مجله علوم آماری، پاییز و زمستان ۱۳۸۸
جلد ۳، شماره ۲، ص ۱۱۹-۱۳۹

شناسایی نقاط دورافتاده در داده‌های نرمال بر اساس مقادیر Z اصلاح شدهٔ مشاهدات

محمدولی احمدی، مجید سرمد
گروه آمار، دانشگاه فردوسی مشهد

تاریخ دریافت: ۱۳۸۸/۵/۷ تاریخ آخرین بازنگری: ۱۳۸۸/۱۲/۱۶

چکیده : در این مقاله، به دلیل اهمیت و گستردگی استفاده از توزیع نرمال، نمونه‌های مبتنی بر این توزیع در نظر گرفته شده، با استفاده از مقادیر برش وابسته به حجم نمونه، نقاط دورافتادهٔ آن‌ها شناسایی می‌شوند. برای به‌دست آوردن مقادیر برش بهینه یک مسأله تصمیم مطرح و به روشی کم‌بیشینه (مینیماکس) حل می‌گردد. در حل این مسأله از روش شبیه‌سازی بهره گرفته شده است .

واژه‌های کلیدی : د رجه بدی، توزیع اسلش، مقادیر Z، مقادیر Z اصلاح شده.

۱ مقدمه

اگر چه زمانی که صحبت از نقطه دورافتاده به میان می‌آید، مفهوم واحدی از آن برداشت می‌شود، اما در بیان تعریف نمی‌توان تعریف جامعی از آن ارائه کرد. باید گفت میزان اعتماد به هر مشاهده، به رابطه آن با سایر مشاهداتی بستگی دارد که تحت شرایط یکسان به دست آمده‌اند. بر این اساس به عقیده اکثر محققین مشاهده‌ای که دور از تودهٔ مشاهدات قرار می‌گیرد، دادهٔ دورافتاده محسوب می‌شود.

° آدرس الکترونیک مسؤول مقاله: محمد ولی احمدی، *ahmadi_m۸۵@yahoo.com*
° کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲G۱۰ و ۶۲G۳۰

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

بنا بر بارنت و لویس (۱۹۸۴) «دادهٔ دورافتاده مشاهده‌ای است که با سایر مشاهدات نمونه ناسازگار و متفاوت است». علاوه بر این در تعریفی مشابه هاوکینز (۱۹۸۰) نقطه دورافتاده را مشاهده‌ای معرفی می‌کند که «از الگوی کلی داده‌ها پیروی نمی‌کند، به گونه‌ای که به نظر می‌رسد بر اساس فرایندی متفاوت با سایر داده‌ها به‌دست آمده است.» با این حال بکمن و کوک (۱۹۸۳) عقیده دارند «مشاهده دورافتاده، یا مشاهده‌ای ناهماهنگ[۱] و متفاوت است یا مشاهده‌ای آلوده[۲]». مشاهده‌ی ناهماهنگ «مشاهده‌ای است که حضور آن در میان مشاهدات غیرطبیعی به نظر می‌رسد». اما مشاهدهٔ آلوده «مشاهده‌ای است که از توزیعی غیر از توزیع فرض شده برای جامعه به‌دست آمده است.» در این نوشتار تعریف بارنت و لویس (۱۹۸۴) را می‌پذیریم. وقوع این دسته از مشاهدات در اکثر نمونه‌های تصادفی تقریباً امری طبیعی‌است. بنابراین اولین دغدغه و نگرانی محققین پس از جمع‌آوری مشاهدات نمونه، چگونگی برخورد با نقاط دورافتاده است. به دلیل اینکه شناسایی علل و عوامل وقوع نقاط دورافتاده، کیفیت فرآیند نمونه‌گیری را ارتقا می‌دهد یا در برخی موارد توزیعی را که برای داده‌ها فرض کرده‌ایم بهبود می‌بخشد، بررسی این مقادیر از اهمیت بسیاری برخوردار است. اما در ابتدا مسأله اساسی، شناسایی این مقادیر در میان توده مشاهدات است و بررسی ظاهری مشاهدات، اولین راه برای تشخیص مقادیر دورافتاده می‌باشد. بر اساس تحقیقی که کالت و لویس (۱۹۷۶) بر داده‌های تک‌متغیره انجام دادند، دریافتند شناسایی نقاط دورافتاده در این داده‌ها به عوامل زیر بستگی دارد.

الف. نحوه بررسی مقادیر نمونه برای شناسایی مقادیر دورافتاده: اگر برای شناسایی نقاط دورافتاده، داده‌های تصادفی در سه حالت مرتب شده، مرتب نشده و نموداری بررسی شوند، شانس تشخیص صحیح این مقادیر در داده‌های مرتب شده بسیار بیشتر است.

ب. مقیاس مقادیر نمونه: اگر مقیاس اندازه‌گیری مشاهدات افزایش یابد، احتمال

---

۱ Discordant

۲ Contaminant

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

اینکه مشاهدات بزرگتر دورافتاده فرض شوند، افزایش خواهد یافت.

بنابراین بررسی ظاهری مشاهدات به تنهایی روشی مطمئن برای شناسایی نقاط دورافتاده نیست، علاوه بر این در نمونه‌های پیچیده‌تر با حجم‌های بزرگ‌تر این کار تقریباً غیرممکن است. پس در این موارد بهتر است به جای بررسی ظاهری داده‌ها، از معیارهای منطقی‌تری بهره گرفته شود. واضح است که با استفاده از این معیارها نمی‌توان مشاهدات دورافتاده را به طور قطعی پیدا کرد، بلکه تنها مقادیر مشکوکی را می‌توان یافت که ارزش بیشتری دارد تا دربارهٔ دورافتاده بودنشان تحقیق شود. چون اکثر پدیده‌های طبیعی از توزیع نرمال پیروی می‌کنند، در این مقاله راه حلی برای شناسایی نقاط دورافتاده در داده‌های نرمال ارائه می‌شود. مشهورترین روش برای شناسایی نقاط دورافتاده در داده‌های نرمال، استفاده از مقادیر Z مشاهدات است. بر اساس نمونه تصادفی $X_1, \ldots, X_n$ از توزیع نرمال این مقادیر برابر با $Z_i = \frac{X_i - \bar{X}}{S}$ هستند که در آن $\bar{X}$ و S به ترتیب میانگین و انحراف معیار نمونه می‌باشند. چون در توزیع نرمال استاندارد، ۹۹/۷ درصد مشاهدات بین ۳- و ۳ قرار می‌گیرند، مشاهده X_i که برای آن $|Z_i| > ۳$ است، نقطه دورافتاده فرض می‌شود. اما این روش به خصوص در نمونه‌های کوچک دقیق نیست. شفلر (۱۹۸۸) نشان داد که مقدار مطلق Z_i در نمونه‌ای به حجم n، حداکثر برابر $\frac{n-۱}{\sqrt{n}}$ است. بنابراین در نمونه‌ای به حجم ۱۰، همواره $|Z_i| < ۲/۸۵ < ۳$ است، در حالی که فرض نرمال بودن اعضای هر نمونه ۱۰ تایی ممکن غیرمنطقی است.

