داده های دور افتاده

استاد راهنما : دکتر وحید رضایی تبار

ارائه دهنده : حسام افشار

فهرست مطالب

- مقدمه
- Q-Q Plot
- Box-Plot
- رگرسیون چندکی

مقدمه

- نقاط پرت مقادیری هستند که با سایر نقاط داده در یک مجموعه داده تفاوت شدیدی دارند. آنها می توانند تأثیر زیادی بر تجزیه و تحلیل های آماری شما داشته باشند و نتایج هر آزمون فرضیه را منحرف کنند. مهم است که به دقت نقاط پرت احتمالی را در مجموعه داده خود شناسایی کنید و برای نتایج دقیق با آنها به شیوه ای مناسب برخورد کنید.
- برخی از نقاط پرت نشان دهنده مقادیر واقعی از تغییرات طبیعی در جمعیت است. سایر نقاط پرت ممکن است ناشی از ورود نادرست داده ها، خرابی تجهیزات، یا سایر خطاهای اندازه گیری باشد.
- داده‌های پرت همیشه نوعی داده کثیف یا نادرست نیست، بنابراین باید در پاکسازی داده‌ها مراقب آن‌ها باشید.

مقدمه

- در مواجه شدن با داده های دور افتاده ای که ناشی از ورود نادرست داده ها، خرابی تجهیزات، یا سایر خطاهای اندازه گیری است، میتوان با آنها مانند داده های از دست رفته برخورد کرد و از آن روش ها برای تخمین آنها استفاده کرد چون این داده ها به نوعی داده کثیف یا از دست رفته هستند.
- اما داده های دور افتاده ای که نشان دهنده مقادیر واقعی از تغییرات طبیعی در جمعیت است مهم هستند و باید با روش هایی مناسب آنها را شناسایی و تحلیل نماییم که در ادامه روش هایی برای این منظور ارائه شده است.

Q-Q PLOT

Q-Q plot یک نمودار پراکندگی است که با ترسیم دو مجموعه چندک در برابر یکدیگر ایجاد می شود.برای این منظور داده ها را به صورت صعودی مرتب کرده سپس توزیع مد نظر را به تعداد n+1 قسمت تقسیم کرده (که n تعداد داده ها است) و انتهای هر قسمت بدست آمده را به داده متناظرش نسبت میدهیم و نمودار پراکندگی آنها را رسم میکنیم. اگر هر دو مجموعه چندک از توزیع یکسانی حاصل شده باشند، باید نقاطی را ببینیم که حول خطی تقریباً مستقیم با زاویه 45 درجه قرار گرفته اند.

Q-Q PLOT

دلیل در نظر گرفتن خط با زاویه 45 هم این است که اگر داده ها از یک توزیع آمده باشند پس چندک مثلا 40 درصد هر دو آنها باید متناظر با یکدیگر باشند.

حال اگر نقطه ای با فاصله زیاد از بقیه مجموعه داده ها قرار گیرد (یعنی در ابتدا یا انتهای خط و با فاصله ای زیاد از بقیه داده ها) میتواند نشانگر داده دور افتاده باشد.

فرض کنید یک مجموعه داده از توزیع نرمال داشته باشیم و نمودار Q-Q plot آن را در برابر توزیع نرمال رسم کنیم. میدانیم احتمال رخداد در چندک های ابتدایی و انتهایی توزیع نرمال بسیار کم است.

خط نمودار Q-Q plot هم نشانگر چندک ها بود پس ابتدا و انتهای این خط همان چندک های ابتدایی یا انتهایی توزیع نرمال هستند که احتمال رخداد کمی در توزیع نرمال دارند در نتیجه این داده ها میتوانند به عنوان داده دور افتاده در نظر گرفته بشوند.

