دانشکده علوم ریاضی

گروه آمار

پایان نامه جهت اخذ درجه کارشناسی ارشد در رشته آمار ریاضی

عنوان پایان نامه

m—برآوردگرهای استوار مکانی و مقیاسی

استاد راهنما

دکتر مجید سرمد

استاد مشاور

دکتر ناصر رضا ارقامی

به اهتمام

محمد ولی احمدی

شهریور ۸۸

مـا بدان مقصـد عـالی نتـوانیـم رسیـد
هم مگر پیش نهد لطف شما گامی چند

پروردگار بزرگ را سپاس می‌گزارم، که توفیق نگارش این رساله را به بنده‌ی حقیر خویش، عنایت فرموده و همواره مرا مورد لطف و رحمت بی‌پایان خود قرار داده است.

رساله‌ی حاضر شرح شاگردی این‌جانب در محضر بزرگانی بوده است که خود را پیوسته مرهون عنایت‌ها و تشویق‌های ایشان می‌دانم و هم اکنون بر خود لازم می‌دانم که از آنان قدردانی کنم.

نخست از بزرگترین سرمایه‌های زندگیم، **پدر و مادر عزیزم**، که همواره مشوق و پشتیبانم بودند، متشکرم.

از استاد فرهیخته و بزرگوارم، **جناب آقای دکتر مجید سرمد**، که هدایت و راهنمایی این جانب را قبول کردند و همواره با دقت نظر در طول نگارش این رساله، چراغ راهم بودند، تشکر می‌کنم.

از **جناب آقای دکتر ناصر رضا ارقامی**، استاد گرانقدر مشاورم، که در نگارش این رساله، همکاری صمیمانه داشتند، کمال تشکر خود را ابراز می‌کنم.

همچنین از اساتید محترم داور، **جناب آقای دکتر خنجری** و **سرکار خانم دکتر حبیبی** که زحمت مطالعه‌ی این رساله را بر خود هموار کردند و نکات لازم را یادآور شدند، تشکر می کنم.

در پایان، از همه‌ی دوستان و عزیزانی که به نحوی مرا در نوشتن این پایان نامه یاری نمودند، سپاس‌گزاری می‌کنم.

محمّد ولی احمدی
شهریور ماه ۱۳۸۸

باسمه تعالی

مشخصات پایان نامه تحصیلی دانشجویان

دانشگاه فردوسی مشهد

عنوان پایان نامه: $m-$برآوردگرهای استوار مکانی و مقیاسی

نویسنده : محمّد ولی احمدی **استاد راهنما:** مجید سرمد

دانشکده: علوم ریاضی	**گروه:** آمار
رشته تحصیلی: آمار ریاضی	**تاریخ دفاع:** ۱۳۸۸/۶/۲۱
مقطع تحصیلی: کارشناسی ○ کارشناسی ارشد ● دکتری ○	**تعداد صفحات:** ۱۳۹

چکیده پایان نامه:

داده‌های جمع‌آوری شده در قالب موارد شامل یک یا چند نقطه‌ی پرت هستند. نقاطی که از توده‌ی مشاهدات نمونه جدا قرار گرفته‌اند و یا از الگوی کلی که برای داده‌ها فرض کرده‌ایم، پیروی نمی‌کنند. علاوه بر این ثابت شده است برآوردگرهای کلاسیکی که تاکنون از ویژگی‌های بهینه برای برآورد پارامتر جامعه برخوردار بوده‌اند، در اثر انحراف از پذیره‌های اساسی، کارایی و اعتبار خود را از دست می‌دهند. در حالی که با استفاده از روش‌های آماری استوار، در صورت انحراف از پذیره‌های اساسی نیز همچنان می‌توان برآورد دقیقی را از پارامتر جامعه به‌دست آورد.

در این پایان‌نامه علاوه بر این که سعی کرده‌ایم روشی را برای شناسایی نقاط پرت در داده‌های تصادفیِ به‌دست آمده از جامعه‌ای نرمال بیان کنیم، به معرفی و بیان ویژگی‌های دسته‌ای از مشهورترین برآوردگرهای استوار مکانی و مقیاسی با عنوان $m-$برآوردگرها پرداخته‌ایم. علاوه بر این معیارهایی را برای سنجش استواری برآوردگرهای مختلف معرفی کرده و در نهایت $m-$برآوردگری را می‌یابیم که علاوه بر استواری از کارایی بالایی نیز برخوردار باشد. $m-$برآوردگر حاصل به $m-$برآوردگر همپل معروف است.

