داده های پرت، مفهوم و کاربرد آن

مهدی جباری نوقابی[۱]

چکیده

در این مقاله ابتدا داده های پرت را تعریف و با ارائه‌ی مثال هایی مفهوم داده پرت یا دور افتاده را شرح می دهیم و در آخر مدل هایی را معرفی می کنیم که با استفاده از این مدل ها می توان پارامتر های توزیع را وقتی داده پرت در سری مشاهدات وجود دارد برآورد کرد.

واژه های کلیدی: داده پرت، طرح آزمایش ها، رگرسیون.

۱– مقدمه

اغلب دانشجویان و محققین رشته آمار و سایر رشته ها وقتی در مواجهه با تحقیقاتی که جنبه کاربردی دارند و به نوعی برای بررسی اهداف آن ها نیاز به تجزیه و تحلیل اطلاعات و داده ها مبتنی بر روش های آمار توصیفی یا استنباطی می باشد، قرار می گیرند، به خصوص در مباحث مرتبط با رگرسیون یا طرح آزمایش ها، با مفهومی روبرو می شوند که به آن داده دور افتاده یا پرت[۲] می-گویند، با توجه به دیدگاه محقق، اغلب تعاریف متفاوتی از آن مطرح و تصمیم های مختلفی برای این گونه مشاهدات در نظر گرفته می شود. لذا در این پژوهش، سعی شده است تا نکاتی را در ارتباط با موضوع داده های پرت و کاربرد آن و همچنین چگونگی مواجهه با آن در مسائل مختلف، بیان گردد. قبل از این که به تعریف و مفهوم داده های پرت بپردازیم، با چند مثال که در درک مفهوم داده های پرت به ما کمک می کند، آشنا می شویم.

در یک کلاس درس، ممکن است چند دانشجو (عده ای کم در مقایسه با عمده دانشجویان) در هنگام تدریس استاد، با یک دیگر صحبت نمایند و باعث اختلال در کلاس درس شوند. در نظر بگیرید یک فروشنده برنج، برای سود بیش تر می خواهد مقدار کمی از برنج های با کیفیت پایین اش را با مقدار زیادی از برنج های با کیفیت، مخلوط نماید و سپس آن ها را با قیمت برنج های با کیفیت بفروشد. وقتی در خیابان های شهر در حال تردد هستید عمده ماشین هایی که می بینید، مانند هم دیگر و با قیمت تقریبا یکسانی هستند. اما تعدادی خودروی گران قیمت نیز مشاهده می کنید که قیمت آن ها به مراتب بیش تر از بقیه می باشد. وقتی استاد درس در پایان ترم می خواهد از دانشجویان ارزیابی به عمل آورد، پس از تصحیح اوراق امتحانی ممکن است تعداد کمی از دانشجویان نمرات خیلی کم و تعدادی نیز نمرات بالا داشته باشند.

[۱] استادیار گروه آمار، دانشکده علوم ریاضی، دانشگاه فردوسی مشهد

[2] Outlier

در همه مثال های فوق، داده هایی غیر عادی در بین مشاهدات، وجود دارد. بعضی از این مشاهدات ممکن است جزء داده های دور افتاده محسوب شوند.

اغلب اوقات وقتی محققین با این گونه مشاهدات مواجه شوند، یا به ماهیت آن داده ها توجهی ننموده و همان روش های آماری ای را در پیش می گیرند که برای داده هایی که شامل مشاهده پرت نیست، انجام می دهند و یا گاهی اوقات بدون در نظر گرفتن عواقب کار، داده هایی که فکر می کنند، داده دور افتاده هستند، را حذف نموده و سپس داده ها را تجزیه و تحلیل می کنند.

در بعضی مسایل داده هایی را داریم که تقریبا همگن و یک دست بوده و همانند داده های بدون مشاهده دور افتاده هستند. بنابراین فکر می کنیم که داده ها دارای مشاهدات پرت نیست. اما وقتی به ماهیت اصلی موضوع تحقیق، مراجعه می کنیم، مشاهده می شود که ممکن است تعدادی داده پرت در بین داده ها وجود داشته باشند، ولی بدون داشتن اطلاعاتی در مورد موضوع تحقیق، صرفا با مشاهده داده ها نمی توان، مشاهدات دور افتاده را تشخیص داد. برای روشن شدن موضوع، فرض کنید در یک کلاس درس که برای دانشجویان دوره کارشناسی آمار، ارائه شده است، چند دانشجوی دوره کارشناسی ارشد نیز به دلیل انتخاب درس بعنوان درس پیش نیاز، ثبت نام کرده باشند. در پایان ترم پس از امتحان درس مربوطه، تعدادی عدد در فاصله صفر تا بیست را خواهیم داشت که در واقع نمرات دانشجویان می باشد. ممکن است نمرات اخذ شده توسط این تعداد دانشجوی دوره کارشناسی ارشد، تفاوت چندانی با بقیه نمرات دانشجویان نداشته باشد، لذا نمی توان با فقط داشتن نمرات، تشخیص داد که کدام نمرات مربوط به این چند دانشجوی خاص است. اما در واقع می توان نمره این چند دانشجو را جزء داده های پرت در نظر گرفت.

۲- تعریف داده های پرت

با توجه به مثال ها و توضیحات ارائه شده، مفهوم داده های پرت تقریبا مشخص گردید. دانشمندان زیادی در مواجهه با مسائل مختلف، تعاریفی از داده های پرت را ارائه کرده اند که در این جا قبل از تعریف جامع داده های پرت، به بررسی تعدادی از آن ها می پردازیم.

گروبس [۱۶] معتقد است که لزوماً همیشه تمامی داده ها به طور واضحی با یکی از مشاهدات سازگار نیستند. بدین معنی که ممکن است که در مطالعه داده ها، به وضوح ناهمگنی دیده شود و گروهی از مشاهدات به صورت کلی یک نمونه تصادفی از یک توزیع مثلا نرمال، تولید نشده باشند. در این مورد، یک داده یا تعداد بیشتری ممکن است داده پرت باشند. کندال و بوکلند[۲۶] تعریف کردند که در یک نمونه از n مشاهده، تعداد معدودی از داده ها ممکن است به لحاظ مقدار، از بقیه مشاهدات دور باشند. مسأله مزبور، این سؤال را پیش می آورد که فکر کنیم آیا این مشاهدات از جامعه دیگری نیستند یا این که روش نمونه گیری اشتباه است؟ این مقادیر، داده پرت هستند.

