Data Analysis

Core

Engage Education Foundation

Univariate data: one variable
Bivariate data: two variables

Numerical data: a quantity that is measured or counted
- discrete – fixed value, "how many?" e.g. 5 cars drive past
- continuous – don't necessarily fixed values, "how much?"

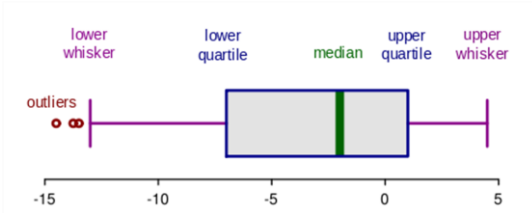
Categorical data: categories of values. e.g. eye colour
- nominal – no specific order in categories
- ordinal data – specific order of categories

Dependent variable: the variable being influenced
Independent variable: the variable that does the

Five point summary

Minimum, Q_1, median, Q_3, maximum

Boxplots



Q_1 = first 25% of values, Q_3 = last 25% of values

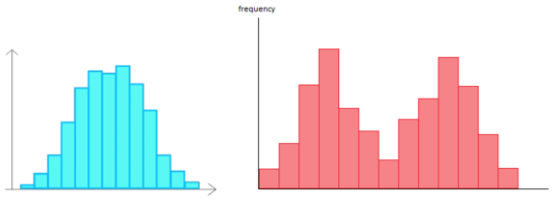
$$IQR = Q_3 - Q_1$$

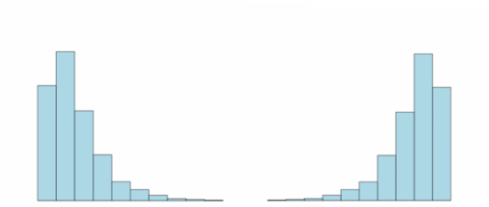
Outliers = less than $Q1 - 1.5 * IQR$ or more than $Q3 + 1.5 * IQR$

Describing the shape of graphical data

Shape

Symmetrical (both below, right one is called bimodal)





Positively skewed (above)

Negatively skewed (above)

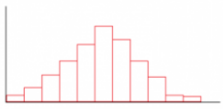
Centre

Three measures of centre that can be used to describe histogram—mean, median and mode. Inferred from a labelled diagram.

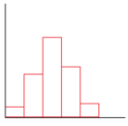
Spread

Spread of distribution is referred to as the maximum range of the distribution.

$$range = largest\ value - smallest\ value$$



Broad peak / wide spread



Narrow peak

mean: the 'average'. sum of all the data values divided by total number of data values
mode: the most frequently occurring value or category
median: the value in the data set that divides the data evenly into two. given *n* terms, the median is the (n+1)/2 th term

standard deviation: is a measure of spread.

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Normal distributions

The 68-95-99.7% rule

Least squares regression line

Fitting by eye

Three median regression line

$$y = a + bx$$

where b is the slope, given by $b = \frac{rs_y}{s_x}$

and a is the intercept, given by $a = \bar{y} - b\bar{x}$

Three line median regression line

1. Divide data into three groups. Data points in the outside groups must always be the same
2. Locate median of each group of points
3. Place ruler between left and right medians
4. Move 1/3rd of the way towards middle median
5. Find gradient, find *y*-intercept.

Residuals

residual value = actual value – predicted value

Residuals randomly scattered around the zero regression line indicates likely linear

Time series

- **trend**: increasing/decreasing
- seasonal: repetitive and evenly spaced within year
- **cycles**: long term, usually over more than a year
- **random**: not classified as any of the above

Seasonal index

$$\text{seasonal index} = \frac{\text{actual figure}}{\text{deasonalised figure}}$$

- seasonal indices have an *average* value of 1(or 100%)
- a seasonal index of 1.3 (or 130%) indicates that that season had 30% more than the seasonal index
- a seasonal index of 0.6 would indicate 40% less than the seasonal index