کلید واژه‌ها: درجه‌ی بدی، مقادیر Z، مقادیر Z اصلاح شده، توزیع اسلش، $m-$برآوردگر، نقطه‌ی شکست، تابع تأثیر، خطای ناخالص حساسیت

امضای استاد راهنما

تاریخ

In the name of God
Dissertation details
Ferdowsi University of Mashhad

Title : *Location and Scale Robust M-estimators*

Student's name : *Mohammad Vali Ahmadi* **Major Professor** : *Majid Sarmad*

School : Mathematical Sciences	**Department:** Statistics
Major : Mathematical Statistics	**Date of defense :** 12/9/2009
Number of pages: 139	**For degree :** BS MS Ph.D

Abstract :

The collected data often contains one or more outlier which are either isolated from the bulk of data or do not follow the general model that we have supposed for data.

Furthermore, it has been proved that the classic estimators which are used to estimate the distribution parameter and have the optimal properties, will lose their efficiency and validity due to deviation from fundamental assumptions. An exact estimation of distribution parameter can be obtained using robust statistical methods,even there exists a deviation from fundamental assumptions .

In this thesis, we want to find a method to identify outlier in random data which is given from a normal distribution. In addition, a class of robust estimator of location and scale as m-estimators is studied. Also, two quantities for measuring robustness of different estimators and finally an m-estimator with high efficiency and robustness, i.e. Hampel's m-estimator will be studied.

Signature of major professor:	**Keywords :** *badness, Z-scores, modified Z-scores, slash distibution, m-estimator, breakdown point, influence function, gross error sensitivity*

فهرست مندرجات

فهرست جداول

فهرست اشکال

۰-۱) پیش گفتار

در بسیاری از داده‌های جمع آوری شده به مشاهدات پرتی برمی‌خوریم که تحلیل داده‌ها را با مشکل مواجه می‌کنند. از این رو آماردانان همواره به دنبال راهی بوده‌اند تا تأثیر این نقاط را بر نتیجه‌گیریها کاهش دهند و در عین حال از اطلاعاتی که نقاط پرت در خود نهفته دارند، به بهترین نحو استفاده کنند. در همین راستا در این پایان‌نامه سعی کرده‌ایم تا روشهایی را برای برخورد با نقاط پرت ارائه دهیم.

در فصل اول، ابتدا مقادیر Z اصلاح شده مشاهدات نمونه را معرفی می‌کنیم. سپس بر اساس روشی کم‌بیشینه که به تفصیل آن را شرح می‌دهیم، سعی می‌کنیم مقادیر برشی را متناسب با حجم مشاهدات نمونه به‌دست آوریم. بر این اساس می‌توانیم از مقایسه‌ی مقادیر Z اصلاح شده‌ی مشاهدات و مقادیر برش به دست آمده، مقادیر پرت را در میان توده مشاهدات تشخیص دهیم.

در فصل دوم ابتدا با ذکر چند مثال به مقایسه برآوردگرهای استوار و غیر استوار می‌پردازیم. سپس دو معیارِ نقطه‌ی شکست و تابع تأثیر را برای سنجش میزان استواری برآوردگرها معرفی می کنیم. در توضیح برآوردگرهای استوار باید گفت که این برآوردگرها در صورتی که مشاهدات نمونه حاوی نقاط پرت باشند، نیز همچنان برآورد صحیحی را از پارامتر جامعه حاصل می‌کنند.

در فصل سوم $m-$برآوردگرهای مکانی و مقیاسی را – که دسته‌ای از برآوردگرهای استوار هستند- معرفی می‌کنیم و چگونگی محاسبه آنها را شرح می‌دهیم. در نهایت نقطه‌ی شکست و تابع تأثیر این برآوردگرها را به دست می آوریم.

در فصل چهارم خواص مجانبی $m-$برآوردگرهای استوار را بررسی می کنیم و توزیع مجانبی آنها را به‌دست می آوریم.

و در فصل پنجم به دنبال $m-$برآوردگری هستیم که علاوه بر استواری از کارایی بالایی نیز برخوردار باشد. در همین راستا نیز $m-$برآوردگر بهینه هَمپِل را معرفی کرده و آن را به دست می آوریم.

