مجله دانشکده پزشکی، دانشگاه علوم پزشکی تهران، دوره ۶۵ ، شماره ۷، مهر ۱۳۸۶، ۲۷-۲۴

روش‌های تعیین داده‌های پرت در مطالعات پزشکی

غلامرضا بابایی[1*]
فیروز امانی[1]
اکبر بیگلریان[1]
مریم کشاورز[2]

۱. گروه آمار زیستی، دانشگاه تربیت مدرس تهران
۲. دانشگاه علوم پزشکی ایران، دانشکده پرستاری و مامایی

*نویسنده مسئول: تهران، خیابان جلال آل احمد، صندوق پستی ۳۳۱-۱۴۱۱۵ تلفن: ۷۷۸۸۳۲۸۵
email: babaee_g@modares.ac.ir

چکیده

زمینه و هدف: مشاهده‌ای که معمولاً نسبت به مقادیر دیگر در بین مجموعه داده‌ها بزرگتر یا کوچکتر است، داده پرت نامیده می‌شود. وجود داده‌های پرت در اکثر موارد منجر به اختلال در نتیجه‌گیری از اطلاعات خواهد شد. شناسایی داده‌های پرت توسط پژوهشگران و کلیه کسانی‌که به نوعی با اطلاعات جمع‌آوری شده سر و کار دارند، حائز اهمیت است باید از وجود یا عدم وجود داده‌های پرت، چگونگی تاثیرگذاری و نحوه رفع داده‌های پرت اطلاع حاصل نموده و داده‌ها را کنترل کنند. در این مقاله سعی شده با ارائه تکنیک‌های شناسایی داده‌های پرت و نحوه برخورد با این نوع از داده‌ها، خطای ناشی از وجود چنین داده‌هایی را به حداقل برسانیم. ***روش بررسی:*** در این مقاله تکنیک‌های مختلف تعیین داده‌های پرت بر روی قد ۳۰ نفر از دانشجویان دانشکده پزشکی دانشگاه تربیت مدرس تهران که توسط متر خیاطی و با قرار دادن فرد در روی یک سطح صاف، اندازه‌گیری شدند مورد بررسی قرار گرفتند. از جمله این تکنیک‌ها می‌توان به آزمون z، آزمون گراب و روش‌های گرافیکی اشاره نمود. ***یافته‌ها:*** تکنیک‌های فوق بیانگر وجود داده‌های پرت در مشاهدات ۱۵۳ و ۱۱۰ مربوط به قد افراد، بودند که با استفاده از جدول و نمودار نشان داده شد. ***نتیجه‌گیری:*** نتایج پژوهش نشان داد که همه تکنیک‌ها در تعیین داده‌های پرت مفید بودند و از این میان استفاده از چارک‌ها در شناسایی داده‌های پرت خفیف و شدید از اهمیت بالایی برخوردار هستند. همچنین آزمون گراب با در اختیار گذاشتن سطح معنی‌داری (p-value)، در شناسایی داده‌های پرت بسیار مفید است.

کلمات کلیدی: داده پرت، چارک‌ها، آزمون گراب.

مقدمه

مشاهده‌ای که معمولاً نسبت به مقادیر دیگر در بین مجموعه داده‌ها بزرگ‌تر یا کوچک‌تر است، یک داده پرت نامیده می‌شود. همچنین در تعریف ساده دیگر می‌توانیم بگوییم دادهٔ پرت مشاهده‌ای است که در فاصله دورتری از سایر داده‌ها قرار می‌گیرد و از مقدار مورد انتظاری که داریم بیشتر می‌باشد.[۱-۳] داده‌های پرت اساساً به یکی از دلایل زیر اتفاق می‌افتند: ۱- غیر صحیح بودن اندازه‌گیری مشاهده شده، ثبت شده یا وارد شده در کامپیوتر، ۲- جمع‌آوری اندازه‌گیری‌ها از جوامع مختلف، ۳- اندازه‌گیری برای بیان یک حادثه یا رویداد نادر، ۴- چولگی بیشتر مجموعه داده‌ها در منحنی توزیع فراوانی نسبی. با توجه به اینکه داده‌های پرت در تمام مراحل مربوط به آنالیز و تفسیر اطلاعات چه از لحاظ ساختاری و چه از لحاظ مفهومی تاثیرگذار هستند و بعضی موارد امکان نتیجه‌گیری منطقی از اطلاعات جمع‌آوری شده وجود ندارد و دچار خطاهای علمی آماری از لحاظ پایایی، روایی و غیره می‌شویم، لذا آشنایی با روش‌های شناسایی و تشخیص داده‌های پرت در اطلاعات جمع‌آوری شده می‌تواند برای پژوهشگران، صاحبان علم، تحلیل‌گران داده، محققین علوم پزشکی، علوم اجتماعی و غیره مفید باشد. هدف از این مقاله، ارائه تکنیک‌های موجود در زمینه شناسایی داده‌های پرت می‌باشد.