برای رفع این مشکل ایگلیویکس و هوگلین (۱۹۹۳) از برآوردگرهای استوار بهره گرفتند. به طور کلی با استفاده از برآوردگرهای استوار در صورتی که درصدی از مشاهدات نمونه نیز دورافتاده باشند می‌توان برآورد صحیحی را از پارامتر جامعه به‌دست آورد. آن‌ها برای برآورد پارامتر مرکزی جامعه از میانه مشاهدات ($\tilde{X}$) و برای برآورد پراکندگی در جامعه از برآوردگری که همپل (۱۹۷۴) به صورت $MAD = median\{|X_i - \tilde{X}|\}$ معرفی کرده است استفاده کردند. بر این اساس مقادیر Z اصلاح شدهٔ مشاهدات به صورت $M_i = \frac{X_i - \tilde{X}}{\frac{MAD}{d}}$ تعریف می‌شود که در آن d موجب می‌شود، $\frac{MAD}{d}$ برآوردگری نااریب برای σ باشد. به علاوه می‌توان نشان

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

داد بر اساس نمونه‌های با حجم زیاد از توزیع $N(\mu,\sigma^2)$، $E(MAD)=0/6745\sigma$ است، لذا ثابت d را برابر ۰/۶۷۴۵ در نظر گرفتند. به علاوه بر اساس بررسی‌های شبیه‌سازی بر روی داده‌های نرمال، تصمیم گرفتند مشاهداتی را که در شرط $|M_i|>3/5$ صدق می‌کنند، نقطه دورافتاده در نظر بگیرند.

ثابت d را زمانی می‌توان برابر ۰/۶۷۴۵ در نظر گرفت که حجم نمونه به بی‌نهایت میل کند. بنابراین در نمونه‌های با حجم کم مقدار واقعی این ثابت متفاوت می‌باشد. از این رو در مقاله حاضر برخلاف ایگلیویکس و هوگلین (۱۹۹۳) که d را برای نمونه‌های با حجم‌های متفاوت یکسان در نظر گرفتند، این ثابت برای نمونه‌های تصادفی به حجم‌های $n=5,6,\ldots,29,30$ به صورت مجزا شبیه‌سازی می‌شود. اگر از مقادیر ثابت d وابسته به حجم نمونه در محاسبهٔ مقادیر $|M_i|$ استفاده شود، برای شناسایی مقادیر دورافتاده مقادیر برشی مورد نیاز است که آن‌ها نیز به حجم نمونه بستگی داشته باشند. بنابراین در این مقاله، مقادیر برش وابسته به حجم نمونه به روشی کم‌بیشینه و مبتنی بر شبیه‌سازی به‌دست آورده می‌شود. بر اساس این روش می‌توان با اعمال ثابت d برای هر حجم نمونه، مقادیر Z اصلاح شده مشاهدات را به‌دست آورد و در نهایت از مقایسه مقادیر Z اصلاح شده با مقدار برش وابسته به حجم نمونه، مقادیر دورافتاده را در میان مشاهدات شناسایی نمود.

از این رو ابتدا چگونگی یافتن مقادیر برش در بخش ۲ توضیح داده می‌شود. در بخش ۳ نحوه شبیه‌سازی مسألهٔ مطرح شده در بخش ۲ بیان می‌گردد و در بخش ۴ نتایج شبیه‌سازی‌ها به تفصیل شرح داده می‌شوند. در بخش آخر کارآیی مقدار برش ثابت ۳/۵ و مقادیر برش به‌دست آمده در این مقاله مقایسه می‌شوند.

۲ چگونگی یافتن مقادیر برش بر اساس حجم نمونه تصادفی

نمونه‌ای تصادفی را تصور کنید که از جامعه‌ای نرمال به‌دست آمده است. اما در میان مشاهدات این نمونه امکان رخداد مقادیر دورافتاده به دلیل عواملی که تحت کنترل نیستند وجود دارد. بنابراین در پی یافتن مقدار برشی هستیم که از مقایسه آن با مقادیر Z اصلاح شدهٔ مشاهدات نمونه به دست آمده، بتوان مقادیر دورافتاده را

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

شناسایی کرد. برای این منظور ابتدا یک تابع زیان معرفی می‌شود تا با استفاده از آن میزان زیان ناشی از به کارگیری مقادیر برش مختلف با یکدیگر مقایسه شوند. سپس دو حالت برای نمونه تصادفی به‌دست آمده در نظر گرفته می‌شود. حالت اول زمانی که هیچ نقطه دورافتاده‌ای در میان مشاهدات نمونه وجود ندارد و حالت دوم وقتی که نمونه به‌دست آمده شامل مشاهدات دورافتاده است. مقدار برشی بهینه است که ماکسیمم زیان ناشی از به کارگیری آن در شناسایی نقاط دورافتاده در دو حالت فوق کمترین گردد. (قاعده مینیماکس[۳])

تعریف ۱ : درجه بدی: درجه بدی مقدار برش λ مبتنی بر نمونهٔ تصادفی $X_1, \ldots, X_n$ از توزیعی با میانگین μ، عبارت از $B(\lambda) = (\bar{X}_\lambda - \mu)^2$ است، که در آن $\bar{X}_\lambda$ میانگین مشاهداتی از نمونه است که بر اساس نقطه برش λ در نمونه باقی می‌مانند.

در این تعریف فرض شده است که هدف از جمع‌آوری داده‌ها برآورد میانگین جامعه می‌باشد و برای این منظور از میانگین نمونه پس از حذف مقادیر دورافتاده به عنوان برآوردگری منطقی استفاده می‌شود.

اگر فرض شود $b(\lambda) = E(B(\lambda))$، در این صورت λ_m مقداری است که به ازای آن $b(\lambda)$ مینیمم می‌شود، یعنی $b_m = b(\lambda_m) = \min_\lambda b(\lambda)$. بنابراین درجه بدی مقیاس‌بندی شده به صورت

$$b_{SC}(\lambda) = \frac{b(\lambda)}{b_m}$$

تعریف می‌شود. بنابر دلایل زیر از درجه بدی مقیاس‌بندی شده برای یافتن مقدار برش بهینه استفاده می‌شود:

الف. بهترین مقدار برش برای مشاهدات نمونه وقتی است که $b(\lambda)$ کمترین مقدار خود یعنی b_m را اختیار کند. از طرفی $b(\lambda)$ معیاری است که میزان ناکارآیی میانگین نمونه را، پس از حذف مقادیر دورافتاده بر اساس مقدار برش λ، برای

[۳] Minimax rule

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

برآورد میانگین جامعه نشان می‌دهد. بنابراین $b_{SC}(\lambda)$ ناکارآیی نسبی برآوردگر میانگین را در مقایسه با بهترین حالت ممکن آن نشان می‌دهد.

ب. برای این که بتوان در ازای λ های معین، درجات بدی توزیع‌های مختلف را با یکدیگر مقایسه نمود، استفاده از $b_{SC}(\lambda)$ بر $b(\lambda)$ ترجیح داده می‌شود.