فصل ۱

شناسایی نقاط پرت بر اساس مقادیر Z اصلاح شده

در این فصل، ابتدا مقادیر Z اصلاح شده مشاهدات نمونه را معرفی می کنیم. سپس بر اساس روشی کم بیشینه که به تفصیل آن را شرح می دهیم، سعی می کنیم مقادیر برشی را که در تعیین نقاط پرت استفاده می‌شوند، متناسب با حجم مشاهدات نمونه به دست آوریم. بر این اساس می توانیم از مقایسه مقادیر Z اصلاح شده مشاهدات و مقادیر برش به دست آمده، مقادیر پرت را در میان توده مشاهدات تشخیص دهیم.

۱- ۱) مقدمه

اگرچه زمانی که صحبت از نقطه پرت می‌شود، مفهوم واحدی از آن برداشت می‌شود، اما بررسی‌ها نشان می‌دهد، تعاریف متنوع و متفاوتی درباره‌ی این نقطه از سوی آماردانان ارائه شده است.

میزان اعتماد به هر مشاهده به رابطه آن با سایر مشاهداتی بستگی دارد که تحت شرایط یکسان به‌دست آمده‌اند. به عقیده‌ی اکثر محققین مشاهده‌ای که دور از توده مشاهدات قرار می‌گیرد، نقطه پرت محسوب می‌شود. از این میان می توان به بارنت و لویس (۱۹۸۴، صفحه ۴) و هاوکینز (۱۹۸۰، صفحه ۱) اشاره کرد.
بنا بر بارنت و لویس (۱۹۸۴) "نقطه‌ی پرت مشاهده‌ای است که با سایر مشاهدات نمونه ناسازگار و متفاوت است." علاوه بر این در تعریفی مشابه هاوکینز (۱۹۸۰) نقطه پرت 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."
اما بکمن و کوک (۱۹۸۳، صفحه ۱۲۱) عقیده دارند "مشاهده پرت، یا مشاهده‌ای ناهماهنگ و متفاوت است یا مشاهده‌ای آلوده". مشاهده ناهماهنگ "مشاهده‌ای است که حضور آن در میان مشاهدات غیر طبیعی به نظر می‌رسد". اما مشاهده آلوده "مشاهده‌ای است که از توزیعی غیر از توزیع فرض‌شده برای جامعه به‌دست آمده است."

در این نوشتار تعریف بارنت و لویس (۱۹۸۴) را از مشاهده پرت می‌پذیریم. وقوع این دسته از مشاهدات در اکثر نمونه‌های تصادفی تقریباً امری طبیعی است. بنابراین اولین دغدغه و نگرانی محققین پس از جمع‌آوری

مشاهدات نمونه، چگونگی برخورد با نقاط پرت است. به دلیل این‌که شناسایی علل و عوامل وقوع نقاط پرت، کیفیت فرآیند نمونه‌گیری را ارتقا می‌دهد و یا در برخی موارد توزیعی را که برای داده‌ها فرض کرده‌ایم بهبود می‌بخشد، بررسی این مقادیر از اهمیت بسیاری برخوردار است. اما آنچه در ابتدا مهم است، شناسایی این مقادیر در میان توده مشاهدات است.

بررسی ظاهری مشاهدات، اولین راه برای تشخیص مقادیر پرت است. اما این روش در مجموع از اعتبار کافی برخوردار نیست. بر اساس تحقیقی که کالت و لویس (۱۹۷۶) بر داده‌های تک‌متغیره انجام دادند، دریافتند شناسایی نقاط پرت در این داده ها به عوامل زیر بستگی دارد:

۱) نحوه بررسی مقادیر نمونه برای شناسایی مقادیر پرت: اگر برای شناسایی نقاط پرت، داده‌های تصادفی را در سه حالت مرتب شده، مرتب نشده و نموداری بررسی کنیم، شانس تشخیص صحیح این مقادیر در داده‌های مرتب شده بسیار بیشتر است.

۲) مقیاس مقادیر نمونه: اگر مقیاس اندازه‌گیری مشاهدات را افزایش دهیم، احتمال اینکه مشاهدات بزرگ‌تر را، پرت در نظر بگیریم افزایش خواهد یافت.

بنابراین دلائل می‌توان گفت بررسی ظاهری مشاهدات به تنهایی روش مطمئنی برای شناسایی نقاط پرت نیست. علاوه بر این در نمونه‌های پیچیده‌تر با حجم‌های بزرگ‌تر (رگرسیون چند متغیره، طرح‌های چند عاملی و ...) این کار تقریباً غیر ممکن است. پس در این موارد بهتر است به جای بررسی ظاهری داده‌ها، از معیارهای منطقی‌تری در این زمینه بهره بگیریم. واضح است که با استفاده از این معیارها نمی‌توان مشاهدات پرت را به طور قطعی پیدا کرد، بلکه تنها مقادیر مشکوکی را می‌توان یافت که ارزش بیشتری دارد تا درباره پرت بودنشان تحقیق شود.