روش بررسی

با توجه به بررسی‌های وسیع میدانی در زمینه داده‌های پرت و روش‌های مورد نظر و بحث‌های مرتبط در این زمینه، چندین روش عمومی را برای تعیین داده‌های پرت ارائه می‌شود:[۴-۶] استفاده از روش

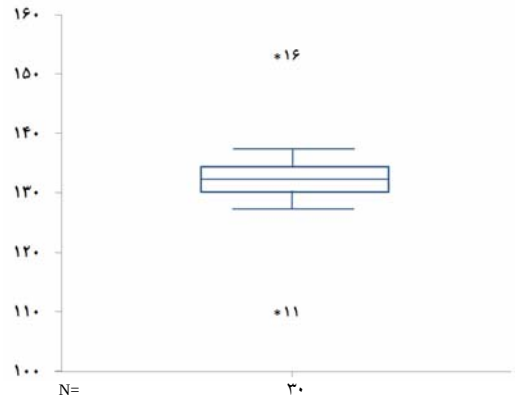
استانداردسازی Z، استفاده از روش نمودار جعبه‌ای، نمودار پراکنش، آزمون گراب، استفاده از نمودار هیستوگرام، استفاده از چارک‌ها در تعیین داده‌های پرت خفیف و شدید. مطابق قضیه چی‌بی‌شف همه مشاهدات (تقریبا" در حدود ۹۹/۸٪) در مجموعه داده‌ها دارای مقادیر (نمرات Z) کمتر از سه هستند که در فاصله ($\bar{X} \pm 3S$) قرار می‌گیرند که در آن $\bar{X}$ میانگین و S انحراف معیار نمونه می‌باشد. بنابراین مشاهدات دارای مقدار قدرمطلق Z بزرگتر از سه پرت خواهند بود. روش دیگر در تعیین داده‌های پرت رسم نمودار جعبه‌ای است. البته روش‌های آماری زیادی برای رسم نمودار جعبه‌ای وجود دارند. بعد از رسم نمودار جعبه‌ای برای هر مجموعه داده‌ای، مشاهداتی که مابین دیواره‌های داخلی و خارجی قرار می‌گیرند، مشکوک به پرت هستند. مشاهداتی که بیرون از دیواره‌های خارجی قرار می‌گیرند، جزء داده‌های پرت هستند. برای داده‌های بزرگ نمودار جعبه‌ای با استفاده از برنامه کامپیوتری ساخته می‌شود. نمودار پراکنش از شیوه‌های ساده دیگر برای تعیین داده‌های پرت است. بدین صورت که بعد از رسم نمودار پراکنش بین دو متغیر وابسته Y و متغیر مستقل X، اگر داده‌ای پرت باشد روی خط برازش شده برای داده‌ها قرار نگرفته و در فاصله دورتری از خط قرار می‌گیرد. لذا خط رگرسیونی رسم شده برای داده‌ها با احتساب مقدار داده پرت دارای برازش خوب نبوده و بدون در نظر گرفتن داده پرت برازش مدل دقیق‌تر خواهد شد. آمار دانان راه‌های مختلفی را برای تعیین داده‌های پرت بیان کرده‌اند. آزمون گراب یکی از شیوه‌های ساده در این زمینه است که توسط گراب در سال ۱۹۶۹ [۶-۸] بیان گردید. آزمون گراب دارای مقدار آماره‌ای است که بر اساس اطلاعات موجود در نرم افزارهای مربوطه قابل محاسبه می‌باشد لذا بعد از محاسبه مقدار آماره آزمون گراب اگر این مقدار از مقدار استاندارد تک‌تک داده‌ها بیشتر باشد آن داده از دید آزمون گراب پرت به حساب می‌آید. انجام آزمون گراب با استفاده از برنامه‌هایQuick Calc و Data Plot صورت می‌گیرد که Online قابل دسترس می‌باشد. کلیه روش‌ها را بر روی اطلاعات مربوط به قد یک نمونه ۳۰ نفری از دانشجویان پزشکی به‌کار برده و بررسی نمودیم. در داده‌های دارای توزیع نرمال اگر نمودار هیستوگرام رسم گردد، وجود یک داده پرت می‌تواند الگوی توزیع داده‌ها را تغییر داده و داده‌ها شکل یک توزیع چوله‌دار را به خود بگیرند. همچنین وجود یک داده پرت باعث می‌گردد که آن داده در فاصله دورتری از بقیه داده‌ها قرار گرفته و بعد از رسم نمودار هیستوگرام به‌عنوان یک نقطه دور در طرف چپ و راست توزیع قرار بگیرد و نوع شکل رسم شده را تحت تاثیر قرار دهد. اگر فرض کنیم که داده‌ها دارای چارک اول Q_1 و چارک سوم Q_3 با دامنه میان چارکی $IQR=Q_3-Q_1$ باشند، در این‌صورت مشاهداتی را که در نامساوی‌های $X_i< Q_1-$ ۱/۵ IQR یا $X_i> Q_1+$ ۱/۵ IQR قرار دارند جزء داده‌های پرت خفیف و مشاهداتی را که در نامساوی‌های $X_i< Q_1-$ ۳ IQR یا $X_i> Q_1+$ ۳ IQR قرار می‌گیرند جزء داده‌های پرت شدید هستند.[۹و۱۰]

