

روش‌های تعیین داده‌های پرت در مطالعات پزشکی

چکیده

زمینه و هدف: مشاهده‌ای که معمولاً^۱ 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 کسانی که 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.

غلامرضا بابایی^{۱*}

فیروز امانی^۱

اکبر بیگلریان^۱

مریم کشاورز^۲

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 ۱۴۱۱۵-۳۳۱ تلفن: ۷۷۸۸۳۲۸۵
email: babaee_g@modares.ac.ir

مقدمه

هستند و بعضی موارد امکان نتیجه‌گیری منطقی از اطلاعات جمع‌آوری شده وجود ندارد و دچار خطاهای علمی آماری از لحاظ پایایی، روایی و غیره می‌شوند، لذا آشنایی با روش‌های شناسایی و تشخیص داده‌های پرت در اطلاعات جمع‌آوری شده می‌تواند برای پژوهشگران، صحابان 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 موردنظر و بحث‌های مرتبط در این زمینه، چندین روش عمومی را برای تعیین داده‌های پرت ارائه می‌شود^{۳-۴}. استفاده از روش

مشاهده‌ای که معمولاً نسبت به مقادیر دیگر در بین مجموعه داده‌ها بزرگتر است، یک داده پرت نامیده می‌شود. همچنین در تعریف ساده دیگر می‌توانیم بگوییم داده 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 می‌افتد: 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

گرای اگر این مقدار از مقدار استاندارد تک تک داده ها بیشتر باشد آن داده از دید آزمون گرای پرت به حساب می آید. انجام آزمون گرای با استفاده از برنامه های Quick Calc و Data Plot صورت می گیرد که Online قابل دسترس می باشد. کلیه روش ها را بر روی اطلاعات مربوط به قدیمی نمونه 30 نفری از دانشجویان پژوهشی به کار برد و بررسی نمودیم. در داده های دارای توزیع نرمال اگر نمودار هیستوگرام رسم گردد، وجود یک داده پرت می تواند الگوی توزیع داده ها را تغییر داده و داده ها شکل یک توزیع چوله دار را به خود بگیرند. همچنین وجود یک داده پرت باعث می گردد که آن داده در فاصله دورتری از بقیه داده ها قرار گرفته و بعد از رسم نمودار هیستوگرام به عنوان 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 داده ها دارای چارک اول Q_1 و چارک سوم Q_3 با دامنه میان چارکی $IQR = Q_3 - Q_1$ باشند، در این صورت مشاهداتی را که در نامساوی های $X_i < Q_1 - 1/5 IQR$ یا $X_i > Q_1 + 1/5 IQR$ قرار دارند جزء داده های پرت خفیف و مشاهداتی را که در نامساوی های $X_i < Q_1 - 3 IQR$ یا $X_i > Q_1 + 3 IQR$ قرار می گیرند جزء داده های پرت شدید هستند.^۹

یافته ها

اگر برای نمونه 30 نفری از دانشجویان پژوهشی اطلاعات قدر به صورت $Z = \frac{X - \text{Mean}}{S}$ می باشد، در این صورت قدر مطلق مقادیر Z برای قدهای 110 و 110 بیشتر از سه هستند. لذا قدهای 103 و 110 در مجموعه داده ها پرت به حساب می آیند. با رسم نمودار جعبه ای برای اطلاعات مربوط به قدر نمونه 30 نفری (شکل ۱) مشاهده شد که