چون اکثر پدیده‌های طبیعی از توزیع نرمال پیروی می‌کنند، در این مقاله راه‌حلی را برای شناسایی نقاط پرت در داده‌های نرمال ارائه داده‌ایم. مشهورترین روش برای شناسایی نقاط پرت در داده‌های نرمال، استفاده از مقادیر Z مشاهدات است. بر اساس نمونه تصادفی $X_1,\ldots,X_n$ این مقادیر برابرند با:

$$Z_i = \frac{X_i - \bar{X}}{S},$$

که $\bar{X}$ و S به ترتیب میانگین و انحراف معیار نمونه‌اند. چون در توزیع گوسین (نرمال استاندارد)، ۹۹/۷ درصد مشاهدات بین ۳- و ۳ قرار می‌گیرند، مشاهداتی که $|Z_i| > ۳$ ، نقطه پرت فرض می‌شوند. اما این

روش به‌خصوص در نمونه‌های کوچک دقیق نیست. شفلر (۱۹۸۸) نشان داد که مقدار مطلق Z_i در نمونه‌ای به حجم n، حداکثر برابر $(n-۱)/\sqrt{n}$ است. بنابراین در نمونه‌ای به حجم ۱۰،

$$\forall i \quad ; \quad |Z_i| < ۲ / ۸۵ < ۳$$

در حالی‌که غیرمنطقی است هر نمونه ۱۰ تایی ممکن را از توزیع نرمال فرض کنیم.

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 پرت باشند نیز می‌توان برآورد دقیقی را از پارامتر جامعه به‌دست آورد. ایگلیویکس و هوگلین برای برآورد پارامتر مکان در جامعه از میانه مشاهدات $(\tilde{X})$ و برای برآورد پراکندگی در جامعه از برآوردگری که همپل (۱۹۷۴) به صورت زیر معرفی کرده است و به میانه انحراف از میانه شهرت دارد، استفاده کردند.

$$MAD_n = Med\{|X_i - \tilde{X}|\}.$$

و در نهایت مقادیر Z اصلاح شده مشاهدات را برابر زیر تعریف کردند:

$$M_i = \frac{X_i - \tilde{X}}{\frac{MAD_n}{d}},$$

که ثابت d در این عبارت سبب می‌شود، MAD_n/d برآوردگری ناریب برای σ باشد. چون بر اساس نمونه‌های با حجم زیاد از توزیع $N(\mu,\sigma^۲)$، $E(MAD_n) = ۰ / ۶۷۴۵\sigma$ است، ایگلیویکس و هوگلین 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 $|M_i| > ۳ / ۵$ صدق می‌کنند، نقطه پرت در نظر بگیرند.

در مثال زیر دو روشی را که تاکنون برای شناسایی نقاط پرت معرفی کرده‌ایم، مقایسه می‌کنیم.

*** مثال۱-۱-۱)** داده های فرضی زیر را در نظر بگیرید:

۲/۱, ۲/۶ , ۲/۴ , ۲/۵ , ۲/۳ , ۲/۱ , ۲/۳ , ۲/۶ , ۸/۲ , ۸/۳

میانگین و انحراف معیار داده‌ها برابرند با:

$$\bar{X} = ۳ / ۵۴ \quad , \quad S = ۲ \,/\, ۴۹$$

وجود مقادیر ۸/ ۲ و ۸/ ۳ در بین داده‌ها سبب کاهش دقت دو بر آوردگر میانگین و انحراف معیار شده است. مقادیر Z داده‌ها در جدول (۱-۱) درج شده است. مشاهده می‌کنیم که:

$$\forall i \quad ; \quad \left|z_i\right| < ۳$$

جدول۱-۱ : جدول مقادیر Z مشاهدات مثال۱-۱-۱

i	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
x_i	۲/۱	۲/۶	۲/۴	۲/۵	۲/۳	۲/۱	۲/۳	۲/۶	۸/۲	۸/۳
z_i	-۰/۵۸	-۰/۳۸	-۰/۴۶	-۰/۴۲	-۰/۵۰	-۰/۵۸	-۰/۵۰	-۰/۳۸	۱/۸۷	۱/۹۱

در حالی که دو مقدار ۸/ ۲ و ۸/ ۳ مقادیری کاملاً پرت در میان داده‌ها هستند، اما مقادیر z_i مشاهدات، این امر را تأیید نمی‌کنند. حال اگر دو مقدار $\tilde{X} = ۲ / ۴۵$ و $MAD_n = ۰ / ۱۵$ را برای برآورد میانگین و انحراف معیار جامعه در نظر بگیریم، مقادیر M_i مشاهدات برابر جدول(۱-۲) خواهد بود.