یافته‌ها

اگر برای نمونه ۳۰ نفری از دانشجویان پزشکی اطلاعات قد به صورت Mean= ۱۳۲/۷۷، S= ۶/۰۶، Z(۱۵۳)= ۳/۳۴ و Z(۱۱۰)= -۳/۷۶ باشد، در این صورت قدرمطلق مقادیر Z برای قدهای ۱۵۳ و ۱۱۰ بیشتر از سه هستند. لذا قدهای ۱۵۳ و ۱۱۰ در مجموعه داده‌ها پرت به‌حساب می‌آیند. با رسم نمودار جعبه‌ای برای اطلاعات مربوط به قد نمونه ۳۰ نفری (شکل ۱) مشاهده شد که

جدول-۱: خروجی آزمون گراب، اجرا شده در برنامه Quick Calc

آماره‌های توصیفی		خروجی آزمون	
تعداد مشاهدات	۲۸	مقدار مشاهده شده	مقدار استاندارد شده	معنی‌داری نقطه پرت
		۱۳۲	۰/۳۸	–
میانگین	۱۳۳/۶۸	۱۳۵	۰/۳	–
انحراف معیار	۴/۳۷	۱۳۴	۰/۰۷	–
		۱۳۱	۰/۶۱	–
سطح معنی‌داری آزمون	۰/۰۵	۱۵۳	۴/۴۲	p< ۰/۰۵
		۱۳۱	۰/۶۱	–
آماره آزمون	۲/۸۸	۱۳۶	۰/۵۳	–



شکل-۱: نمودار جعبه‌ای رسم شده برای اطلاعات مربوط به قد دانشجویان

داده‌های مربوط به قد نفرات شش و ۱۱ یعنی اعداد ۱۱۰ و ۱۵۳ نیز جزء داده‌های پرت هستند. برای مثال در اطلاعات مربوط به قد دانشجویان داریم: Q_1=۱۳۱، Q_2=۱۳۳، Q_3=۱۳۵، IQR=۱۳۵-۱۳۱=۴. پس با احتساب مقادیر فوق اگر مشاهدات از اعداد ۱۲۵ کمتر و ۱۴۱ بیشتر باشند، جزء داده‌های پرت خفیف ولی اگر از اعداد ۱۱۹ کمتر و ۱۴۷ بیشتر باشند، جزء داده‌های پرت شدید هستند. لذا با در نظر گرفتن شرایط فوق مقادیر ۱۱۰ و ۱۵۳ جزء داده‌های پرت شدید هستند که با نتایج سایر روش‌ها مشابه می‌باشند.