انسکومب[۱] می گوید، یک مشاهده با باقیمانده بزرگ و غیر طبیعی، یک مشاهده پرت را تلقی می کند. فرگوسن[۱۵] اعتقاد دارد که در یک نمونه با اندازه اصلاح شده از یک جامعه، معمولا یک یا دو مشاهده به طور شگفت آوری از گروه اصلی فاصله دارند، که به آن ها داده پرت گوییم. گروبس [۱۷] تعریف می کند که یک مشاهده دور افتاده، آن داده ای است که به طور علامت داری مشخص است که انحرافی از بقیه داده های نمونه، تحت شرایطی که اتفاق می افتد، دارد. هاوکینز[۱۸] معتقد بود، مشاهده ای که آن قدر از بقیه داده ها انحراف داشته باشد به طوری که این ایده را در ذهن بیاورد که توسط روش و مکانیزم دیگری تولید شده است، داده دور افتاده است. میلر [۲۸] ابراز نمود که داده پرت یک مشاهده یا میانگین ساده است که از روند بقیه داده ها پیروی نمی کند. بکمن و کوک [۳] اعتقاد داشتند

که مشاهده ناسازگار هر داده ای است که در هنگام بررسی مشاهدات، شگفت آور بوده یا مورد اختلاف باشد. همچنین، آنان یک مشاهده برجسته را داده ای که توسط توزیع هدف قابل بررسی می باشد، تعریف کردند. بارنت و لویس [۲] داده پرت را مشاهده ای می دانند که به طور آشکاری نسبت به باقیمانده مجموعه داده ها ناسازگار می باشد. بررسی این تعریف نشان می دهد که تصویر مشترک برای همه مشاهدات طوری است که یک مشاهده پرت توسط میانگین رابطه اش با بقیه مشاهدات، مشخص می شود. براساس مثال های فوق و تعریف های مورد بحث، می توان گفت که این تعریف تقریباً مناسب می باشد. زیرا بکمن و کوک [۳] بیان نمودند که قابلیت اعتماد احتمالی هر مشاهده توسط رابطه اش با دیگر مشاهدات که تحت شرایط یکسانی بدست می آید، مشخص می گردد.

تعداد دیگری از محققین، معتقدند که اگر در یک مجموعه مشاهدات، اختلاف داده ای نسبت به میانگین مشاهدات از پنج برابر انحراف معیار، بیش تر باشد، آن مشاهده یک داده پرت خواهد بود. براساس این تعریف، اگر سرمایه آقای بیل گیتس در مقایسه با سرمایه نه نفر دیگر که به تصادف از جامعه انتخاب شده اند، مقایسه شود برای هر ترکیب تصادفی از افراد، سرمایه ایشان یک مشاهده پرت خواهد بود.

بعضی دیگر از دانشمندان، سعی نموده اند داده های پرت را مشخصاً به صورت زیر تعریف نمایند. فرض کنید که q_1 و q_3 به ترتیب چارک اول و سوم یک نمونه تصادفی باشد، آن گاه دامنه میان چارکی که آن را با IQR نشان می دهیم برابر است با $q_3 - q_1$. لذا اگر یک مشاهده حداقل به اندازه سه برابر IRQ از فاصله (q_1، q_3) دور باشد، آن مشاهده، یک داده پرت خواهد بود.

با توجه به توضیحات فوق و براساس تعریف های مطرح شده، می توانیم داده های پرت را به صورت زیر تعریف کنیم.

تعریف: یک داده پرت، مشاهده ای است که به طور غیرعادی یا اتفاقی از وضعیت عمومی داده های تحت آزمایش و نسبت به قاعده ای که براساس آن آنالیز می شود، انحراف داشته باشد.

۳– مثال هایی از داده های پرت

برای مشخص شدن و درک بیش تر داده های پرت، مثال هایی را در نظر بگیرید که در آن $n-k$ مشاهده از توزیع $F(x)$ تولید شده و k مشاهده باقیمانده از توزیع $G(x)$ می آیند. همچنین هر دو توزیع وابسته به پارامتر $\theta \in \Theta$ می باشند.

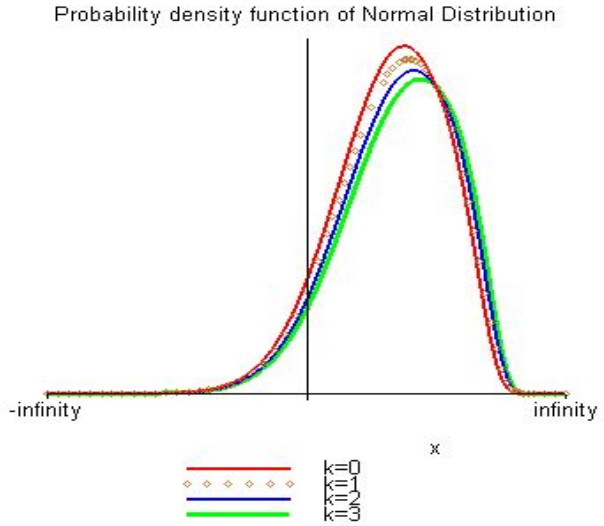
مثال ۱: فرض کنید مشاهدات $X_1, X_2, ..., X_n$ طوری به تصادف انتخاب شده باشند که k تا از آن ها از تابع توزیع $G(x)$ و $n-k$ تای باقی مانده از توزیع $F(x)$ پیروی کند. در نظر می گیریم $n=10$ و $k=1,2,3$. آنگاه تابع چگالی احتمال داده ها در مواجهه با داده های پرت در هر کدام از حالات زیر را رسم می کنیم.

الف) ۹ مشاهده به تصادف از توزیع نرمال با میانگین ۳ و واریانس ٤ (N(۳,٤)) و یک مشاهده از یکی از توزیع های N(٤,٥) یا N(۳,٦) یا N(٥,٦) تولید شده باشند.

ب) ۸ مشاهده به تصادف از توزیع نرمال با میانگین ۳ و واریانس ٤ (N(۳,٤)) و دو مشاهده از یکی از توزیع های N(٤,٥) یا N(۳,٦) یا N(٥,٦) تولید شده باشند.

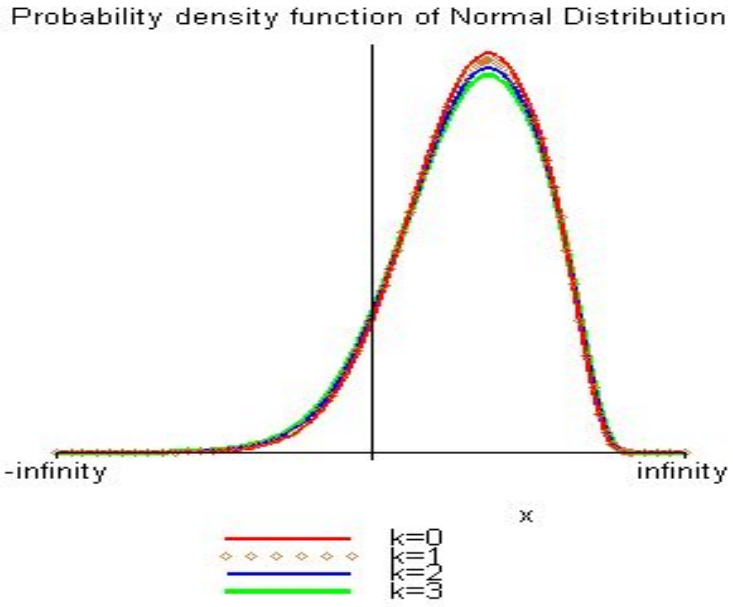
ج) ۷ مشاهده به تصادف از توزیع نرمال با میانگین ۳ و واریانس ٤ (N(۳,٤)) و سه مشاهده از یکی از توزیع های N(٤,٥) یا N(۳,٦) یا N(٥,٦) تولید شده باشند.