Data transformations

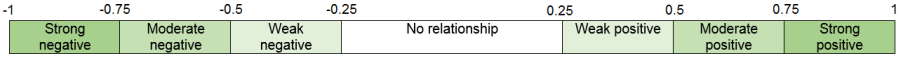
Refer to attachment

Scatterplots

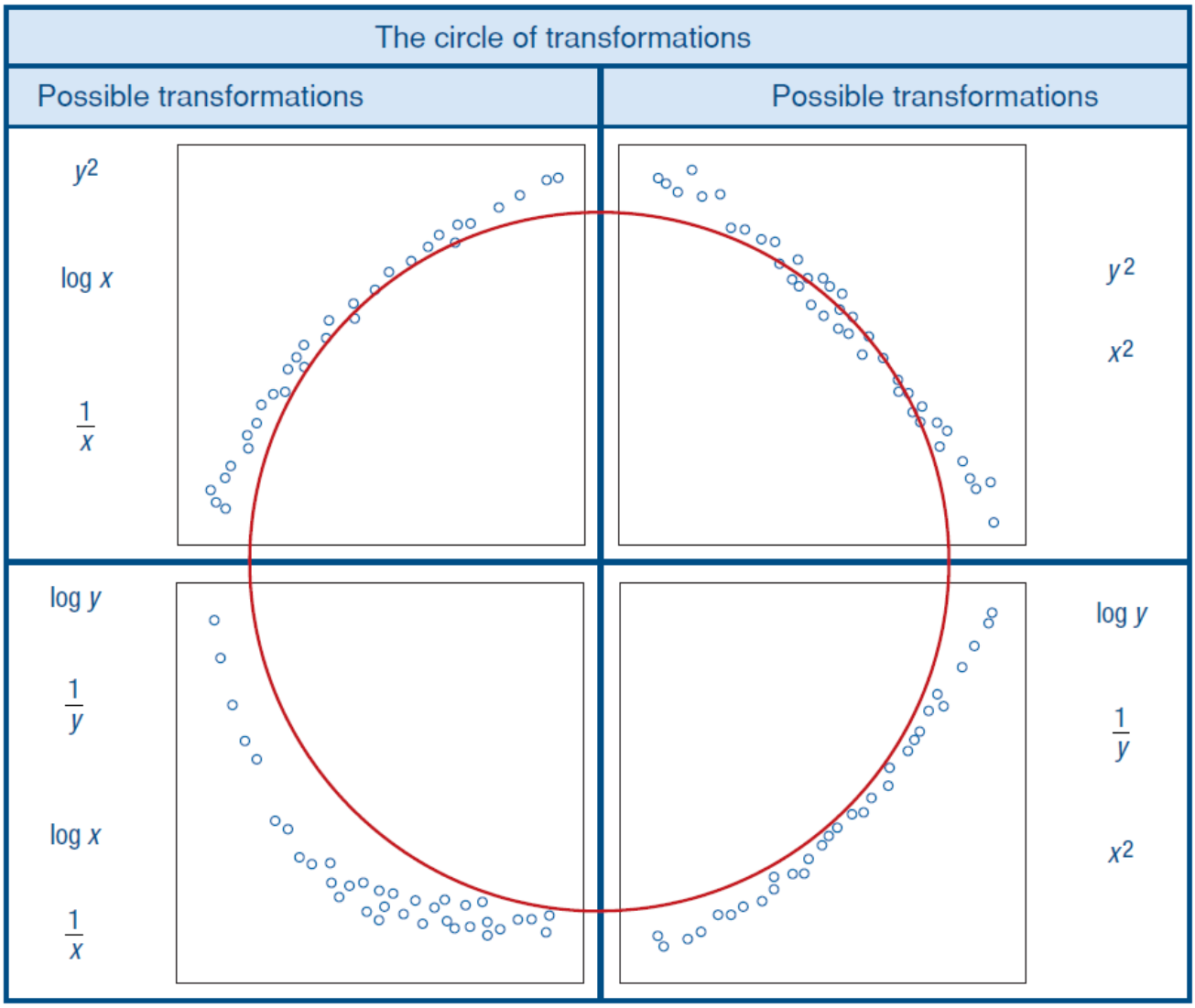
Direction: positive or negative

Form: linear or non-linear

Strength: the correlation coefficient, *r*



Data transformations:


The circle of transformations
Possible transformations
Possible transformations
y^2
$\log x$
$\frac{1}{x}$
y^2
x^2
$\log y$
$\frac{1}{y}$
$\log x$
$\frac{1}{x}$
$\log y$
$\frac{1}{y}$
x^2