بحث

همان‌طور که از بررسی نتایج مربوط به هر کدام از روش‌های مورد بررسی در تعیین داده‌های پرت در بین اطلاعات موجود بر می‌آید، مشاهده می‌شود که همه روش‌ها می‌توانند به‌نوعی در تعیین مشاهدات دورافتاده به ما کمک کنند. لذا توصیه می‌شود هر فرد پژوهشگر و هر تحلیل‌گر آماری قبل از بکارگیری تحلیل‌های آماری، پس از مدیریت داده‌ها در ابتدا با کمک آماره‌های توصیفی شمای کلی از توزیع مشاهدات را بررسی نماید. در ادامه در صورت وجود داده‌های مشکوک در بین داده‌های جمع‌آوری شده، وضعیت آنها را (به لحاظ پرت بودن) با تکنیک‌های مورد بحث بررسی و در صورت شناسایی داده‌های از نوع پرت در جهت حذف آنها تصمیم‌گیری نماید. پیشنهاد می‌شود که در کنار روش‌های ترسیمی از آزمون داده‌های پرت، نظیر آزمون گراب، چارک‌ها جهت تصمیم‌گیری متقن در مورد این نوع از داده‌ها استفاده و در مرحله پایانی اقدام به تحلیل اطلاعات و نتیجه‌گیری کلی شود. از این میان، استفاده از چارک‌ها در شناسایی داده‌های پرت خفیف و شدید از اهمیت بالایی برخوردار هستند. همچنین آزمون گراب با در اختیار گذاشتن سطح معنی‌داری (p-value) در شناسایی داده‌های پرت بسیار مفید است.

References

1. What are outliers in the data? 2006: [http://www.itl.nist.gov/div-898/handbook/prc/section1/prc16.html]. Available from: URL.
2. Fahy T. Arcadia Financial: The Impact of Outliers on Absolute Returns in Trend Following Trading Systems. *Retrieved on* 2006; 10-29.
3. Moore DS, McCabe GP. Introduction to the Practice of Statistics, 3rd ed. New York: WH Freeman: 1989.
4. Outlier RJ. From MathWorld: A Wolfram Web Resource. created by Weisstein EW. [cited 2004 Sep 22]; [http://mathworld.-wolfram.com/Outlier.html]. Available from: URL.
5. High R. Dealing with Outliers. How to Maintain Your Data's Integrity. [http://cc.uoregon.edu/cnews/spring2000/outliers.html]. Available from: URL.
6. Berthouex, PM, Brown LC. Statistics for Environmental Engineers. 1st ed. Florida: CRC Press, Boca Raton: 1994.
7. Gibbons RD. Statistical Methods for Groundwater Monitoring. 3rd ed. New York: John Wiley & Sons Inc: 1994.
8. Fallon A, Spada C. Detection and Accommodation of Outliers in Normally Distributed Data Sets [cited 1997]; [http://www.cee.vt.edu/ewr/environmental/teach/smprimer/outlier/outlier.html]. Available from: URL.
9. Laurikkala J, Juhola1 M, Kentala E. Informal identification of outliers in medical data. [http://ai.ijs.si/Branax/idamap-2000_AcceptedPapers/Laurikkala.pdf]. Available from: URL.
10. Charu C. Aggarwal, Philip S. Yu. Outlier Detection for High Dimensional Data. [cited 2001]; [http://citeseer.ist.psu.edu/agar-wal01outlier.html]. Available from: URL.

Tehran University Medical Journal; Vol. 65, No. 7, Oct 2007: 24-27

Detection of Outliers methods in medical studies

Babaee Gh. [*1]
Amani F.[1]
Biglarian A.[1]
Keshavarz M.[2]

1- Department of Biostatistics, Faculty of Medical Science, Tarbiat Modares University
2- Faculty of Nursing & Midwifery, Iran University of Medical Sciences

* Corresponding author: Tarbiat Modares University, Jalal Ale Ahmad Ave., Tehran Tel: +98-21-88011001
email: babaee_g@modares.ac.ir

Abstract

***Background*:** An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Outliers sometimes deal with to abnormality in obtained results from collected data and information. known outlier data by researchers, physicians and other persons that work in medical fields and sciences is important and they must control data before getting result about outlier data, effect of them in information bias and how to remove & control to obtain minimum bias and exact data .in this paper we had trying by known technique and tests to control them and minimized the errors related to them.

Methods: This paper has been done on 30 student's height in Tarbiat Modares University that measured by meter in smoothing area. We applied some methods such as; Z-test, Grub test and graphical methods to determine outliers. In this paper the advantage and disadvantage of methods were evaluated and finally compares with each other.

***Results*:** The above tests showed that the data values 153, 110 among collected data were outliers. All of the methods showed that the above data were outliers. Calculation quartiles and intermediate quartiles showed that the observations under 125 and upper 141 were mind outliers and if the observation under 119 and upper 147 is the sever outliers. According to upper situations the amounts of 110 and 153 is the sever outliers and resulted from all methods.

***Conclusion*:** The results showed that all methods were useful in determine outlier data and between them Quartiles were important to known severe and mild outliers. Also Grub test with p-Value is very useful to report outliers.

Keywords: Outlier, quartiles, grubs test.