حل:

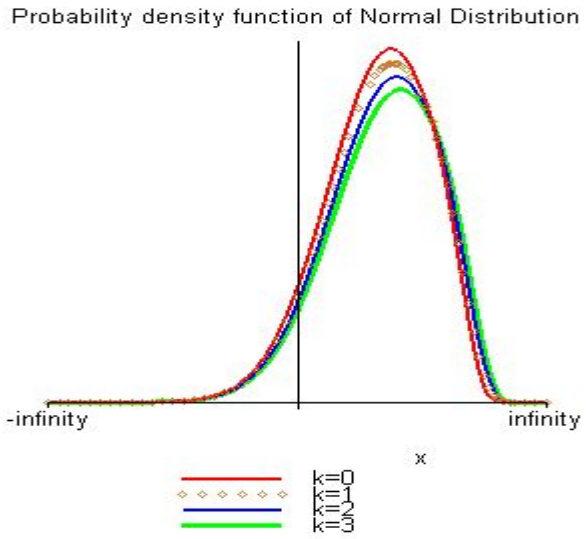


نمودار شماره ۱: تابع چگالی $N(۳,۴)$ ($F(x)$) در مواجهه با $N(۴,۵)$ ($G(x)$)

(نماد infinity همان ∞ می باشد)



نمودار شماره ۲: تابع چگالی $N(۳,۴)$ در مواجهه با $N(۳,۶)$



نمودار شماره ۳: تابع چگالی N(۳,۴) در مواجهه با N(۵,۶)

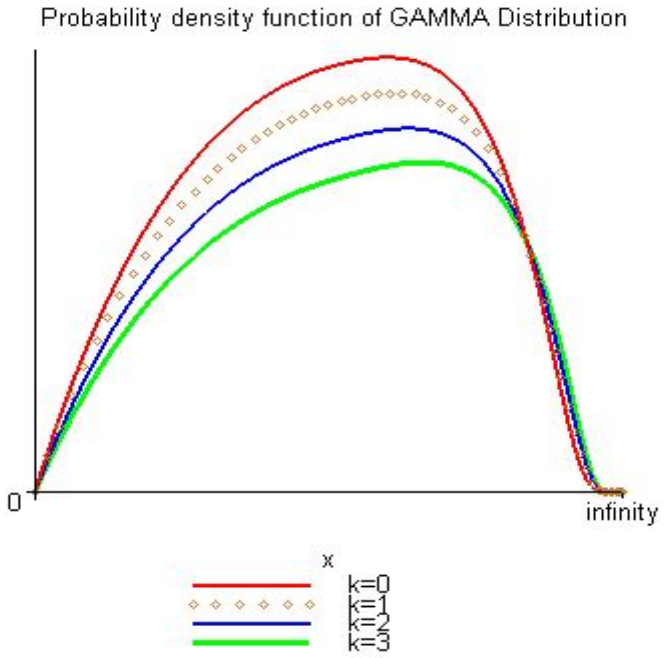
مثال ۲: برای مشاهدات $X_1, X_2, ..., X_n$ همانند مثال قبل، در هر کدام از حالات زیر تابع چگالی در مواجهه با مشاهدات پرت را رسم می کنیم.

الف) ۹ مشاهده به تصادف از توزیع گاما با پارامترهای ۲ و ۵ ((Gamma(۵،۲) و یک مشاهده از یکی از توزیع های Gamma(۵،٤) یا Gamma(۱۵،۲) یا Gamma(۱۵،٤) تولید شده باشند.

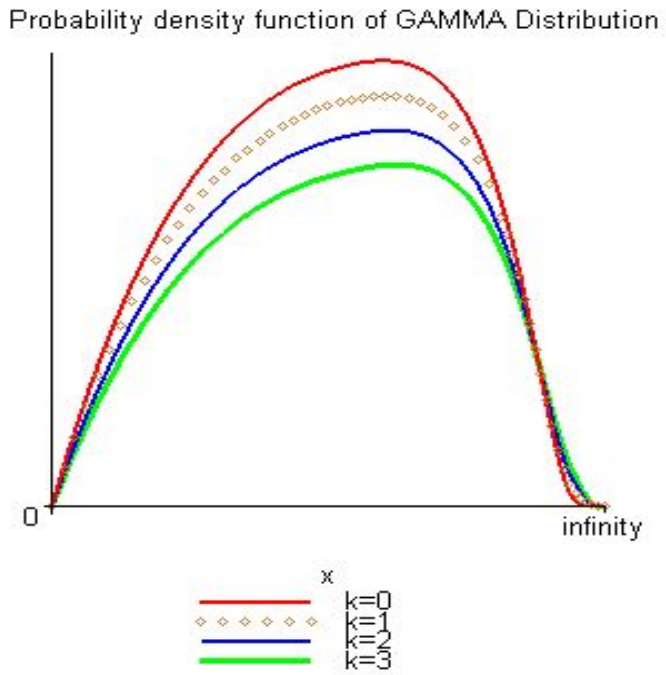
ب) ۸ مشاهده به تصادف از توزیع Gamma(۵،۲) و دو مشاهده از یکی از توزیع های Gamma(۵،٤) یا Gamma(۱۵،۲) یا Gamma(۱۵،٤) تولید شده باشند.

ج) ۷ مشاهده به تصادف از توزیع Gamma(۵،۲) و سه مشاهده از یکی از توزیع های Gamma(۵،٤) یا Gamma(۱۵،۲) یا Gamma(۱۵،٤) تولید شده باشند.

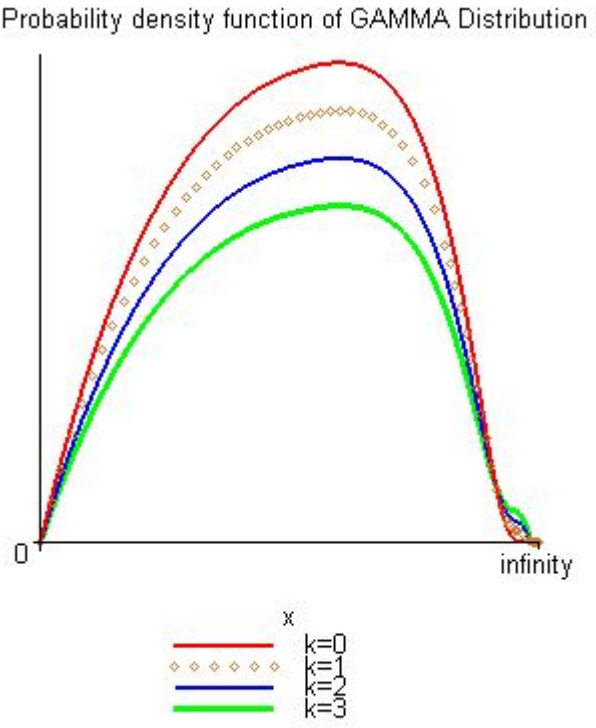
حل:



نمودار شماره ۴: تابع چگالی Gamma(۵،۲) در مواجهه با Gamma(۵،۴)



نمودار شماره ۵: تابع چگالی Gamma(۵،۲) در مواجهه با Gamma(۱۵،۲)



نمودار شماره ۶: تابع چگالی Gamma(۵،۲) در مواجهه با Gamma(۱۵،۴)