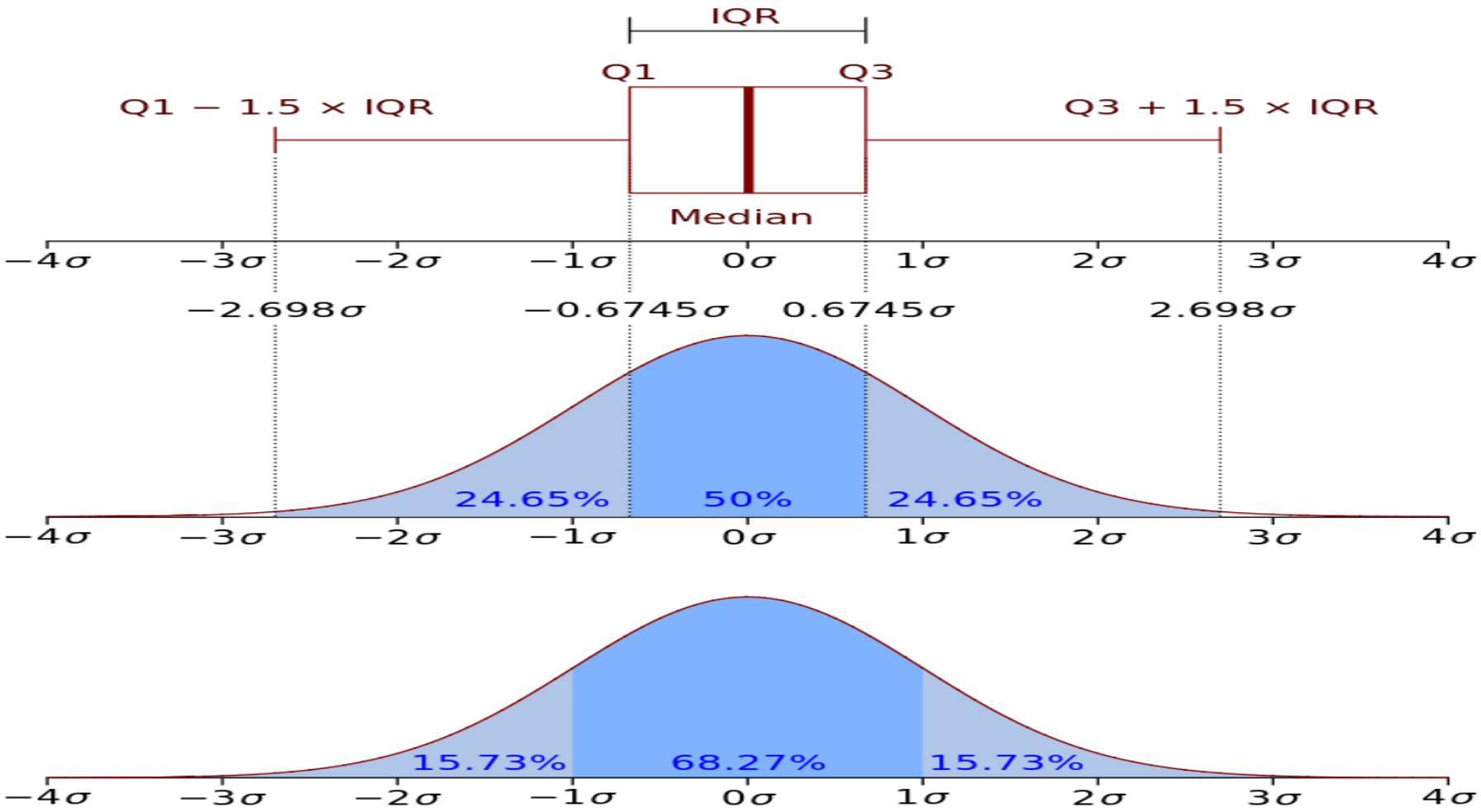
BOX-PLOT

نمودار جعبه ای یا همان Box-Plot یک روش برای نمایش داده ها است که آماردانان را قادر می سازد تا یک بررسی گرافیکی سریع روی یک یا چند مجموعه داده انجام دهند. نمودارهای جعبه همچنین فضای کمتری را اشغال می کنند و بنابراین برای مقایسه توزیع بین چندین گروه یا مجموعه داده به صورت موازی مفید هستند.

نمودار جعبه ای اطلاعاتی در مورد گستردگی و چولگی داده های عدد را از طریق چارک های آنها نمایش میدهد. علاوه بر جعبه در نمودار جعبه، خطوطی در بالا و پایین جعبه وجود دارد که نشان دهنده تغییرپذیری در خارج از چارک های بالایی و پایینی (چارک اول و سوم) است.

نمودارهای جعبه ناپارامتریک هستند. آنها تنوع در نمونه های یک جامعه آماری را بدون هیچ گونه فرضی در مورد توزیع آماری زیربنایی نشان می دهند

BOX-PLOT


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%

BOX-PLOT

معمولا بعد از رسم نمودار جعبه ای داده هایی که بیرون از خطوط افقی که یعنی محدوده بیرون از IQR $\pm 1.5 \times IQR$ هستند به عنوان داده دور افتاده در نظر گرفته میشوند. دلیل در نظر گرفتن عدد 1.5 در محاسبه این محدوده این است که همانطور که در شکل صفحه قبل مشاهده شد اگر مثلا توزیع گاوسی باشد ای طبق روش IQR داده هایی که خارج از محدوده $\mu \pm 2.698\sigma$ باید به عنوان داده دور افتاده در نظر گرفته شوند. در واقع با اینکار داده هایی را که احتمال رخداد آنها بسیار کم است را به عنوان داده دور افتاده در نظر میگیریم.