جدول۱-۲ : جدول مقادیر Z اصلاح شده‌ی مشاهدات مثال۱-۱-۱

i	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
x_i	۲/۱	۲/۶	۲/۴	۲/۵	۲/۳	۲/۱	۲/۳	۲/۶	۸/۲	۸/۳
m_i	-۱/۵۷	۰/۶۷	-۰/۲۲	۰/۲۲	-۰/۶۷	-۱/۵۷	-۰/۶۷	۰/۶۷	۲۵/۸۶	۲۶/۳۱

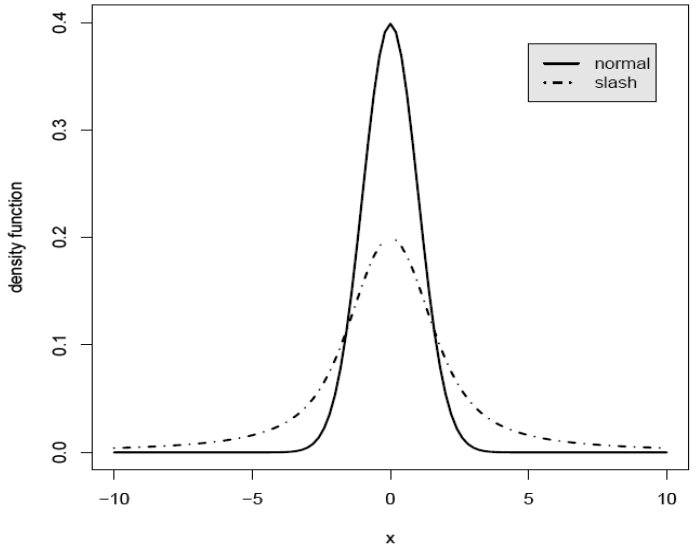
چون m_i مشاهداتِ ۸/ ۲ و ۸/ ۳ بیش از ۳/ ۵ است، بنابراین بر اساس این معیار می‌توان دو مقدار پرت را در میان مشاهدات پیدا کرد. این مثال به وضوح برتری مقادیر Z اصلاح‌شده را بر مقادیر Z معمولی در شناسایی نقاط پرت نشان می‌دهد.

در ادامه قصد داریم، ثابت d را که ایگلویکس و هوگلین برای نمونه‌های با حجم‌های متفاوت برابر مقدار ثابتِ ۶۷۴۵ / ۰ فرض کردند، برای حجم نمونه‌های تصادفی $n = ۵,۶,\ldots,۲۹,۳۰$ به صورت مجزا شبیه‌سازی کنیم. همچنین قصد داریم، بر خلاف ایگلیویکس و هوگلین که مقدار برش ثابتی را برای شناسایی نقاط پرت استفاده کردند، از مقادیر برش وابسته به حجم نمونه در این زمینه بهره بگیریم و بر اساس روشهای شبیه‌سازی این مقادیر را بیابیم.

قبل از پرداختن به این موضوع، توزیع اسلش را که در تولید نقاط پرت استفاده می‌شود، معرفی می‌کنیم.

۱-۲) معرفی توزیع اسلش

توزیع اسلش همانند توزیع گوسین نسبت به مبدأ متقارن است، اما دمهای سنگین آن سبب می‌شود تا با کمک آن بتوان مقادیر بسیار بزرگی را در نمونه‌ی تصادفی تولید کرد، به گونه‌ای که این مقادیر در مقایسه با مقادیر حاصل از توزیع گوسین به راحتی پرت به نظر رسند. در شکل (۱-۱) توابع چگالی توزیع‌های گوسین و اسلش رسم شده است.



شکل۱-۱ : نمودار تابع چگالی احتمال توزیع‌های نرمال استاندارد و اسلش

تعریف۱-۲-۱) اگر X و Y متغیرهای تصادفی مستقل و به ترتیب از توزیع‌های $N(۰,۱)$ و $U(۰,۱)$ برخوردار باشند، متغیر تصادفی

$$S = \frac{X}{Y}$$

دارای توزیع اسلش خواهد بود.