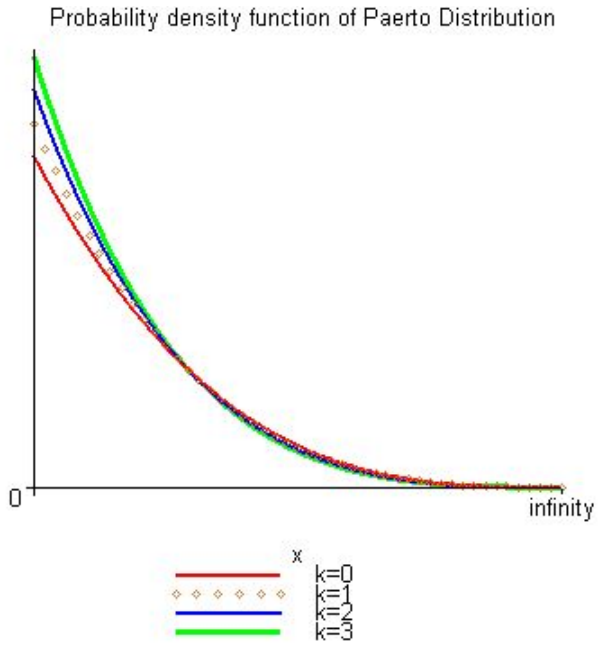
مثال ۳: همانند مثال قبل، در هر کدام از حالات زیر تابع چگالی داده ها در مواجهه با مشاهدات پرت را رسم می کنیم.

الف) ۹ مشاهده به تصادف از توزیع پارتو با پارامترهای ۲ و ۵۰ ((Pareto(۵۰،۲) و یک مشاهده از یکی از توزیع های Pareto(۵۰،٤) یا Pareto(۷۵،۲) یا Pareto(۷۵،٤) تولید شده باشند.

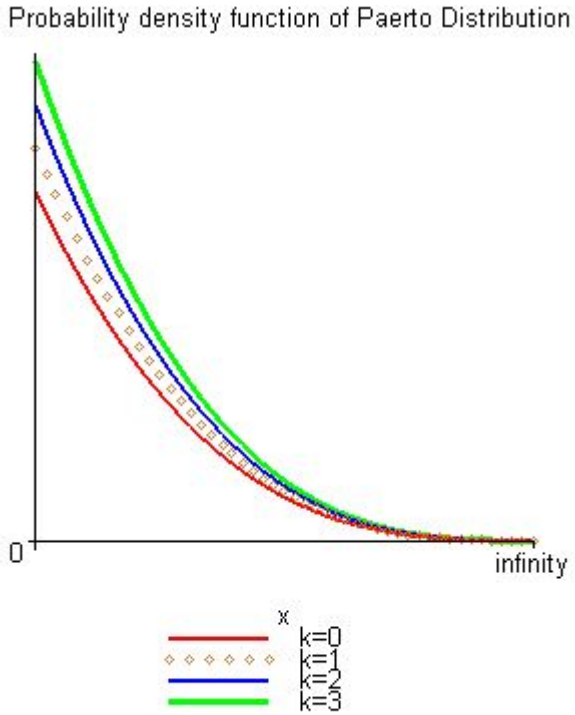
ب) ۸ مشاهده به تصادف از توزیع Pareto(۵۰،۲) و دو مشاهده از یکی از توزیع های Pareto(۵۰،٤) یا Pareto(۷۵،۲) یا Pareto(۷۵،٤) تولید شده باشند.

ج) ۷ مشاهده به تصادف از توزیع Pareto(۵۰،۲) و سه مشاهده از یکی از توزیع های Pareto(۵۰،٤) یا Pareto(۷۵،۲) یا Pareto(۷۵،٤) تولید شده باشند.

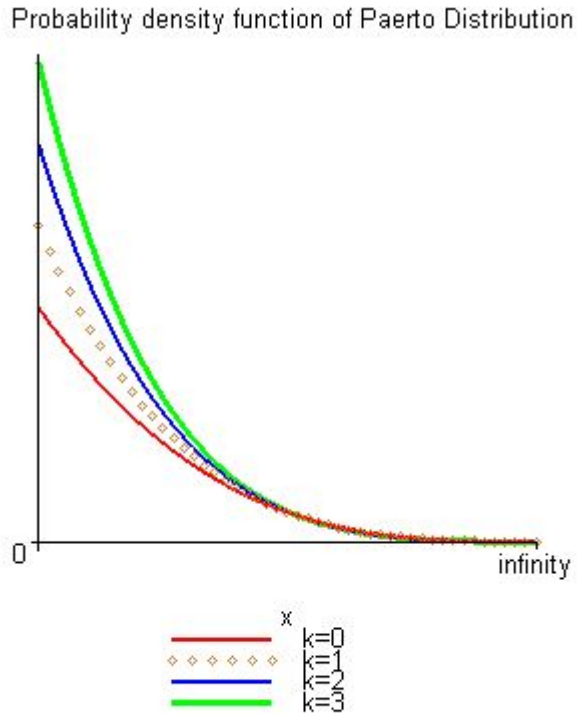
حل:



نمودار شماره ۷: تابع چگالی Pareto(۵۰،۲) در مواجهه با Pareto(۵۰،۴)



نمودار شماره ۸: تابع چگالی Pareto(۵۰،۲) در مواجهه با Pareto(۷۵،۲)



نمودار شماره ۹: تابع چگالی Pareto(۵۰،۲) در مواجهه با Pareto(۷۵،۴)

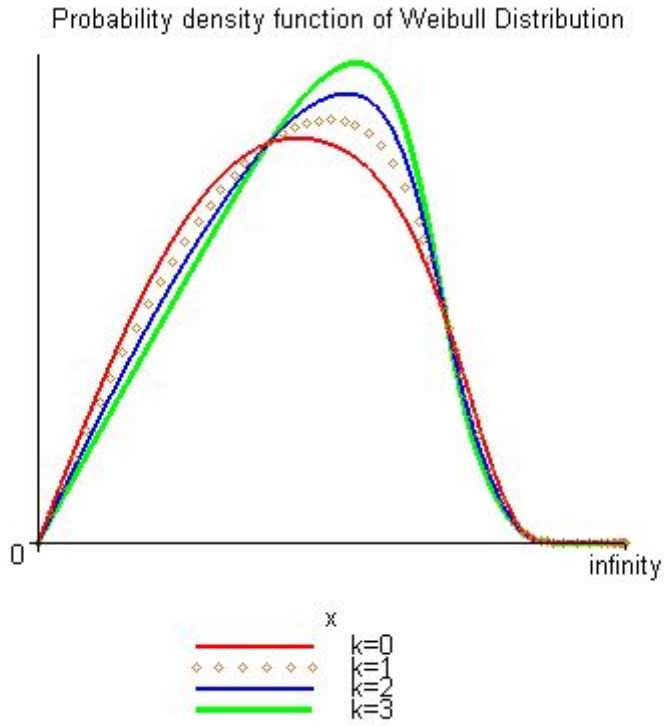
مثال ٤: همانند مثال قبل، در هر کدام از حالات زیر تابع چگالی داده ها در مواجهه با مشاهدات پرت را رسم می کنیم.