پس از معرفی تابع زیان، برای این که بتوان حالات متصور برای نمونهٔ تصادفی را شبیه‌سازی کرد، از دو توزیع کاملاً متفاوت در تولید مقادیر دورافتاده بهره گرفته می‌شود: توزیع نرمال استاندارد برای حالتی که نمونهٔ تصادفی فاقد مقادیر دورافتاده است و توزیع اسلش برای حالتی که نمونهٔ تصادفی از مقادیر دورافتاده برخوردار است. در بیان کوتاهی از توزیع اسلش باید گفت، این توزیع همانند توزیع نرمال استاندارد حول مبدأ متقارن است، اما از دمهای سنگینی نسبت به آن برخوردار است که همین ویژگی سبب وجود مقادیری در میان مشاهدات نمونه می‌شود که دورافتاده بودن آنها نسبت به مشاهدات حاصل از توزیع نرمال استاندارد کاملاً واضح است. در شکل ۱ توابع چگالی توزیع‌های نرمال استاندارد و اسلش رسم شده است.

بنابراین به روش کم‌بیشینه، مسأله تصمیمی حل می‌شود که تابع زیان آن مطابق با درجه بدی مقیاس‌بندی شده باشد و بر این اساس مقدار برش بهینه، که با λ^* نشان داده می‌شود، مقداری است که در رابطه

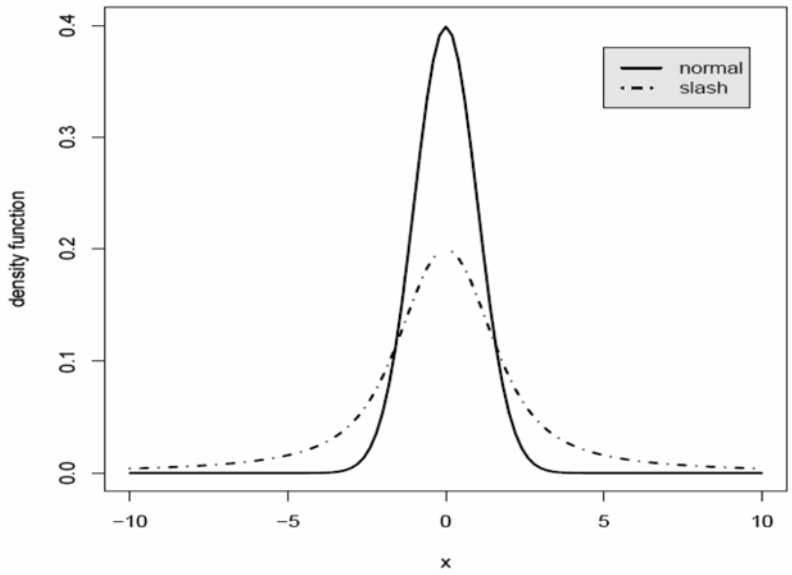
$$\max(b_{SC}^G(\lambda^*), b_{SC}^S(\lambda^*)) = \min_{\lambda} \max(b_{SC}^G(\lambda), b_{SC}^S(\lambda))$$

صدق می‌کند و در آن $b_{SC}^G(.)$ و $b_{SC}^S(.)$ به ترتیب درجات بدی مقیاس‌بندی شده توزیع‌های نرمال استاندارد و اسلش را نشان می‌دهند. به عبارت دیگر به ازای مقادیر برش مختلف، مقداری انتخاب می‌شود که ماکسیمم درجه بدی مقیاس‌بندی شده آن در بین دو توزیع نرمال استاندارد و اسلش کمترین باشد.

۳ نحوه شبیه‌سازی

زمانی که نمونه تصادفی دقیقاً بر اساس توزیع نرمال استاندارد به‌دست آمده است، برای این‌که میانگین نمونه، پس از حذف مقادیر دورافتاده، برآورد دقیقی را از پارامتر

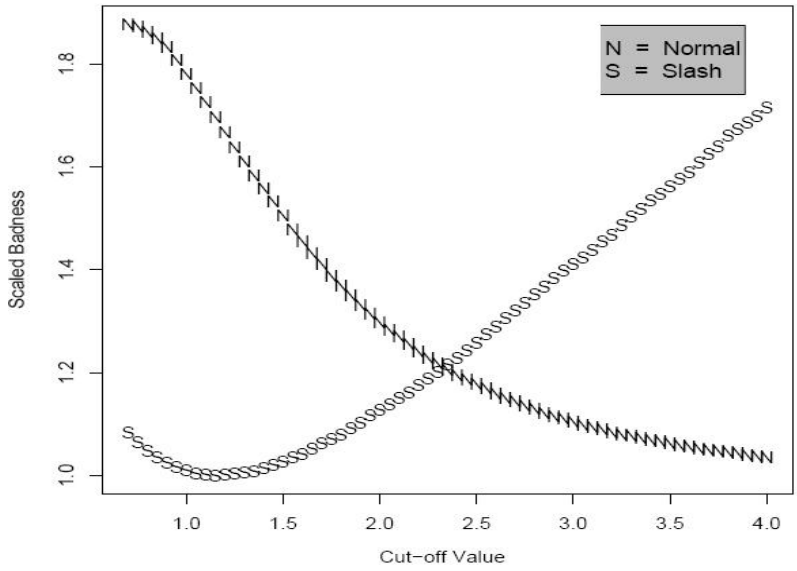
[DOR: 20.1001.1.17358183.1388.3.2.15.8]



شکل ۱: نمودار تابع چگالی توزیع های نرمال استاندارد و اسلش

μ حاصل کند، منطقی است که مقدار برش λ بیشترین مقدار ممکن در نظر گرفته شود تا تمامی مشاهدات نمونه در برآورد پارامتر μ شرکت داشته باشند. در این صورت هر چه λ بزرگتر باشد، درجه بدی $b^G(\lambda)$ کاهش خواهد یافت (همین طور $b^G_{SC}(\lambda)$). بنابراین $b^G_{SC}(\lambda)$ تابعی نزولی بر حسب λ است. اما اگر مشاهدات نمونه دارای نقاط دورافتاده باشند (که در اینجا برای تولید نمونه‌های تصادفی دارای نقاط دورافتاده از توزیع اسلش استفاده شده است.) مقادیر بزرگ λ باعث می‌شود، مشاهدات دورافتاده همچنان در نمونه باقی بمانند و دقت برآورد پارامتر μ را کاهش دهند. در این صورت با افزایش λ، مقدار $b^S(\lambda)$ (و در نتیجه مقدار $b^S_{SC}(\lambda)$) افزایش خواهد یافت. بنابراین $b^S_{SC}(\lambda)$ تابعی صعودی بر حسب λ است. در شکل ۲ مبتنی بر شبیه‌سازی برای حجم نمونهٔ برابر ۵ مقادیر برآورد $b^G_{SC}(\lambda)$ و $b^S_{SC}(\lambda)$ در ازای λ های مختلف رسم شده‌اند، که نحوه رسم آن در بخش بعد توضیح داده خواهد شد. در این شکل صعودی بودن b^S_{SC} و نزولی بودن b^G_{SC} به استثنای همسایگی عدد یک، کاملاً مشهود است.