۱-۲-۱) تابع چگالی احتمال توزیع اسلش

ابتدا توزیع توأم متغیرهای تصادفی $S = \frac{X}{Y}$ و $W = Y$ را به‌دست می‌آوریم. ژاکوبین این تبدیل برابر است با:

$$J = \begin{vmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial w} \end{vmatrix} = \begin{vmatrix} w & s \\ ۰ & ۱ \end{vmatrix} = w \ .$$

بنابراین توزیع توأم متغیرهای تصادفی S و W برابر است با:

$$g_{S,W}(s,w) = f_{X,Y}(sw,w)\big|J\big| = \frac{w}{\sqrt{۲\pi}} e^{-\frac{s^۲w^۲}{۲}} \quad ; \quad ۰ \le w \le ۱ \,, s \in \mathbb{R}$$

و اما تابع چگالی احتمال متغیر تصادفی S برابر است با :

$$\begin{aligned} f_S(s) &= \int_۰^۱ g_{S,W}(s,w)dw = \frac{۱}{\sqrt{۲\pi}} \int_۰^۱ we^{-\frac{s^۲w^۲}{۲}} dw \\ &= \frac{۱}{s^۲}\left(\frac{۱}{\sqrt{۲\pi}} - \frac{۱}{\sqrt{۲\pi}} e^{-\frac{s^۲}{۲}} \right) \\ &= \frac{۱}{s^۲}\big(\phi(۰) - \phi(s) \big), \end{aligned}$$

که در آن ϕ تابع چگالی احتمال نرمال استاندارد است.

۱-۲-۲) تابع توزیع اسلش

ابتدا تابع توزیع شرطی S به شرط $Y = y$ را می‌یابیم:

$$\begin{aligned} \mathrm{P}(S \le s \mid Y = y) &= \mathrm{P}(\frac{X}{Y} \le s \mid Y = y) \\ &= \mathrm{P}(\frac{X}{y} \le s \mid Y = y) \\ &= \mathrm{P}(X \le sy \mid Y = y) \\ &= \mathrm{P}(X \le sy) = \Phi(sy), \end{aligned}$$

که در آن Φ تابع توزیع نرمال استاندارد است. بنابراین تابع توزیع اسلش برابر است با:

$$\begin{aligned} F_S(s) &= \int_۰^۱ \mathrm{P}(S \le s \mid Y = y) f_Y(y)dy = \int_۰^۱ \Phi(sy)dy \\ &= \frac{۱}{s} \int_۰^s \Phi(t)dt = \frac{۱}{s}\left(t\Phi(t)\Big|_۰^s - \int_۰^s t\phi(t)dt \right) \\ &= \frac{۱}{s}\left(s\Phi(s) - \frac{۱}{\sqrt{۲\pi}} e^{-\frac{t^۲}{۲}} \Big|_۰^s \right) = \Phi(s) - \frac{۱}{s}\big(\phi(۰) - \phi(s) \big). \end{aligned}$$

۱-۳) چگونگی یافتن مقادیر برش بر اساس حجم نمونه به‌دست آمده

در ابتدا تأکید می‌کنیم که همانند روش ایگلیویکس و هوگلین قصد داریم از مقادیر Z اصلاح شده‌ی مشاهدات، برای شناسایی نقاط پرت استفاده کنیم. مسأله در این جا یافتن مقادیر برش بهینه مبتنی بر حجم نمونه‌ی به‌دست آمده، است.

برای یافتن مقادیر برش بهینه، استراتژی کم‌بیشینه (مینیماکس) را پیشنهاد می‌کنیم. به عبارت دیگر ابتدا تابع زیانی را تعیین کرده و دو حالت را در نمونه‌ی تصادفی در نظر می‌گیریم. حالت اول زمانی که هیچ نقطه‌ی پرتی در مشاهدات وجود ندارد و حالت دوم زمانی که نمونه دارای نقاط پرت بسیاری است. مقدار برش بهینه حد وسطی است برای این دو حالت، به نحوی که ماکسیمم زیان ما در ازاء مقادیر برش مختلف کمترین گردد. ابتدا معیار زیر را به عنوان تابع زیان مسأله تصمیم تعریف می‌کنیم.