الف) ۹ مشاهده به تصادف از توزیع وایبل با پارامترهای ۲ و ۵ (Weibull(۵،۲)) و یک مشاهده از یکی از توزیع های Weibull(۵،٤) یا Weibull(۱،۲/۵) یا Weibull(۱،٤/۵) تولید شده باشند.

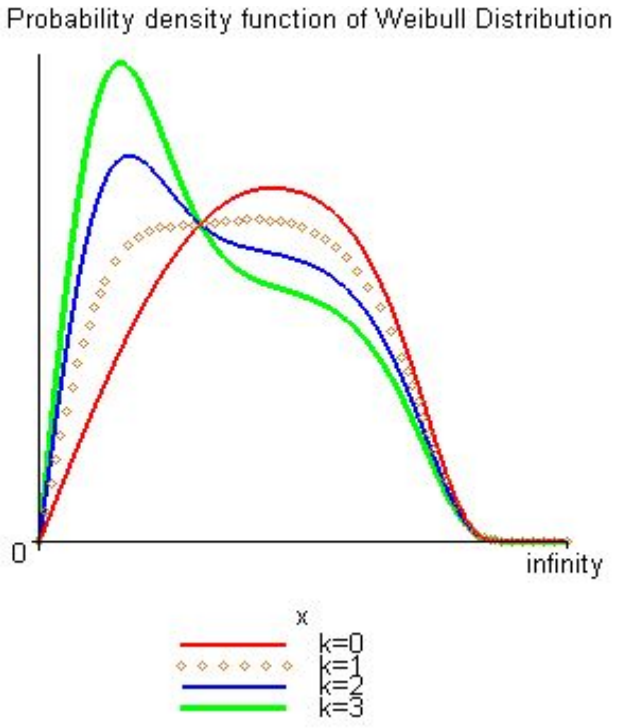
ب) ۸ مشاهده به تصادف از توزیع Weibull(۵،۲) و دو مشاهده از یکی از توزیع های Weibull(۵،٤) یا Weibull(۱،۲/۵) یا Weibull(۱،٤/۵) تولید شده باشند.

ج) ۷ مشاهده به تصادف از توزیع Weibull((۵،۲)) و سه مشاهده از یکی از توزیع های Weibull(۵،٤) یا Weibull(۱،۲/۵) یا Weibull(۱،٤/۵) تولید شده باشند.

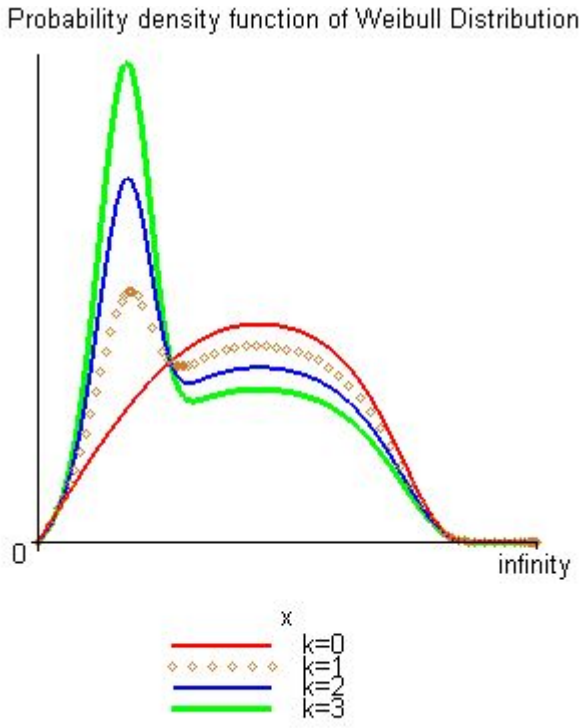
حل:



نمودار شماره ۱۰: تابع چگالی Weibull(۵،۲) در مواجهه با Weibull(۵،۴)



نمودار شماره ۱۱: تابع چگالی Weibull(۵،۲) در مواجهه با Weibull(۱،۲/۵)



نمودار شماره ۱۲: تابع چگالی Weibull(۵،۲) در مواجهه با Weibull(۱،۴/۵)

۴- تشخیص داده های پرت

در بسیاری از داده ها، می توان مشاهده نمود که تعدادی از آن ها پراکندگی زیادتری نسبت به بقیه دارند، به طوری که نشان دهنده منابع غیر طبیعی خطا بوده و این خطاها در بحث های نظری، در نظر گرفته نشده اند. برای اطلاعات بیش تر به بارنت و لویس [۲] مراجعه کنید.

اگر توسط روش های آزمون فرضیه بتوانیم تشخیص دهیم که تعدادی از مشاهدات، داده پرت هستند، آنگاه هر چند ممکن است که مطمئن نباشیم که این روش تشخیص کاملاً درست است، اما می تواند به ما در تجزیه و تحلیل داده ها کمک فراوانی کند. بنابراین در مسأله مواجهه با داده های پرت یکی از مشکلات، روش تشخیص آن ها می باشد. البته با توجه به تعریف جامع داده های پرت، عملاً تشخیص آن ها با روش های آزمون فرضیه امکان پذیر نمی باشد. زیرا که داده های پرت ممکن است از توزیعی همانند توزیع بقیه مشاهدات ولی با یک یا چند پارامتر متفاوت، آمده باشند، یا این که از توزیع متفاوتی پیروی کنند. لذا، در صورتی که بدانیم یا بپذیریم داده ها هم توزیع می باشند، می توان با استفاده از تکنیک های مختلفی، داده های پرت را تشخیص داد. محققین فراوانی تاکنون در مورد روش های تشخیص داده های پرت در توزیع های مختلف نرمال، نمایی، گاما، پارتو، وایبل و غیره مقالاتی را نوشته اند. برای اطلاعات بیش تر می توان به بولشف [٤]، چیکاگودار و کونچور[۵]، کابه[٢٤]، لاکس[۲۷]، زربت و نیکولین[۳۳]، جباری نوقابی و همکاران [۲۱]، دیکشیت و جباری نوقابی [١٤] و جباری نوقابی و جباری نوقابی [۲۲] مراجعه نمود.

۵- برآورد در مواجهه با داده های پرت

کاله و سینها [۲۵]، چیکاگودار و کونچور [۵]، جوشی[۲۳]، راهوت[۳۰]، سینها [۳۱]، ویال [۳۲]، کولت و لویس [٦]، جباری نوقابی و همکاران [۲۰]، جباری نوقابی و جباری نوقابی [۱۹]، دیکشیت [۸]، دیکشیت و جباری نوقابی [۹و۱۰و۱۱و۱۲و۱۳و۱٤] و نصیری و

جباری نوقابی [۲۹] پارامترهای توزیع نرمال را وقتی یک داده پرت در بین آن ها وجود داشته باشد برآورد نمودند. حال اگر بخواهیم به طور کلی مدلی بیان کنیم، تا براساس آن توزیع و چگالی یک متغیر تصادفی را در مواجهه با k داده پرت بدست آورد، می توان از مدل زیر که به مدل دیکشیت [۷] معروف می باشد، استفاده نمود.