[DOR: 20.1001.1.17358183.1388.3.2.15.8]



شکل ۲: نمودار مقادیر شبیه‌سازی‌شده $b^G_{SC}(\lambda)$ و $b^S_{SC}(\lambda)$ در ازای $\underline{\lambda} = \{۰/۷, ۰/۷۵, \cdots, ۳/۹۵, ۴\}$ و $n = ۵$

بر اساس شکل ۲ محل تلاقی توابع $b^G_{SC}(\lambda)$ و $b^S_{SC}(\lambda)$ بر روی محور افقی، مقدار برش بهینه λ^* را مشخص می‌کند. برای به‌دست آوردن λ^* رابطه

$$\delta(\lambda) = b^G_{SC}(\lambda) - b^S_{SC}(\lambda)$$

تعریف می‌شود. مقدار λ^* از حل معادله $\delta(\lambda^*) = ۰$ حاصل می‌شود. اما چون فرم بستهٔ توابع $b^G_{SC}(\lambda)$ و $b^S_{SC}(\lambda)$ نامعلوم هستند، مقدار آن‌ها با شبیه‌سازی و به ازای λ های معینِ $\{\lambda_۱, \ldots, \lambda_k\}$ برآورد می‌شوند.

ابتدا $N_۱$ نمونه تصادفی به حجم n از توزیع نرمال استاندارد و $N_۲$ نمونه تصادفی به حجم n از توزیع اسلش تولید می‌شوند. نکته قابل توجه آن است که به منظور کاهش واریانس تفاوت بین متوسط درجات بدی مقیاس بندی شده در هر یک از نمونه‌های نرمال استاندارد و اسلش، بهتر است ابتدا نمونه‌های اسلش را تولید

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

کرده و سپس با استفاده از رابطه

$$z = \Phi^{-1}(F_S(s)) \qquad (1)$$

نمونه های نرمال استاندارد را به دست آوریم که در آن F_S تابع توزیع اسلش و $\Phi^{-1}(\alpha)$ چندک αام متغیر تصادفی نرمال استاندارد است. واضح است که $z \sim N(0, 1)$.

بنابراین ابتدا N نمونه تصادفی به حجم n از توزیع اسلش تولید می‌شود و پس از آن بر اساس تبدیل (۱)، N نمونه تصادفی دارای توزیع نرمال استاندارد و همبسته با نمونه‌های اسلش به‌دست می‌آید. سپس درجات بدی مقادیر برش $\lambda_1, \ldots, \lambda_k$ به ازای هر یک از نمونه‌های تصادفی از دو توزیع نرمال استاندارد و اسلش به صورت

$$\begin{pmatrix} B_1^G(\lambda_1) & B_1^G(\lambda_2) & \ldots & B_1^G(\lambda_k) \\ B_2^G(\lambda_1) & B_2^G(\lambda_2) & \ldots & B_2^G(\lambda_k) \\ \vdots & \vdots & \ddots & \vdots \\ B_N^G(\lambda_1) & B_N^G(\lambda_2) & \ldots & B_N^G(\lambda_k) \end{pmatrix}$$

و

$$\begin{pmatrix} B_1^S(\lambda_1) & B_1^S(\lambda_2) & \ldots & B_1^S(\lambda_k) \\ B_2^S(\lambda_1) & B_2^S(\lambda_2) & \ldots & B_2^S(\lambda_k) \\ \vdots & \vdots & \ddots & \vdots \\ B_N^S(\lambda_1) & B_N^S(\lambda_2) & \ldots & B_N^S(\lambda_k) \end{pmatrix}$$

محاسبه می‌شود. بر این اساس متوسط درجات بدی برای هر λ_i در توزیع‌های نرمال استاندارد و اسلش به ترتیب برابر است با

$$\begin{cases} B^G(\lambda_i) = \frac{1}{N}\sum_{j=1}^{N} B_j^G(\lambda_i) \\ \\ B^S(\lambda_i) = \frac{1}{N}\sum_{j=1}^{N} B_j^S(\lambda_i) \end{cases} \qquad i = 1, \ldots, k$$

برای محاسبه درجات بدی مقیاس‌بندی شده، لازم است مینیمم درجه بدی در هر دو توزیع نرمال استاندارد و اسلش به‌دست آید.

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

لم ۱ : مینیمم درجه بدی بر اساس نمونه‌های تصادفی به حجم n از توزیع نرمال استاندارد برابر با $\frac{۱}{n}$ است.

برهان : در توزیع نرمال استاندارد، میانگین نمونه، برآوردگری نااریب برای میانگین جامعه است که کمترین واریانس را در بین برآوردگرهای نااریب داراست. در نتیجه حذف حتی یک مشاهده، درجه بدی را افزایش می‌دهد. به عبارت بهتر، مینیمم درجه بدی زمانی اتفاق می‌افتد که هیچ یک از مشاهدات نمونه حذف نشود. بنابراین برای حفظ همه مشاهدات نمونه، λ_m^G را باید برابر بینهایت فرض کرد و در این صورت b_m^G معادل با واریانس میانگین نمونه خواهد بود. زیرا

$$b_m^G = b^G(\infty) = E(B(\infty)) = E(\bar{X})^۲ = Var(\bar{X}) = \frac{۱}{n}$$

اما مقدار دقیق λ_m^S معلوم نیست و به روش شبیه‌سازی دو مقدار λ_m^S و b_m^S برآورد می‌شود. با فرض

$$B_m^{S'} = \min\left(B^S(\lambda_۱), B^S(\lambda_۲), \ldots, B^S(\lambda_k)\right)$$

$B_m^{S'}$ برآوردگری اریب از b_m^S است. $\hat{\lambda}_m^S$ بر اساس رابطه

$$B^S(\hat{\lambda}_m^S) = B_m^{S'}$$

تعریف می‌شود. حال بر اساس N نمونه تصادفی به حجم n از توزیع اسلش، شبیه‌سازی مستقلی انجام می‌گیرد و درجه بدی در $\hat{\lambda}_m^S$ به‌دست می‌آید که برآوردگری نااریب برای $b^S(\hat{\lambda}_m^S)$ است و آن را با B_m^S نشان داده می‌شود. این برآوردگر همچنان برآوردگری اریب برای b_m^S است. اما اریبی آن از $B_m^{S'}$ کمتر است. (سرمد، ۲۰۰۶) در مرحله اول شبیه‌سازی بر اساس دامنه وسیعی از مقادیر برش، مقادیر

$$\hat{b}_{SC}^G(\lambda_i) = \frac{B^G(\lambda_i)}{b_m^G} \quad ; \quad \hat{b}_{SC}^S(\lambda_i) = \frac{B^S(\lambda_i)}{B_m^S}$$

را در یک نمودار رسم کرده و با حل معادله $\delta(\lambda) = ۰$ مقدار تقریبی λ^* به‌دست می‌آید. سپس با استفاده از این λ^* اولیه، بازه محدودتری از مقادیر برش را در نظر گرفته، با تکرار شبیه‌سازی بالا، مقدار دقیق λ^* حاصل می‌گردد.

