Data Processing

Edward Jex

April 14, 2020

Different Types of Data

- Categorical - Generally not numerical.
- Discrete - Only certain values, normally integers.
- Continuous - Numerical, real values.
- Ranked - Numerical, ordered.

Categorical data

For categorical data, the most common summary measure of our data is the modal class. This is the class with the highest frequency. Diagrams:

- Bar chart
- Pie chart
- Pictogram
- Pot chart

Ranked Data

If our data is ranked, we normally use stem and leaf diagrams or box plots to represent the data.

Stem and Leaf Diagrams

Example:

Key: 3|1 means 31

Stem	Leaf
1	9 9
2	0 4 7 8
3	1 2 2 2 6
4	0 5 5
5	5

Note: includes repeats, only a single digit on the right.

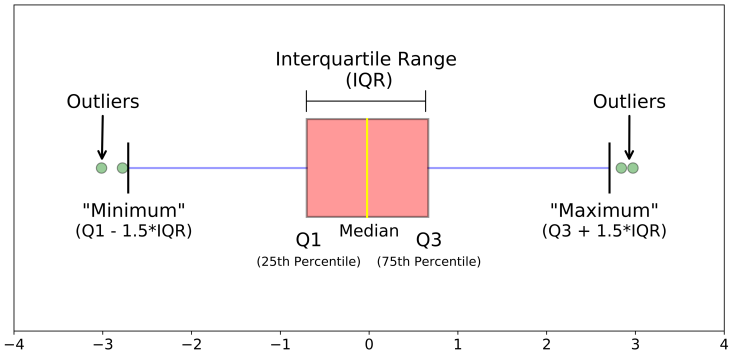
For ranked data, we often use the median, lower and upper quartiles as our summary measures.

- The median value (Q_2) - The middle number
- The lower quartile (Q_1) - The middle number of the lower half
- The upper quartile (Q_3) - The middle number of the upper half

If the number of data-points is off, you just take the middle number. If it is even, take the average between the two middle numbers and when calculating the quartiles, include the middle.

Box Plots

The five key numbers can be shown on a simple diagram known as a box-and-whisker plot.



Outliers

We can say that a data-point is an outlier if the data-point is more than 1.5× IQR beyond or below the lower or upper quartiles.

Product-Moment Correlation Coefficient

The PMCC measures hot close the data is to a straight line.
$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ where

$$\text{Always} > 0 \begin{cases} S_{xx} = \Sigma x^2 - n\bar{x}^2 \\ S_{yy} = \Sigma y^2 - n\bar{y}^2 \end{cases}$$

$$\text{Positive or negative} \begin{cases} S_{xy} = \Sigma xy - n\bar{x}\bar{y} \end{cases}$$

We are assuming that the underlying population has a bivariate normal distribution. If we show the data on a scatter graph it should form an ellipse.