تعریف۱-۳-۱) درجه‌ی بدی: برای نمونه‌ی تصادفی $X_1,\ldots,X_n$ از توزیعی با میانگین μ، درجه‌ی بدی مقدار برش λ برابر است با:

$$B(\lambda) = (\overline{X}_\lambda - \mu)^2,$$

که $\overline{X}_\lambda$ میانگین مشاهدات باقیمانده در نمونه است، پس از حذف مقادیری که $|M_i| > \lambda$.

در تعریف درجه‌ی بدی فرض بر این است که هدف از جمع‌آوری داده‌ها، برآورد میانگین جامعه است. به نظر می‌رسد میانگین نمونه پس از حذف مشاهدات پرت، برآوردگری منطقی برای این منظور باشد.

بنابراین برای هر توزیعی همانند F، تعریف می‌کنیم:

$$b(\lambda) = E_F\big(B(\lambda)\big),$$

که امید ریاضی بر اساس نمونه‌های تصادفی به حجم n از این توزیع به‌دست آمده است.

فرض کنید λ_m مقداری است که در ازاء آن $b(\lambda)$ می‌نیمم می‌شود، یعنی $b(\lambda_m) = \min_\lambda b(\lambda)$. به علاوه فرض می‌کنیم $b_m = b(\lambda_m)$. بنابراین درجه‌ی بدی مقیاس‌بندی‌شده را به صورت زیر تعریف می‌کنیم:

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

بنابر دلائل زیر از درجه‌ی بدی مقیاس‌بندی‌شده برای یافتن مقدار برش بهینه استفاده می‌کنیم:

۱) بهترین مقدار برش برای مشاهدات نمونه زمانی است که $b(\lambda)$ کمترین مقدار خود یعنی b_m را اختیار کند. از طرفی $b(\lambda)$ معیاری است که میزان ناکارائی میانگین نمونه را – پس از حذف مقادیر پرت بر اساس مقدار برش λ – برای برآورد میانگین جامعه نشان می‌دهد. بنابراین $b_{SC}(\lambda)$ ناکارائی نسبی برآوردگر میانگین را در مقایسه با بهترین حالت ممکن آن نشان می‌دهد.

۲) به جهت این که بتوانیم به ازای λ های معین، درجات بدی توزیع‌های مختلف را با یکدیگر مقایسه کنیم، استفاده از $b_{SC}(\lambda)$ را ترجیح می‌دهیم.

حال برای یافتن مقدار برش بهینه، از دو توزیع کاملاً متفاوت در تولید نقاط پرت استفاده می‌کنیم: توزیع گوسین (نرمال استاندارد) و توزیع اسلش که نقاط پرت بسیاری را - در مقایسه با توزیع گوسین - تولید می‌کند.

بنابراین به روش کم‌بیشینه، مسأله‌ی تصمیمی را حل می‌کنیم که تابع زیان آن مطابق با درجه‌ی بدی مقیاس‌بندی‌شده تعریف می‌شود و بر این اساس مقدار برش بهینه - که آن را با λ^* نشان می‌دهیم- مقداری است که در رابطه‌ی زیر صدق می‌کند:

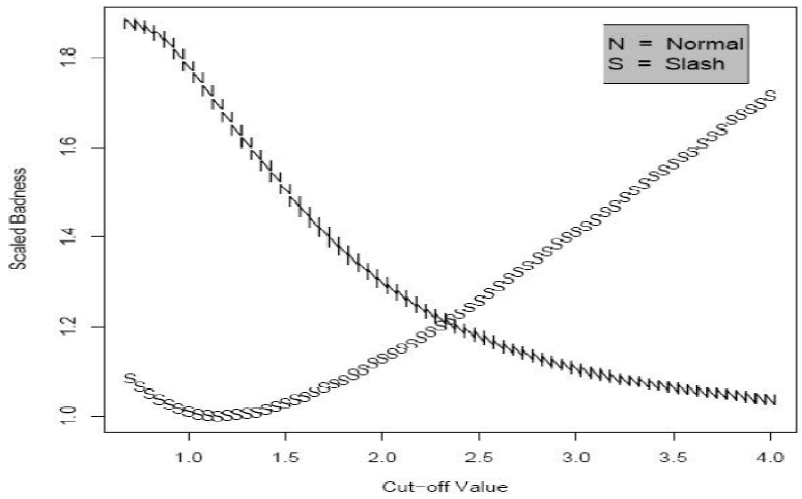
$$\max\Big(b_{SC}^{G}(\lambda^*), b_{SC}^{S}(\lambda^*)\Big) = \min_{\lambda} \max\Big(b_{SC}^{G}(\lambda), b_{SC}^{S}(\lambda)\Big),$$

که $b_{SC}^{G}(\cdot)$ و $b_{SC}^{S}(\cdot)$ به ترتیب درجات بدی مقیاس‌بندی‌شده‌ی توزیع‌های گوسین و اسلش را نشان می‌دهد. به عبارت دیگر به ازاء مقادیر برش مختلف، مقداری را انتخاب می‌کنیم که ماکزیمم درجه بدی مقیاس‌بندی‌شده‌ی آن در بین دو توزیع گوسین و اسلش کمترین باشد.