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

۴ شبیه‌سازی

مراحل شبیه‌سازی برای یافتن λ^* به صورت زیر است:

۱) یافتن $\hat{\lambda}_m^S$

۲) برآورد b_m^S با استفاده از B_m^S

۳) یافتن برآوردی اولیه از λ^*

۴) به‌دست آوردن λ^* دقیق در بازه‌ای محدود حول λ^* اولیه.

مراحل شبیه‌سازی برای نمونه‌های به حجم $n = ۵, ۶, \ldots, ۲۹, ۳۰$ به طور جداگانه انجام می‌گیرد.

نتایج شبیه‌سازی:

الف. برای یافتن $\hat{\lambda}_m^S$ شبیه‌سازی با یک میلیون نمونه تصادفی از توزیع اسلش آغاز شده است. به علت محدودیت حافظه و به منظور افزایش سرعت شبیه‌سازی، ۵ شبیه‌سازی با ۲۰۰ هزار نمونه تصادفی به حجم n در نظر گرفته شده و در هر شبیه‌سازی مقادیر $\hat{\lambda}_m^S$ محاسبه گردیده است. در نهایت میانگین مقادیر $\hat{\lambda}_m^S$ در ۵ مرحله شبیه‌سازی به عنوان مقدار نهایی معرفی شده است.

برای به‌دست آوردن مقادیر $\hat{\lambda}_m^S$ در هر مرحله شبیه‌سازی به این صورت عمل می‌شود که ابتدا برای هر حجم نمونه دامنه‌ای از مقادیر برش در همسایگی عدد یک در نظر گرفته شده است. به عنوان مثال برای $n = ۵$ مجموعهٔ

$$\underline{\lambda} = \{۰/۶, ۰/۶۱, \ldots, ۱/۳۹, ۱/۴۰\}$$

برای $\underline{\lambda}$ فرض گردیده است. مقادیر ابتدا و انتهای این مجموعه بر اساس شبیه‌سازی‌های اولیه با تعداد نمونه‌های تصادفی کم تخمین زده می‌شود. سپس برای مجموعه $\underline{\lambda}$ درجات بدی بر اساس نمونه‌های تصادفی از توزیع اسلش، آن‌طور که در

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

جدول ۱: مقادیر $\hat{\lambda}_m^S$ و B_m^S متناظر با حجم نمونه

حجم نمونه	$\hat{\lambda}_m^S$	B_m^S
۵	۰/۹۷۶۲	۱/۹۴۰۲
۶	۰/۸۳۶۴	۱/۴۱۰۲
۷	۱/۱۰۴۰	۱/۱۰۱۶
۸	۱/۰۵۲۱	۰/۹۰۳۷
۹	۱/۱۸۴۳	۰/۷۷۵۱
۱۰	۱/۱۷۰۰	۰/۶۷۱۰
۱۱	۱/۲۵۶۸	۰/۶۰۰۵
۱۲	۱/۲۷۸۶	۰/۵۳۲۷
۱۳	۱/۳۲۰۰	۰/۴۸۹۳
۱۴	۱/۳۱۳۴	۰/۴۴۴۱
۱۵	۱/۳۵۲۵	۰/۴۱۳۶
۱۶	۱/۳۱۹۲	۰/۳۷۹۵
۱۷	۱/۳۸۹۰	۰/۳۵۹۱
۱۸	۱/۳۷۰۶	۰/۳۳۲۹
۱۹	۱/۴۳۱۴	۰/۳۱۶۲
۲۰	۱/۳۹۳۲	۰/۲۹۵۱
۲۱	۱/۴۴۰۲	۰/۲۸۲۴
۲۲	۱/۴۱۶۸	۰/۲۶۶۲
۲۳	۱/۴۵۲۱	۰/۲۵۵۵
۲۴	۱/۴۳۰۴	۰/۲۴۲۶
۲۵	۱/۴۵۶۴	۰/۲۳۲۸
۲۶	۱/۴۴۱۴	۰/۲۲۱۷
۲۷	۱/۴۶۷۳	۰/۲۱۴۴
۲۸	۱/۴۵۵۴	۰/۲۰۴۴
۲۹	۱/۴۸۲۳	۰/۱۹۷۸
۳۰	۱/۴۶۸۴	۰/۱۹۰۴

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

مرحله سوم توضیح داده خواهد شد، شبیه‌سازی می‌شود. مقداری از λ که دارای کمترین درجه بدی است، به عنوان $\hat{\lambda}_m^S$ معرفی می‌گردد.

ب. با استفاده از $\hat{\lambda}_m^S$، B_m^S براساس یک میلیون نمونه تصادفی متفاوت از توزیع اسلش شبیه‌سازی می‌شود. (نحوه محاسبه این درجات بدی به طور کامل در مرحله سوم شبیه‌سازی شرح داده خواهد شد.) مقادیر $\hat{\lambda}_m^S$ و B_m^S به ازای n های متفاوت در جدول ۱ درج شده است.

ج. برای یافتن λ^* اولیه، ۵۰۰ هزار نمونه تصادفی به حجم n از توزیع اسلش تولید شده و سپس با استفاده از رابطه (۱) ۵۰۰ هزار نمونه تصادفی برای توزیع نرمال استاندارد به‌دست آمده است. در مرحله بعد مقادیر Z اصلاح‌شدهٔ هر نمونه n تایی از توزیع‌های اسلش و نرمال استاندارد محاسبه می‌شود.

پیش‌تر بیان شد که ایگلیویکس و هوگلین (۱۹۹۳) در محاسبه M_i ها به ازای هر حجم نمونه، ثابت d را برابر ۰/۶۷۴۵ فرض کردند که این عدد بر اساس وقتی است که $n \to \infty$. اما در نمونه‌های کوچکتر بر اساس شبیه‌سازی از توزیع نرمال استاندارد، مقدار این ثابت در جدول ۲ داده شده است. نحوه به‌دست آوردن این ثابت به این صورت است که ابتدا ۵۰۰۰۰ نمونهٔ تصادفی به حجم n از توزیع نرمال استاندارد تولید و سپس مقدار MAD هر نمونهٔ تصادفی محاسبه شده است. مقدار شبیه‌سازی شدهٔ ثابت d برابر با میانگین مقادیر MAD به‌دست آمده در هر نمونهٔ تصادفی می‌باشد.

بنابراین در محاسبه مقادیر Z اصلاح‌شده به ازای هر حجم نمونه، ثابت d متناظر با آن در نظر گرفته می‌شود.