۵-۱- مدل دیکشیت

فرض کنید، X_i $(i \geq ۱)$ یک سری از متغیرهای تصادفی نامنفی باشد که برای هر ترکیب $(i_۱, i_۲, ..., i_{n-k})$ از مقادیر صحیح $(۱,۲,...,n)$، شرایط زیر برقرار باشد.

۱. متغیرهای تصادفی $X_{i_۱}, X_{i_۲}, ..., X_{i_{n-k}}$ مستقل و دارای تابع چگالی (جرم) احتمال $f(x)$ باشد.

۲. متغیرهای تصادفی باقی مانده نیز مستقل و دارای تابع چگالی (جرم) احتمال $g(x)$ باشد.

۳. دو مجموعه متغیرهای تصادفی، مستقل باشند.

٤. به علاوه فرض کنیم که ترکیب های $(i_۱, i_۲, ..., i_{n-k})$ از اعداد صحیح $(۱,۲,...,n)$ به طور تصادفی و با احتمال های برابر $[C(n,k)]^{-۱}$ برای هر ترکیب، به طوری که $C(n,k) = \frac{n!}{k!(n-k)!}$، انتخاب شده باشند.

آنگاه تابع چگالی توأم $(X_۱, X_۲, ..., X_n)$ به صورت زیر است.

$$f(x_۱, x_۲, ..., x_n) = \prod_{i=۱}^{n} f(x_i) \sum_{(i_۱, i_۲, ..., i_k)} \prod_{j=۱}^{k} \frac{g(x_{i_j})}{f(x_{i_j})} [C(n,k)]^{-۱}, \quad (۱)$$

که در آن $\sum_{(i_۱, i_۲, ..., i_k)} = \sum_{i_۱=۱}^{n-k+۱} \sum_{i_۲=i_۱+۱}^{n-k+۲} \cdots \sum_{i_k=i_{k-۱}+۱}^{n}$.

همچنین، برای یک تک مشاهده X، تابع چگالی و توزیع متغیر تصادفی به شکل زیر می باشد.

$$h(x) = bg(x) + \bar{b}f(x), \quad (۲)$$

و

$$H(x) = bG(x) + \bar{b}F(x). \quad (۳)$$

که در آن $b = \frac{k}{n}$ و $\bar{b} = ۱ - b$.

با توجه به مدل فوق تابع چگالی توأم دو متغیر تصادفی $(X_i, X_j), i \neq j$ و همچنین سه متغیر تصادفی $(X_i, X_j, X_t), i \neq j \neq t$ به ترتیب عبارتند از

$$\begin{aligned} h(x_i, x_j) = \frac{۱}{\binom{n}{k}} \Big\{ & \binom{n-۲}{k-۲} [g(x_i)g(x_j)] \\ & + \binom{n-۲}{k-۱} [g(x_i)f(x_j) + f(x_i)g(x_j)] \\ & + \binom{n-۲}{k} [f(x_i)f(x_j)] \Big\}, \end{aligned} \quad (٤)$$

و

$$\begin{aligned} f(x_i, x_j, x_t) = & \\ & \frac{۱}{\binom{n}{k}} \Big\{ \binom{n-۳}{k-۳} [g(x_i)g(x_j)g(x_t)] \\ & + \binom{n-۳}{k-۲} [f(x_i)g(x_j)g(x_t) \\ & + g(x_i)f(x_j)g(x_t) \\ & + g(x_i)g(x_j)f(x_t)] \\ & + \binom{n-۳}{k-۱} [f(x_i)f(x_j)g(x_t) \\ & + f(x_i)g(x_j)f(x_t) \\ & + g(x_i)f(x_j)f(x_t)] \\ & + \binom{n-۳}{k} [f(x_i)f(x_j)f(x_t)] \Big\}. \end{aligned} \quad (۵)$$

تذکر: لازم به ذکر است که در مدل بالا، متغیرهای تصادفی $(X_۱, X_۲, ..., X_n)$ مستقل نمی باشند.

مثال ۵: فرض کنید متغیرهای تصادفی $(X_۱, X_۲, ..., X_n)$ به صورتی باشند که هر k تا از آن ها مستقل و هم توزیع پارتو با پارامترهای α و $\beta\theta$ به صورت زیر باشد.

$$g(x) = \frac{\alpha(\beta\theta)^{\alpha}}{x^{\alpha+۱}}, \quad \beta\theta \leq x, \beta \geq ۱, \theta > ۰, \alpha > ۰,$$

و $n-k$ تای باقی مانده نیز مستقل و هم توزیع پارتو و دارای تابع چگالی زیر باشد.

$$f(x) = \frac{\alpha\theta^{\alpha}}{x^{\alpha+۱}}, \quad \theta \leq x, \ \theta > ۰, \alpha > ۰.$$

آنگاه مطلوب است، محاسبه تابع چگالی توأم $(X_۱, X_۲, ..., X_n)$ و همچنین محاسبه تابع چگالی X و چگالی توأم $(X_i, X_j), i \neq j$.

حل: با توجه به مدل بیان شده می توان نوشت

$$f(x_1,x_2,\dots,x_n)=\prod_{i=1}^{n}\frac{\alpha\theta^{\alpha}}{x_i^{\alpha+1}}\times \sum_{(i_1,\dots,i_k)}\prod_{j=1}^{k}\frac{\frac{\alpha(\beta\theta)^{\alpha}}{x_{i_j}^{\alpha+1}}I\left(x_{i_j}-\beta\theta\right)}{\frac{\alpha\theta^{\alpha}}{x_{i_j}^{\alpha+1}}I\left(x_{i_j}-\theta\right)}[C(n,k)]^{-1}.$$

آن گاه عبارت زیر بدست می آید

$$f(x_1,x_2,\dots,x_n)=\frac{\alpha^n\theta^{n\alpha}\beta^{k\alpha}}{C(n,k)}\left(\prod_{i=1}^{n}x_i\right)^{-(\alpha+1)}\times \sum_{(i_1,\dots,i_k)}\prod_{j=1}^{k}I\left(x_{i_j}-\theta\right)I\left(x_{i_j}-\beta\theta\right).$$

از طرفی داریم

$$h(x)=\frac{k}{n}\frac{\alpha\left(\beta\theta\right)^{\alpha}}{x^{\alpha+1}}+\frac{n-k}{n}\frac{\alpha\theta^{\alpha}}{x^{\alpha+1}}=\frac{\alpha\theta^{\alpha}}{x^{\alpha+1}}\left(\frac{k}{n}\beta^{\alpha}+\frac{n-k}{n}\right).$$

همچنین برای $i\neq j$ رابطه ی زیر برقرار است

$$h(x_i,x_j)=\frac{1}{\binom{n}{k}}\left\{\binom{n-2}{k-2}\left[\frac{\alpha(\beta\theta)^{\alpha}}{x_i^{\alpha+1}}\times\frac{\alpha(\beta\theta)^{\alpha}}{x_j^{\alpha+1}}\right]+\binom{n-2}{k-1}\left[\frac{\alpha(\beta\theta)^{\alpha}}{x_i^{\alpha+1}}\times\frac{\alpha\theta^{\alpha}}{x_j^{\alpha+1}}+\frac{\alpha\theta^{\alpha}}{x_i^{\alpha+1}}\times\frac{\alpha(\beta\theta)^{\alpha}}{x_j^{\alpha+1}}\right]+\binom{n-2}{k}\left[\frac{\alpha\theta^{\alpha}}{x_i^{\alpha+1}}\times\frac{\alpha\theta^{\alpha}}{x_j^{\alpha+1}}\right]\right\},$$

و پس از ساده کردن داریم

$$h(x_i,x_j)=\frac{\alpha^2\theta^{2\alpha}}{n(n-1)(x_ix_j)^{\alpha+1}}\{k(k-1)\beta^{2\alpha}+2k(n-k)\beta^{\alpha}+(n-k)(n-k-1)\};i\neq j.$$

مثال ٦: در مثال ٥، اگر فرض کنید $n=$۱۰ و $k=$۱، تابع چگالی توأم ده متغیر تصادفی و تابع چگالی حاشیه ای X و $(X_i,X_j),i\neq j$ را می نویسیم.