استانداردسازی Z استفاده از روش نمودار جعبه ای، نمودار پراکنش، آزمون گرای، استفاده از نمودار هیستوگرام، استفاده از چارک ها در تعیین داده های پرت خفیف و شدید. مطابق قضیه چی بی شف همه مشاهدات (تقریباً در حدود $99/8\%$) در مجموعه داده ها دارای مقادیر (نمرات Z) کمتر از سه هستند که در فاصله $(\bar{X} \pm 3S)$ قرار می گیرند که در آن \bar{X} میانگین و S انحراف معیار نمونه می باشد. بنابراین مشاهدات دارای مقدار قدر مطلق Z بزرگتر از سه پرت خواهد بود. روش دیگر در تعیین داده های پرت رسم نمودار جعبه ای است. البته روش های آماری زیادی برای رسم نمودار جعبه ای وجود دارند. بعد از رسم نمودار جعبه ای برای هر مجموعه داده ای، مشاهداتی که مابین دیواره های داخلی و خارجی قرار می گیرند، مشکوک به پرت هستند. مشاهداتی که بیرون از دیواره های خارجی قرار می گیرند، جزء داده های پرت هستند. برای داده های بزرگ نمودار جعبه ای با استفاده از برنامه کامپیوتری ساخته می شود. نمودار پراکنش از شیوه های ساده دیگر برای تعیین داده های پرت است. بدین صورت که بعد از رسم نمودار پراکنش بین دو متغیر وابسته Y و متغیر مستقل X ، اگر داده های پرت باشد روی خط برآذش شده برای داده ها قرار نگرفته و در فاصله دورتری از خط قرار می گیرد. لذا خط رگرسیونی رسم شده برای داده ها با احتساب مقدار داده پرت دارای برآذش خوب نبوده و بدون در نظر گرفتن داده پرت برآذش مدل دقیق تر خواهد شد. آمار دانان راه های مختلفی را برای تعیین داده های پرت بیان کرده اند. آزمون گرای یکی از شیوه های ساده در این زمینه است که توسط گرای در سال 1969^{6-8} بیان گردید. آزمون گرای دارای مقدار آماره ای است که بر اساس اطلاعات موجود در نرم افزارهای مربوطه قابل محاسبه می باشد لذا بعد از محاسبه مقدار آماره آزمون

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

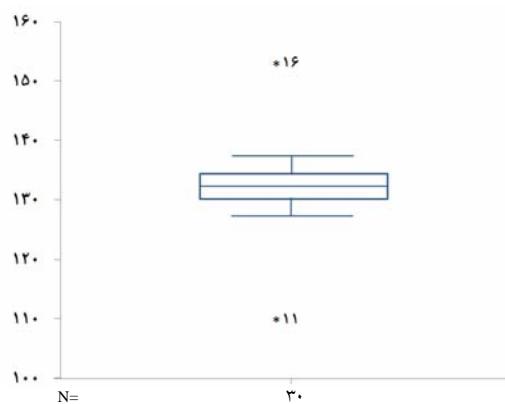
خروجی آزمون		آماره های توصیفی		تعداد مشاهدات
معنی داری نقطه پرت	مقدار استاندارد شده	مقدار مشاهده شده	۲۸	
-	۰/۳۸	۱۳۲		
-	۰/۳	۱۳۵	۱۳۳/۶۸	میانگین
-	۰/۰۷	۱۲۴	۴/۳۷	انحراف معیار
-	۰/۶۱	۱۲۱		
$p < 0/05$	۴/۴۲	۱۵۳	۰/۰۵	سطح معنی داری آزمون
-	۰/۶۱	۱۲۱		
-	۰/۵۳	۱۳۶	۲/۸۸	آماره آزمون

بحث

همان طور 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 به نوعی 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 داده‌ای مشکوک در بین داده‌های جمع‌آوری شده، وضعیت آنها را (به لحاظ پرت بودن) با تکنیک‌های مورد بحث بررسی و در صورت شناسایی داده‌ای 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 متمن 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 گраб با در اختیار گذاشتن سطح معنی‌داری هستند که با نتایج سایر روش‌ها مشابه می‌باشند.

References

1. What are outliers in the data? 2006; [<http://www.itl.nist.gov/div898/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, Juhola M, Kentala E. Informal identification of outliers in medical data. [http://ai.ijs.si/Branax/IDMAP-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/aggarwal01outlier.html>]. Available from: URL.



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: $IQR = 135 - 131 = 4$, $Q_3 = 135$, $Q_1 = 133$, $n = 30$. پس با احتساب مقادیر فوق اگر مشاهدات از اعداد ۱۲۵ کمتر و 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.

Detection of Outliers methods in medical studies

Babaei 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

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.

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