم نمونه می‌تواند به تابع زیان نمایی خطی تعمیم داده شود و چند حالت خاص را مورد توجه قرار دادیم. حالاتی را اشاره کردیم که بنابر هزینه بالای جمع‌آوری داده و یا اطلاعات پیشین قوی، نمونه‌گیری ارزش انجام ندارد.
در تعیین حجم نمونه برای یک تجزیه و تحلیل بیزی، بعضی از نویسندگان اخیرا از یک شیوه دارای دو توزیع پیشین طرفداری کرده‌اند. بر اساس این ایده، باید از یک توزیع پیشین برازشی استفاده کنیم، که نمایانگر دانش ما درباره مقدار محتمل پارامتر در حین طراحی آزمایش، یعنی انتخاب حجم نمونه در این حالت باشد؛ اما از یک توزیع پیشین دارای اطلاعات کمتر برای گزارش استنباط‌های ما استفاده کند. اگر به یک حجم نمونه مناسب بر اساس یک توزیع پیشین با یک حجم نمونه معادل $n_{۰}$ برسیم، سپس برای استنباط این اطلاعات

پیشین را نادیده بگیریم، ریسک پسین ما افزایش پیدا می‌کند و فعالیت ما منطقی نخواهد بود.

در این مقاله معیارهایی برای تعیین حجم نمونه بیزی که به محدودیت‌های محاسبات کلاسیک پاسخ می‌دهد معرفی کردیم. اما راه حل‌ها یکتا نیستند، نه فقط به خاطر اینکه انتخاب پارامترهای پیشین ممکن است واضح نباشد، بلکه همچنین به این دلیل که معیارهای متفاوت می‌توانند اساسا منجر به حجم‌های نمونه متفاوت شوند. می‌توان گفت انتخاب بین یک معیار که روی توزیع ماقبل پسین داده‌های مجهول میانگین می‌گیرد و یک معیار بدترین حالت در هر آزمایش داده شده، به میزان تمایل ما به خطر کردن بستگی دارد. انتخاب بین ACC و ALC تا حدی دلخواه است، گرچه به نظر می‌رسد که قرارداد گزارش ۹۵ درصد بازه‌ها بدون در نظر گرفتن داده‌ها از ALC پشتیبانی می‌کند. توجه به همه معیارها باید منجر به انتخابی آگاهانه‌تر شود.

در همه حالات، محاسبات می‌توانند به صورت سرراست و سریع، هم به صورت دقیق و هم از طریق شبیه‌سازی انجام شوند. تکنیک‌های این مقاله می‌توانند به توزیع‌های پیشین نامزدوج از طریق الگوریتم‌های شبیه‌سازی مناسب تعمیم داده شوند. پیشرفت اخیر در الگوریتم‌های محاسباتی بیزی در این زمینه می‌تواند کمک کننده باشد.

مراجع

[1] Saiful Islam, A. F. M. and Pettit, L. I. (2012). Bayesian sample size determination using linex loss and linear cost. *Communication in Statistics - Theory and Methods*, 41, 223-240.

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

[3] Adcock, C. J. (1997). Sample size determination: A review. *Statistician*., 46, 262-283.

[4] Degroot, M. H. (1970). *Optimal Statistical Decision*. New York: McGraw-Hill Inc.

[5] Joseph, L. and Belisle, P. (1997). Bayesian sample size determination for normal means and differences between normal means. *Statistician*, 46, 209-226.

[6] Joseph, L., du Berger, R. and Belisle, P. (1997). Bayesian and mixed Bayesian likelihood criteria for sample size determination. *Statist. Med.*, 16, 769-781.

[7] Joseph, L., Wolfson, D. B. and du Berger, R. (1995). Sample size calculations for binomial proportions via height posterior density intervals. *Statistician*, 44, 143-154.

[8] Lindley, D. V. (1972). Bayesian Statistics, A Review. Philadelphia, Society for Industrial and Applied Mathematics.

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

[10] Santis, F. D. (2006). Sample size determination for robust Bayesian analysis. *J. Amer. Statist. Assoc.*, 101, 278-291.

[11] Santis, F. D. (2007). Using historical data for Bayesian sample size determination. *J. Roy. Statist. Soc. Ser. A*, 170, 95-113.

[12] Weiss, R. (1997). Bayesian sample size calculation for hypothesis testing. *Statistician*, 46, 185-191.

[13] Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *J. Amer. Statist. Assoc.*, 81, 446-451.