حل: با جای گذاری در فرمول ها داریم

$$f(x_1,x_2,\dots,x_{10})=\frac{\alpha^{10}\theta^{10\alpha}\beta^{\alpha}}{9}\left(\prod_{i=1}^{10}x_i\right)^{-(\alpha+1)}\times\sum_{i=1}^{10}I(x_i-\theta)I(x_i-\beta\theta),$$

$$h(x)=\frac{\alpha\theta^{\alpha}}{x^{\alpha+1}}\left(\frac{1}{10}\beta^{\alpha}+\frac{9}{10}\right),$$

$$h(x_i,x_j)=\frac{\alpha^2\theta^{2\alpha}}{90\left(x_ix_j\right)^{\alpha+1}}\left\{18\beta^{\alpha}+72\right\},\quad i\neq j.$$

مثال ۷: موضوع خسارت وارده به ماشین های در حال تردد در داخل شهرها در اثر تصادف را در نظر بگیرید. اگر اتومبیل ها دارای بیمه باشند، وقتی تصادفی صورت می گیرد، شرکت بیمه متعهد است که از طرف راننده مقصر به طرف غیر مقصر مبلغی را بعنوان خسارت پرداخت نماید. می دانیم توزیع مشاهدات مبلغ خسارت از مدل پارتو پیروی می کند. اما اگر در نظر بگیریم که تعدادی از اتومبیل هایی که تصادف نموده اند، گران قیمت بوده و یا خسارت وارده به آن ها شدید می باشد (طبیعی است که تعداد این اتومبیل ها نسبت به عمده اتومبیل ها کم است)، لذا این مشاهدات خسارت دارای توزیع پارتو ولی با پارامتر متفاوت است (با توجه به توضیحات قبلی این مشاهدات، داده های پرت هستند). می دانیم که خسارت وارده کمتر از ۵۰۰ هزار ریال از نظر اقتصادی به صرفه نیست تا این که برای دریافت خسارت از شرکت بیمه اقدام صورت گیرد. لذا مبالغ خسارت وارده دارای توزیع پارتو با پارامتر معلوم $\theta=$۵۰۰۰۰۰ ریال و پارامتر نامعلوم α می باشد. اما اگر فرض کنیم برای آن اتومبیلی هایی که تصادف شدیدتر داشته و یا گران تر بوده اند، مبلغ خسارت حداقل۱/۵ برابر می باشد، این مشاهدات دارای توزیع پارتو با پارامتر $\beta\theta=$۱.۵×۵۰۰۰۰۰=۷۵۰۰۰۰ ریال و پارامتر مجهول α خواهد بود. بنابراین اگر یک نمونه تصادفی به حجم ۱۰ از داده های خسارت وارده به اتومبیل ها را داشته باشیم، کل مشاهدات بدون آن که بدانیم کدامیک داده پرت است، دارای توزیع پارتو در مواجهه با داده های پرت با پارامترهای معلوم $\theta=$۵۰۰۰۰۰، $\beta=$۱.۵ و پارامتر مجهول α می باشد. فرض کنید داده به شرح ذیل باشند.

۷۵۰۰۰۰، ۷۸۰۰۰۰، ۱۲٤۰۰۰۰، ۷٦۵۰۰۰۰، ۱۷۵۰۰۰۰، ۱٤۵۰۰۰۰، ۱٦۳۰۰۰۰، ٤۷۲۵۰۰۰، ۷٦۰۰۰۰، ۱۲۸۰۰۰۰.

برآورد روش گشتاوری و درست نمایی ماکزیمم پارامتر α را می خواهیم با فرض این که یک مشاهده پرت

داشته باشد، بیابیم و آن را با برآورد های متناظر بدون در نظر گرفتن یک مشاهده پرت مقایسه کنیم.

حل: با توجه به مثال ٦ و از روی گشتاور مرتبه یک توزیع داریم

$$\mu_1' = E(X) = \int xh(x)dx = \frac{1}{10}\int_{\beta\theta}^{\infty} x\frac{\alpha(\beta\theta)^{\alpha}}{x^{\alpha+1}}dx + \frac{10-1}{10}\int_{\theta}^{\infty} x\frac{\alpha\theta^{\alpha}}{x^{\alpha+1}}dx = \frac{\alpha\theta}{\alpha-1}\left(\frac{1}{10}\beta+\frac{9}{10}\right).$$

بنابراین با حل معادله فوق می توان مقدار پارامتر مجهول α را برحسب پارامترهای معلوم به صورت زیر یافت

$$\hat{\alpha}_{mm} = \frac{\hat{\mu}_1'}{\hat{\mu}_1' - \theta(\frac{1}{10}\beta+\frac{9}{10})}, \quad (٦)$$

حال با جای گذاری مقادیر معلوم پارامترهای β و θ داریم

$$\hat{\alpha}_{mm} = \frac{\hat{\mu}_1'}{\hat{\mu}_1' - ۵۰۰۰۰۰(\frac{1}{10}\times ۱.۵+\frac{9}{10})} = \frac{\hat{\mu}_1'}{\hat{\mu}_1' - ۵۲۵۰۰۰}.$$

با توجه به داده ها گشتاور مرتبه یک نمونه، یعنی میانگین مشاهدات، $\bar{X} = ۲۲۰۱۵۰۰$ است. لذا با جای گذاری میانگین در رابطه قبل می توان برآورد روش گشتاوری پارامتر مجهول را بدست آورد. یعنی

$$\hat{\alpha}_{mm} = \frac{۲۲۰۱۵۰۰}{۲۲۰۱۵۰۰-۵۲۵۰۰۰} = ۱.۳۱۳.$$

برای محاسبه برآورد روش درست نمایی ماکزیمم پارامتر مجهول بایستی با استفاده از تابع چگالی توام، تابع درستنمایی را محاسبه و ماکزیمم نمود. از این رو با کمی محاسبات خواهیم داشت.

$$\hat{\alpha}_{ml} = \frac{n}{\sum_{i=1}^{n}\ln(x_i) - n\ln(\theta) - k\ln(\beta)} \quad (٧)$$

که در آن ln نماد لگاریتم طبیعی است. در این صورت با استفاده از داده ها داریم

$$\hat{\alpha}_{ml} = \frac{۱۰}{۱۴۲.۸۱۳۸-۱۰\times\ln(۵۰۰۰۰۰)-۱\times\ln(۱.۵)} = ۰.۸۹۴۱.$$