۱-۴) نحوه شبیه‌سازی

زمانی که نمونه تصادفی دقیقاً بر اساس توزیع گوسین به‌دست آمده است، برای این‌که میانگین نمونه – پس از حذف مقادیر پرت – برآورد دقیقی را از پارامتر μ حاصل کند، منطقی است که مقدار برش λ را بزرگترین مقدار ممکن در نظر بگیریم تا تمامی مشاهدات نمونه در برآورد پارامتر μ شرکت داشته باشند. در این صورت هر چه λ بزرگ‌تر باشد، درجه‌ی بدی $b^G(\lambda)$ کاهش خواهد یافت. (همین طور $b_{SC}^{G}(\lambda)$) بنابراین $b_{SC}^{G}(\lambda)$ تابعی نزولی بر حسب λ است. اما اگر مشاهدات نمونه دارای نقاط پرت باشند (که در اینجا برای تولید نمونه‌های تصادفی دارای نقاط پرت بسیار از توزیع اسلش استفاده کرده‌ایم) مقادیر بزرگ λ باعث می‌شود، مشاهدات پرت همچنان در نمونه باقی مانند و دقت برآورد پارامتر μ را کاهش دهند. در این

صورت با افزایش λ، $b^S(\lambda)$ افزایش خواهد یافت. (همین طور $b^S_{SC}(\lambda)$) بنابراین $b^S_{SC}(\lambda)$ تابعی صعودی بر حسب λ است. در شکل (۱-۲) مبتنی بر شبیه‌سازی برای حجم نمونه‌ی برابر ۵، مقادیر برآوردشده $b^G_{SC}(\lambda)$ و $b^S_{SC}(\lambda)$ را در ازای λ های مختلف رسم کرده‌ایم. (نحوه رسم این شکل را در بخش‌های بعد توضیح می-دهیم.) در این شکل صعودی بودن b^S_{SC} و نزولی بودن b^G_{SC} (به استثناء همسایگی عدد یک) کاملاً مشهود است.



شـــــکل ۱-۲ : نمـــــودار مقـــــادیر شـــــبیه‌ســـــازی‌شـــــده $b^G_{SC}(\lambda)$ و $b^S_{SC}(\lambda)$ در ازای $\underline{\lambda} = \{۰ / ۷,۰ / ۷۵,...,۳ / ۹۵,۴\}$ برای $n = ۵$

بر اساس شکل (۱-۲) محل تلاقی توابع $b^G_{SC}(\lambda)$ و $b^S_{SC}(\lambda)$ بر روی محور افقی، مقدار برش بهینه λ^* را مشخص می‌کند. حال کافی است نحوه به‌دست آوردن λ^* را شرح دهیم.

رابطه‌ی زیر را تعریف می‌کنیم:

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

مقدار λ^* از حل معادله‌ی $\delta(\lambda^*) = ۰$ حاصل می‌شود. اما چون $b^G_{SC}(\lambda)$ و $b^S_{SC}(\lambda)$ مقادیری نامعلوم‌اند، به روش شبیه‌سازی و به ازای λ های معینِ $\{\lambda_۱,\lambda_۲...,\lambda_k\}$ آنها را برآورد می‌کنیم. قبل از بیان نحوه شبیه‌سازی b^G_{SC} و b^S_{SC} تأکید می‌کنیم که مقدار دقیق λ^* به حجم نمونه‌های تصادفی (n) وابسته است.

ابتدا $N_۱$ نمونه تصادفی به حجم n از توزیع گوسین و $N_۲$ نمونه تصادفی به حجم n توزیع اسلش به‌دست می‌آوریم. به منظور کاهش واریانس تفاوت بین متوسط درجات بدی در نمونه‌های گوسین و اسلش، بهتر است ابتدا نمونه‌های اسلش را تولید کرده و سپس با استفاده از رابطه‌ی زیر نمونه‌های گوسین را به‌دست آوریم.