تحديد القيم المبتعدة او الشاذة Detecting outliers

يعد تحديد القيم الشاذة مهما لاسباب عدة، اولها تمييز التسجيل الخاطئ في النتائج . ان وجود قيمة شاذة واحدة يؤدي الى تاثير كبير في قيمة الوسط الحسابي والتغايير. ويمكن معرفة القيم الشاذة من خلال فحص البيانات، او الاطلاع للمخطط البياني لتوزيع البيانات. هناك طريقتين لفحص البيانات:-

1. A classic outlier detection method

A classic outlier detection technique illustrates the problem of masking. This classic technique declares the value X an outlier if $\frac{|x-\bar{x}|}{S} \geq 2$

حيث ان X = قيمة المتغير ، $\bar{x}$ = الوسط الحسابي ، S = الانحراف المعياري

Example 1

Consider the values 2,2,2,2,2,3,3,3,3,3,4,4,4,4,4,1000.

The sample mean is $\bar{x}$= 65.94, the sample standard deviation is s = 249.1,

$$\frac{|1000-\overline{65.94}|}{249.1} = 3.75 \geq 2$$

القيمة 3.75 هي اكبر من 2 ، لذلك فان القيمة 1000 تعد قيمة شاذة، احيانا هذه الطريقة لا تكون كافية لتحديد القيمة الشاذة.

2. The boxplot rule

هذه طريقة محسنة بالمقارنة مع الطريقة الكلاسيكية الاولى في تحديد القيم الشاذة، وذلك بتحديد الربعيات Q1 , Q3 ومقارنة قيمة المتغير X بحسب المعادلات الاتية :-

$X < Q1 - 1.5(Q3-Q1)$
or
$X > Q3 + 1.5(Q3 - Q1)$.

مثال:-

1,2,3,4,5,6,7,8,9,10,11,12,13,14,100,500.

lower quartile is $q1$ =4.417, the upper quartile is $q3$ = 12.583, so $q3 + 1.5(q3-q1) = 12.583+1.5(12.583-4.417) = 24.83$.That is, any value greater than 24.83 is declared an outlier. In particular, the values 100 and 500 are labeled outliers.

مقاييس التغاير

التغاير variance : هو مقياس للتغير في البيانات وهو يمثل مربع الانحراف المعياري. يحسب التغايير للمجتمع وفق المعادلة ادناه:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}$$

حيث ان μ = الوسط الحسابي للمجتمع

ويتم حساب التغاير للعينة وفق المعادلة ادناه :-

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

حيث ان $\bar{x}$ = الوسط الحسابي العينة

الانحراف المعياري standard deviation

هو مقياس للتغاير في البيانات حول الوسط الحسابي، ويقاس عن طريق حساب الجذر التربيعي للتغاير. كلما كبرت قيمته تعني وجود تغاير كبير في البيانات وحودات القياس له هي نفس وحدات المتغيرات الاصلية ، وهو يتاثر بالقيم الشاذة outliers.

- الانحراف المعياري للمجتمع يحسب وفق الاتي :- $\sigma = \sqrt{\sigma^2}$

- الانحراف المعياري للعينة يحسب وفق الاتي :- $s = \sqrt{s^2}$

الدرجة المعيارية Z- Score

الدرجة المعيارية تمثل عدد الانحرافات المعيارية لقيمة المتغير x التي تقع بعيدا عن الوسط الحسابي، ويتم حسابه وفق الاتي:-

$$z = \frac{x - \bar{x}}{s}$$

حيث ان $\bar{x}$ يمثل الوسط الحسابي و s= مثل الانحراف المعياري

وقيمة الدرجة المعيارية تكون موجبة او سالبة او صفرية. فاذا كانت سالبة تعني القيمة x هي اقل من الوسط الحسابي، والموجبة تعني قيمة x اعلى من الوسط الحسابي ، والصفر تعني قيمة المتغير x مساوية للوسط الحسابي.

وتفسير نتائجه يكون كلاتي ان قيمة المتغير x هي تنحرف عن المعدل (الوسط الحسابي) بمقدار z من الانحراف المعياري .

The z-score that corresponds to each speed is calculated below.

$x = 62$ mph: $z = \frac{62 - 56}{4} = 1.5$

$x = 47$ mph: $z = \frac{47 - 56}{4} = -2.25$

$x = 56$ mph: $z = \frac{56 - 56}{4} = 0$

في هذا المثال سرعة السيارة x=62 يعني انها تبتعد عن بمقدار 1.5 انحراف معياري عن المعدل لسرعة السيارات الثلاث وهكذا لبقية السيارات.