از طرفی، اگر فرض کنیم همه داده ها از توزیع پارتو با پارامتر معلوم $\theta = ۵۰۰۰۰۰$ و پارامتر مجهول α آمده باشد (بدون در نظر گرفتن داده های پرت)، برای یافتن برآورد گشتاوری و درستنمایی ماکزیمم پارامتر مجهول کافی است که در فرمول های (٦) و (٧) مقدار β را برابر یک قرار دهیم. لذا

$$\tilde{\alpha}_{mm} = \frac{\hat{\mu}_1'}{\hat{\mu}_1' - \theta} = \frac{۲۲۰۱۵۰۰}{۲۲۰۱۵۰۰-۵۰۰۰۰۰} = ۱.۲۹۳۹,$$

$$\tilde{\alpha}_{ml} = \frac{n}{\sum_{i=1}^{n} ln(x_i) - nln(\theta)} = \frac{10}{142.8138-10\times\ln(500000)} = 0.8628.$$

بنابراین مشاهده می شود که مقادیر برآورد پارامتر مجهول در هر دو روش، در حالتی که داده ها از توزیع پارتو با داده های پرت آمده باشند، بیش تر از حالتی است که داده های پرت را در نظر نگیریم.

نتیجه گیری

با استفاده از مثال ها و توضیحات بیان شده می توان درک نمود که اگر از نظر تئوری به این نتیجه برسیم که ممکن است در داده ها، مشاهده پرت وجود داشته باشد، نمی توان آن ها را حذف نمود. بلکه باید با استفاده از مدل های مناسب در حضور داده های پرت، برآورد یا استنباط را انجام داد.

سپاسگزاری

بر خود لازم می بینیم تا از زحمات داوران و سردبیر محترم بخاطر ارائه نکات موثر و مفید در ارتقای این مقاله، کمال تشکر و قدردانی را داشته باشیم.

منابع

[1]. Anscombe, F. J. (1960). Rejection of outliers. Technometrics 2, 123-147.
[2]. Barnett, V. and Lewis, T. (1994). Outliers in Statistical Data, 3rd ed. Wiley.
[3]. Beckman, R. and Cook, R. D. (1983). Outliers (with discussion). Technometrics 25, 119–149.

[4]. Bol'shev, L. N. (1969). On tests for rejecting outlying observations. Trudy In-ta prikladnoi Mat. Tblissi Gosudart. Univ., 2, 159–177 (in Russian).

[5]. Chikkagoudar, M. S. and Kunchur, S. H. (1980). Estimation of the mean of an exponential distribution in the presence of an outlier. Canadian Journal of Statistics, 8, 59-63.

[6]. Collet, D. and Lewis, T. (1976). The subjective nature of outlier rejection procedures. Appl. Statist., 25, 228-237.

[7]. Dixit, U. J. (1987). Characterization of the gamma distribution in the presence of k outliers. Bull. Bombay Mathematical Colloquium, 4, 54-59.

[8]. Dixit, U. J. (1989). Estimation of parameters of the gamma distribution in the presence of outliers. Communications in Statistics Theory and Methods, 18, 3071-3085.

[9]. Dixit, U. J. and Jabbari Nooghabi, M. (2010). Efficient estimation of the parameters of the Pareto distribution in the presence of outliers. Submitted.

[10]. Dixit, U. J. and Jabbari Nooghabi, M. (2010). Testing the parameters of a Pareto distribution in the presence of outliers. Submitted.

[11]. Dixit, U. J. and Jabbari Nooghabi, M. (2011). Efficient estimation in the Pareto distribution with the presence of outliers. Statistical Methodology, 8(4), 340-355.

[12]. Dixit, U. J. and Jabbari Nooghabi, M. (2010). Characterizations of the Pareto distribution in the presence of outliers. Submitted.

[13]. Dixit, U. J. and Jabbari Nooghabi, M. (2010). Bayesian inference for the Pareto lifetime model in the presence of outliers under progressive censoring with binomial removals. Submitted.

[14]. Dixit, U. J. and Jabbari Nooghabi, M. (2011). Estimation of parameters of gamma distribution in the presence of outliers in right censored samples. Aligar Journal of Statistics, 31.

[15]. Ferguson, T. S. (1961). On the rejection of outliers. Proc. 4th Berkeley symp. 1, 253-287.

[16]. Grubbs, F. E. (1950). Sample criteria for testing outlying observations. Ann. Math. Statist., 21, 27-58.

[17]. Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. Technometrics, 11, 1-21.

[18]. Hawkins, D. M. (1980). Identification of outliers, London: Chapman and Hall.

[19]. Jabbari Nooghabi, M. and Jabbari Nooghabi, H. (2009). Efficient estimation of pdf and cdf for the exponentiated Pareto distribution. ICCS-X Conference, The American University in Cairo, December 20–23.

[20]. Jabbari Nooghabi, M., Jabbari Nooghabi, H. and Nasiri, P. (2009). Estimation of parameters of the gamma distribution in the presence of outliers generated from uniform distribution. Pakistan Journal of Statistics, 25(1), 15-26.

[21]. Jabbari Nooghabi, M., Jabbari Nooghabi, H. and Nasiri, P. (2010). Detecting outliers in the Gamma distribution. Communication in Statistics Theoy and Methods, 39(4), 698–706.

[22]. Jabbari Nooghabi, M. and Jabbari Nooghabi, H. (2011). Detecting outliers in the Pareto distribution. Submitted.

[23]. Joshi P. C. (1972). Efficient estimation of the mean of an exponential distribution when an outlier is present. Technometrics, 14(1), 137-143.

[24]. Kabe, D. G. (1970). Testing outliers from an exponential population. Metrika, 15, 15–18.

[25]. Kale, B. K. and Sinha, S. K. (1971). Estimation of expected life in the presence of an outlier observation. Technometrics, 13, 755-759.

[26]. Kendall, M. G. and Buckland, W. R. (1957). A Dictionary of Statistical Terms. London: Longman.

[27]. Likes, J. (1966). Distribution of Dixon's statistics in the case of an exponential population. Metrika, 11, 46–54.

[28]. Miller, R. G. Jr. (1981). Simultaneous Statistical Inference, 2nd ed. New York: Springer Verlag.

[29]. Nasiri, P. and Jabbari Noogabi, M. (2010). Estimation of the parameters of the generalized exponential distribution in the presence of outliers generated from uniform distribution. Journal of Statistical Sciences, 1(1), 185-195.

[30]. Rauhut, B. O. (1982). Estimation of the mean of an Exponential distribution with an outlying observation. Communications in Statistics Theory and Methods, 11, 1439-1452.

[31]. Sinha, S.K. (1973). Distribution of order statistics and estimation of mean life when an outlier may be present. The Canadian Journal of Statistics, 1(1), 119-121.

[32]. Veale, J. R. (1975). Improved Estimation of expected life when one identified spurious observation may be present. J. Amer. Statist. Ass., 70, 398-401.

[33]. Zerbet, A. and Nikulin, M. N. (2003). A new statistic for detecting outliers in exponential case. Communications in Statistics Theory and Methods, 32(3), 573–583.