در مرحله بعد دامنه وسیعی از مقادیر برش در نظر گرفته شده است:

$$\underline{\lambda} = \{۱, ۱/۱, \ldots, ۳/۹, ۴\}$$

سپس برای هر مقدار از $\underline{\lambda}$ درجه بدی نمونه‌های تصادفی تولید شده از دو توزیع نرمال استاندارد و اسلش محاسبه می‌شود. بدین طریق که مقادیری در نمونه را که $|M_i| > \lambda$، از نمونه کنار گذاشته، میانگین مشاهدات باقی‌مانده در نمونه به توان دو می‌رسد. سپس با استفاده از مقادیر $b_m^G = \frac{۱}{n}$ و $\hat{b}_m^S = B_m^S$، که در مرحله قبل به‌دست

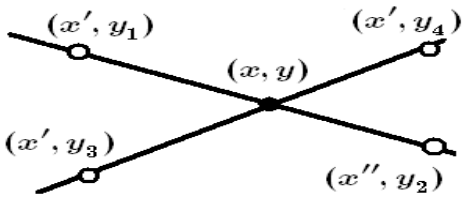
[DOR: 20.1001.1.17358183.1388.3.2.15.8]

جدول ۲: مقادیر ثابت d متناظر با حجم نمونه

حجم نمونه	d	حجم نمونه	d
۵	۰/۵۵۴۶	۶	۰/۵۶۷۶
۷	۰/۵۹۲۴	۸	۰/۵۹۸۵
۹	۰/۶۱۲۵	۱۰	۰/۶۱۵۶
۱۱	۰/۶۲۴۷	۱۲	۰/۶۲۶۵
۱۳	۰/۶۳۲۷	۱۴	۰/۶۳۴۰
۱۵	۰/۶۳۸۵	۱۶	۰/۶۳۹۲
۱۷	۰/۶۴۳۰	۱۸	۰/۶۴۳۶
۱۹	۰/۶۴۶۵	۲۰	۰/۶۴۶۹
۲۱	۰/۶۴۹۲	۲۲	۰/۶۴۹۵
۲۳	۰/۶۵۱۵	۲۴	۰/۶۵۱۸
۲۵	۰/۶۵۳۳	۲۶	۰/۶۵۳۶
۲۷	۰/۶۵۴۹	۲۸	۰/۶۵۵۵
۲۹	۰/۶۵۶۱	۳۰	۰/۶۵۶۷

آمده است، درجات بدیِ مقیاس‌بندی شده محاسبه می‌شوند.

از آنجا که توابع $B^S_{SC}(\lambda)$ و $B^G_{SC}(\lambda)$ به ترتیب بر حسب λ صعودی و نزولی هستند، با رسم آن‌ها در یک نمودار، دو مقدار برشی را که در آن‌ها $\delta(\lambda)$ تغییر علامت می‌دهد مشخص می‌شوند. اگر این دو مقدار برش با x' و x'' نشان داده شوند و مقادیر y_1 ، y_2 ، y_3 و y_4 به ترتیب نشان دهندهٔ $B^G_{SC}(x')$ و $B^G_{SC}(x'')$ و $B^S_{SC}(x')$ و $B^S_{SC}(x'')$ باشند، مقدار اولیه λ^* همان محل تقاطع دو پاره‌خط رسم شده در شکل ۳ می‌باشد که بر اساس رابطه (۲) قابل محاسبه است و در آن علامت $|A|$ نشان دهندهٔ دترمینان ماتریس A است.



شکل ۳: در این شکل مقدار x همان مقدار برش بهینه مورد نظر است.

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

$$x = \frac{\begin{vmatrix} \begin{vmatrix} x' & y_1 \\ x'' & y_2 \end{vmatrix} & x' - x'' \\ \\ \begin{vmatrix} x' & y_3 \\ x'' & y_4 \end{vmatrix} & x' - x'' \end{vmatrix}}{\begin{vmatrix} x' - x'' & y_1 - y_2 \\ x' - x'' & y_3 - y_4 \end{vmatrix}} \qquad (۲)$$

د. فرض کنید λ^* اولیه برابر ۲/۳۶ گردد، در این صورت دامنه محدودتر زیر برای مقادیر برش در نظر گرفته می‌شود:

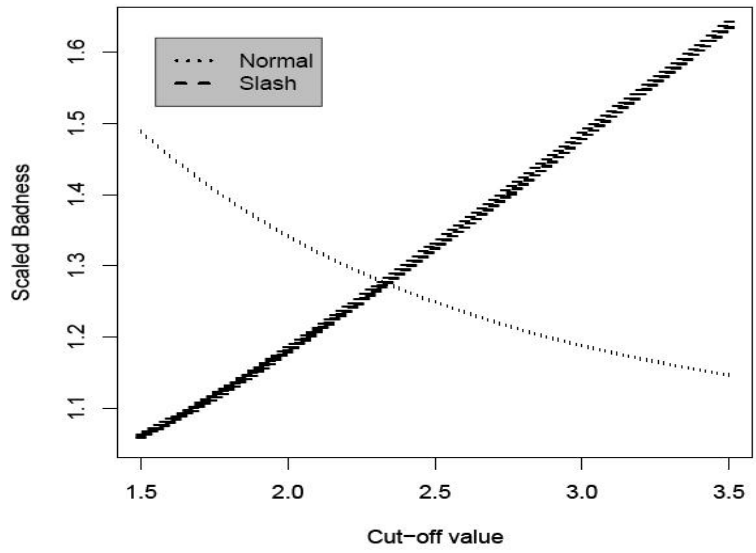
$$\underline{\lambda} = \{۲/۳۰, ۲/۳۱, \ldots, ۲/۳۹, ۲/۴۰\}$$

با تولید ۱۰ میلیون نمونه تصادفی به حجم n از دو توزیع نرمال استاندارد و اسلش مبتنی بر رابطه (۱) و با استفاده از مقادیر برش بردار $\underline{\lambda}$، شبیه‌سازی مرحله قبل تکرار شده و λ^* نهایی به‌دست می‌آید.

در مرحله آخر به منظور اطمینان از پایداری نتایج به‌دست آمده مراحل شبیه‌سازی چندین مرتبه تکرار می‌شود. به عنوان مثال در شکل ۴ به ازای $n = ۵$، نتایج حاصل از سه مرتبه تکرار شبیه‌سازی با یک میلیون نمونه تصادفی رسم شده است. همان‌طور که ملاحظه می‌شود، نتایج در توزیع نرمال استاندارد بسیار مقاوم و پایدار است، در حالی که در توزیع اسلش تفاوت جزئی بین سه مرتبه شبیه‌سازی وجود دارد که به نظر قابل چشم‌پوشی است.

در جدول ۳ به ازای n های متفاوت، مقادیر λ^* نهایی به همراه تعداد تکرار شبیه‌سازی برای رسیدن به نتیجه‌ای پایدار درج شده‌است. در نگاه اول نمی‌توان گفت با افزایش حجم نمونه، مقادیر λ^* دارای روند خاصی هستند. اما اگر حجم نمونه‌های زوج و فرد به صورت جداگانه بررسی شوند، مشاهده می‌گردد که در هر یک از این دو دسته داده، λ^* ها به صورت جداگانه افزایش می‌یابند. شکل ۵ مؤید این مطلب است.