این روش در مواردی بسیار مفید است که مقادیر کمی در انتهای مجموعه داده خود داشته باشید، اما مطمئن نیستید که آیا هر یک از آنها ممکن است به عنوان مقادیر پرت محسوب شوند.

رگرسیون چندکی

همانطور که قبلاً اشاره کردیم داده های پرت بسیار مهم هستند و به راحتی نمیتوان از آنها صرف نظر کرد.

یکی از مزایای استفاده از رگرسیون چندکی نسبت به روش معمول رگرسیون کمترین مربعات(OLS) ، پایداری در مقابل داده های دورافتاده است زیرا در رگرسیون چندکی، به جای محاسبه میانگین شرطی متغیر پاسخ، از میانه یا چندک‌های شرطی متغیر پاسخ استفاده می‌شود.

دیگر مزیت رگرسیون چندکی این است که هیچ فرضی در مورد توزیع متغیر هدف ایجاد نمی کند.

رگرسیون چندکی

فرض کنید Y یک متغیر تصادفی با تابع توزیع تجمعی $F_Y(y)=P(Y \le y)$ است. چندک τام متغیر Y به صورت زیر تعریف می‌شود:

$$Q_Y(\tau) = F_Y^{-1}(\tau) = \inf\{ y : F_Y(y) \ge \tau\}$$

در اینجا τ مقداری بین 0 و 1 در نظر گرفته می‌شود. به این ترتیب مشخص است که مثلا منظور از چندک 0.1، کوچکترین مقدار از مقادیر Y است که مقدار تابع توزیع تجمعی بزرگتر از 0.1 است. برای پیدا کردن چندک τام از روشی که در ادامه معرفی می‌شود استفاده خواهیم کرد. تابع زیان را به صورت زیر در نظر می‌گیریم:

$$\rho_\tau(y) = y\,(\tau - \mathbb{I}_{(y<0)})$$

رگرسیون چندکی

این ترتیب برای پیدا کردن چندک، از کمینه‌سازی امید ریاضی Y−u نسبت به u استفاده می‌کنیم. بنابراین خواهیم داشت :

$$\min_u E(\rho_\tau(Y-u)) = \min_u \left\{ (\tau-1)\int_{-\infty}^{u}(y-u)dF_Y(y) + \tau\int_{u}^{\infty}(y-u)dF_Y(y) \right\}$$

حال اگر تابع توزیع تجمعی را نداشته باشیم و فقط n مونه از توزیع متغیر داشته باشیم آنگاه چندک متغیر را با بهینه‌سازی پایین می‌توان به‌دست‌آورد.

$$\hat{Q}_Y(\tau) = \arg\min_{u\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i-u) = \arg\min_{u\in\mathbb{R}} \left[(\tau-1)\sum_{y_i<u}(y_i-u) + \tau\sum_{y_i\geq u}(y_i-u) \right]$$

رگرسیون چندکی

با مفهومی که از چندک و تابع زیان درک کردیم، حالا می‌توانیم به چندک شرطی و رگرسیون چندکی بپردازیم. فرض کنید چندک شرطی Y نسبت به متغیر X را به صورت $Q_{Y|X}(\tau)$ نشان داده‌ایم. به کمک این رابطه رگرسیون یا مدل خطی رگرسیون چندکی را به شکل زیر بیان می‌کنیم:

$$Q_{Y|X}(\tau) = X\beta_\tau$$

به منظور برآورد پارامترهای این مدل خطی کافی است که تابع زیان معرفی شده را برحسب β کمینه کنیم. بیان ریاضی این مسئله را به صورت زیر می‌نویسیم.

$$\beta_\tau = arg \min_{\beta \epsilon R^k} E(\rho_\tau(Y - X\beta))$$

رگرسیون چندکی

حال این معادله منجر به برآورد پارامترهای β به صورت زیر خواهد شد:

$$\widehat{\beta_\tau} = arg \min_{\beta \epsilon R^k} \sum_{i=1}^{n} \rho_\tau (Y_i - X_i \beta)$$

رگرسیون چندکی

نمودار ترسیم شده به بررسی رابطه میزان درآمد و میزان سرانه خوراک پرداخته است. البته مشخص است که خطوط منقطع همان رگرسیون چندکی و خط قرمز رنگ نیز رگرسیون خطی ساده (OLS) است. داده ها نیز به صورت دایره‌های آبی رنگ در نمودار دیده می‌شوند.