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

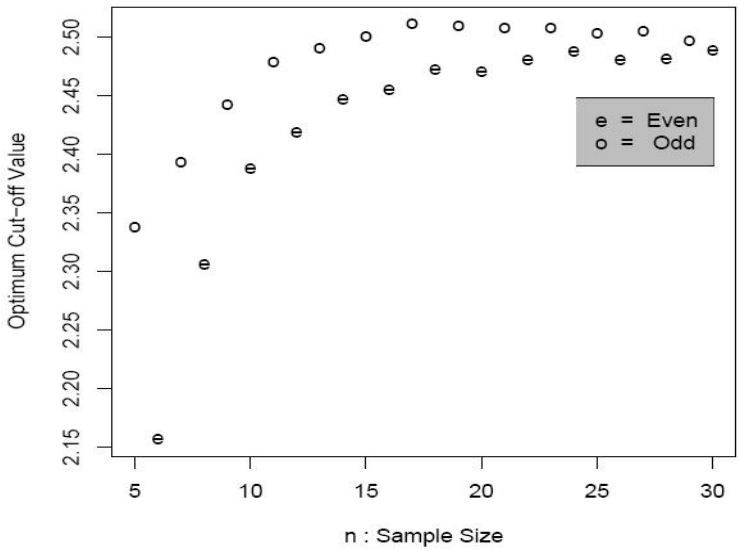


شکل ۴: نمودار ۳ مرتبه تکرار شبیه‌سازی مقادیر $B^S_{SC}(\lambda)$ و $B^G_{SC}(\lambda)$ برای $n = 5$ به جهت اطمینان از پایداری نتایج

جدول ۳: مقادیر λ^* (نقطه برش) نهایی وابسته به حجم نمونه به همراه تعداد شبیه‌سازی‌های صورت گرفته برای رسیدن به نتیجه‌ای پایدار

حجم نمونه	λ^*	N	حجم نمونه	λ^*	N
۵	۲/۳۳۷۷	۳×۱۰۰۰۰۰۰	۶	۲/۱۵۵۸	۳×۱۰۰۰۰۰۰
۷	۲/۳۹۳۲	۴×۶۰۰۰۰۰	۸	۲/۳۰۵۰	۴×۶۰۰۰۰۰
۹	۲/۴۴۱۹	۴×۶۰۰۰۰۰	۱۰	۲/۳۸۷۲	۴×۶۰۰۰۰۰
۱۱	۲/۴۷۸۳	۴×۲۰۰۰۰۰	۱۲	۲/۴۱۸۵	۴×۲۰۰۰۰۰
۱۳	۲/۴۹۰۰	۴×۲۰۰۰۰۰	۱۴	۲/۴۴۶۴	۴×۲۰۰۰۰۰
۱۵	۲/۵۰۰۵	۴×۲۰۰۰۰۰	۱۶	۲/۴۵۵۰	۴×۲۰۰۰۰۰
۱۷	۲/۵۱۱۵	۴×۲۰۰۰۰۰	۱۸	۲/۴۷۱۶	۴×۲۰۰۰۰۰
۱۹	۲/۵۰۹۵	۴×۲۰۰۰۰۰	۲۰	۲/۴۷۰۰	۴×۲۰۰۰۰۰
۲۱	۲/۵۰۷۸	۴×۲۰۰۰۰۰	۲۲	۲/۴۷۹۷	۴×۲۰۰۰۰۰
۲۳	۲/۵۰۷۸	۴×۲۰۰۰۰۰	۲۴	۲/۴۸۷۷	۴×۲۰۰۰۰۰
۲۵	۲/۵۰۲۹	۴×۲۰۰۰۰۰	۲۶	۲/۴۸۰۲	۴×۲۰۰۰۰۰
۲۷	۲/۵۰۴۹	۴×۲۰۰۰۰۰	۲۸	۲/۴۸۱۴	۴×۲۰۰۰۰۰
۲۹	۲/۴۹۶۲	۴×۲۰۰۰۰۰	۳۰	۲/۴۸۸۲	۴×۲۰۰۰۰۰

[DOR: 20.1001.1.17358183.1388.3.2.15.8]



شکل ۵: مقادیر برش بهینه در ازای حجم نمونه‌های زوج و فرد

تذکر ۱ : در نمونه‌های تصادفی تولید شده از توزیع اسلش، گاه مقادیری به بزرگی ۷۰۰۰۰۰ مشاهده می‌شود که تأثیر بسیاری بر درجات بدی مقیاس‌بندی شدهٔ این توزیع دارد. تولید چنین مقادیری ناپایداری شبیه‌سازی‌ها را در پی خواهد داشت. هدف از به کارگیری توزیع اسلش، تولید مقادیری است که در مقایسه با توزیع نرمال استاندارد کاملاً دورافتاده به نظر رسند. اما لزومی ندارد مقادیر دورافتاده به بزرگی اعدادی همچون ۷۰۰۰۰۰ باشند و حتی مشاهداتی نظیر ۲۰ نیز برای توزیع نرمال استاندارد بسیار غیرمنتظره است. بنابراین می توان مشاهدات فوق‌العاده دورافتاده را از نمونه‌های اسلش حذف کرد. برای این منظور از توزیع اسلش بریده‌شده با نقطه برش M که به صورت

$$F_{S_{trM}}(s) = \frac{F_S(s) - F_S(-M)}{F_S(M) - F_S(-M)}$$

تعریف می‌شود، برای تولید نمونهٔ تصادفی حاوی نقاط دورافتاده استفاده خواهد شد. از آنجا که در توزیع اسلش احتمال اینکه قدر مطلق مشاهده‌ای از ۸۰ بیشتر باشد، کمتر از یک درصد است، لذا در توزیع اسلش بریده شده نقطه برش برابر ۸۰

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

در نظر گرفته خواهد شد.

۵ مقایسه استفاده از مقادیر برش وابسته به حجم نمونه با مقدار برش ثابت ۳/۵

پس از محاسبه مقادیر M_i مشاهدات، می‌توان از مقادیر برش وابسته به حجم نمونه یا از مقدار ثابت ۳/۵ برای شناسایی نقاط دورافتاده استفاده کرد. اما مسأله اساسی مقایسه عملکرد مقادیر برش وابسته به حجم نمونه با مقدار برش ثابت ۳/۵ در ردیابی نقاط دورافتاده است. چون برای یافتن مقادیر λ_n^* از روش مینیماکس استفاده شده است، می‌توان از تابع زیان این روش به عنوان معیاری برای مقایسه کارآیی مقادیر برش بهره برد.

همانطور که از جدول ۳ ملاحظه می‌شود به ازای هر حجم نمونه، $\lambda_n^* < ۳/۵$ است. از طرفی چون $b_{SC}^S(\lambda)$ و $b_{SC}^G(\lambda)$ به ترتیب توابعی صعودی و نزولی بر حسب λ هستند، داریم

$$n = ۵, ۶, \ldots, ۳۰ \quad : \quad \begin{cases} b_{SC}^G(۳/۵) < b_{SC}^G(\lambda_n^*) \\ \\ b_{SC}^S(۳/۵) > b_{SC}^S(\lambda_n^*) \end{cases} \qquad (۳)$$

اما λ_n^* مقداری است که به ازای آن $b_{SC}^G(\lambda_n^*) = b_{SC}^S(\lambda_n^*)$. بنابراین با فرض

$$b_{SC}(\lambda_n^*) = b_{SC}^G(\lambda_n^*) = b_{SC}^S(\lambda_n^*)$$

داریم

$$b_{SC}^G(۳/۵) < b_{SC}(\lambda_n^*) < b_{SC}^S(۳/۵)$$

حال اگر فرض شود

$$b_{SC}(\lambda_n^*) - b_{SC}^G(۳/۵) = d^G(\lambda_n^*) \quad , \quad b_{SC}^S(۳/۵) - b_{SC}(\lambda_n^*) = d^S(\lambda_n^*)$$

بر اساس رابطه (۳) بدیهی است که $d^S(\lambda_n^*) > ۰$, $d^G(\lambda_n^*) > ۰$.

به راحتی می‌توان همانند شبیه‌سازی‌های گذشته مقادیر $B_{SC}^G(۳/۵)$ و $B_{SC}^S(۳/۵)$ و $B_{SC}(\lambda_n^*)$ را به‌دست آورد و در نهایت $d^G(\lambda_n^*)$ و $d^S(\lambda_n^*)$ را برآورد

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

نمود. (که مقدار شبیه‌سازی شده آن‌ها به ترتیب به صورت $D^G(\lambda_n^*)$ و $D^S(\lambda_n^*)$ فرض می‌شود.) در جدول ۴ مقادیر $D^G(\lambda_n^*)$ و $D^S(\lambda_n^*)$ به ازای n های مختلف درج شده‌اند.

جدول ۴: مقادیر $D^G(\lambda_n^*)$ و $D^S(\lambda_n^*)$ وابسته به حجم نمونه

حجم نمونه	$D^G(\lambda_n^*)$	$D^S(\lambda_n^*)$	حجم نمونه	$D^G(\lambda_n^*)$	$D^S(\lambda_n^*)$
۵	۰/۱۲۸۶	۰/۳۶۳۵	۶	۰/۱۵۸۷	۰/۴۴۶۵
۷	۰/۱۳۵۰	۰/۳۴۷۲	۸	۰/۱۴۶۰	۰/۳۷۸۵
۹	۰/۱۳۳۷	۰/۳۱۳۸	۱۰	۰/۱۳۶۳	۰/۳۳۵۵
۱۱	۰/۱۲۹۵	۰/۲۹۴۱	۱۲	۰/۱۳۶۵	۰/۳۱۲۲
۱۳	۰/۱۲۷۴	۰/۲۸۶۱	۱۴	۰/۱۳۲۴	۰/۲۹۳۵
۱۵	۰/۱۲۶۰	۰/۲۷۷۰	۱۶	۰/۱۳۱۹	۰/۲۹۱۴
۱۷	۰/۱۲۲۹	۰/۲۶۷۷	۱۸	۰/۱۲۶۹	۰/۲۸۱۷
۱۹	۰/۱۲۳۱	۰/۲۶۴۹	۲۰	۰/۱۲۸۴	۰/۲۸۱۱
۲۱	۰/۱۲۲۳	۰/۲۶۵۰	۲۲	۰/۱۲۳۴	۰/۲۷۴۳
۲۳	۰/۱۲۳۴	۰/۲۶۴۲	۲۴	۰/۱۲۳۰	۰/۲۶۸۷
۲۵	۰/۱۲۲۱	۰/۲۵۸۵	۲۶	۰/۱۲۴۶	۰/۲۷۰۳
۲۷	۰/۱۲۱۱	۰/۲۵۹۵	۲۸	۰/۱۲۴۷	۰/۲۶۹۳
۲۹	۰/۱۲۲۰	۰/۲۶۴۳	۳۰	۰/۱۲۱۹	۰/۲۶۴۴

چون به ازای هر n، $D^S(\lambda_n^*) \gg D^G(\lambda_n^*)$ است، بنابراین استفاده از مقدار برش λ_n^* بر ۳/۵ ترجیح داده می‌شود. زیرا با وجود این که در نمونه‌های تصادفی از توزیع نرمال استاندارد، درجه بدی ۳/۵ از λ_n^* کمتر است، اما در نمونه‌های تصادفی از توزیع اسلش درجه بدی از λ_n^* به ۳/۵ افزایش می‌یابد و این افزایش درجه بدی در توزیع اسلش چند برابر کاهش درجه بدی در توزیع نرمال استاندارد است. بنابراین منطقی است که برای شناسایی نقاط دورافتاده مقدار برش λ_n^* بر ۳/۵ ترجیح داده شود.

بحث و نتیجه‌گیری

استفاده از مقادیر مختلف نقطه‌ی برش متناظر با حجم نمونه در روش Z اصلاح شده، دارای کارآیی بیشتری نسبت به نقطهٔ برش ثابت ۳/۵ که پیشتر توسط

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

ایگلویکس و هوگلین (۱۹۹۳) معرفی شده بود، می‌باشد. همچنین توصیه می‌شود مقدار ثابت d را نیز متناظر با حجم نمونه در نظر گرفت که قبلاً این ثابت برابر ۰/۶۷۴۵ (متناظر با حجم نمونه زیاد) بوده است.

تقدیر و تشکر

نویسندگان از اصلاحات پیشنهادی داوران محترم که موجب بهبود این مقاله گردید، کمال تشکر و قدردانی را دارند.

مراجع

Barnett, V. and Lewis, T. (1984), *Outliers in Statistical Data*, John Wiely, New York.

Beckman, R. J. and Cook, R. D. (1983), Outliers, *Technometrics*, **25**, 119-149.

Collett, D. and Lewis, T. (1976), The Subjective Nature of Outlier Rejection Procedures, *Applied Statistics*, **25**, 228-237.

Hampel, F. R. (1974), The Influence Curve and its Role in Robust Estimation, *American Statistical Association*, **69**, 383-393.

Hawkins, D. M. (1980), *Identification of Outliers*, Chapman and Hall, New York.

Iglewicz, B. and Hoaglin, D. C. (1993), *How to Detect and Handle Outliers*, Quality Press, Wisconsin.

Rogers, W. H. and Tukey, J. W. (1972), Understanding Some Long-Tailed Symmetrical Distributions, *Statistica Neerlandica*, **26**, 211-226.

[DOR: 20.1001.1.17358183.1388.3.2.15.8]

Sarmad M. (2006), *Robust Data Analysis for Factorial Experimental Designs: Improved Methods and Software*, Ph.D. Thesis. University of Durham, England

Shiffler, R. E. (1988), Maximum Z Scores and Outliers, *The American Statistician*, **42**, 79-80.

Yatracos, G. Y. (1991), A Note on Tukey's Polyefficiency, *Biometrika*, **78**, 702-703.

Journal of Statistical Sciences, Autumn & Winter 2009 **229**

Vol. 3, No. 2, pp 119-139.

Detecting Outliers in Normal Data Using Modified Z-Scores

Ahmadi, M. V. and Sarmad, M.

Department of Statistics, Ferdowsi University, Mashhad, Iran.

Abstract: Because of importance and popularity of the Normal distribution, the samples based on this distribution has been considered and the outliers are identified using cut-off values which are dependent on the sample size. A decision problem has been structured to obtain the optimal cut-off value. The problem is solved by a simulation study with a minimax rule.

Keywords: Badness, Slash Distribution, Z-Score, Modified Z-Score.

Mathematics Subject Classification (2000): 62G10, 62G30