Quantitative Reasoning

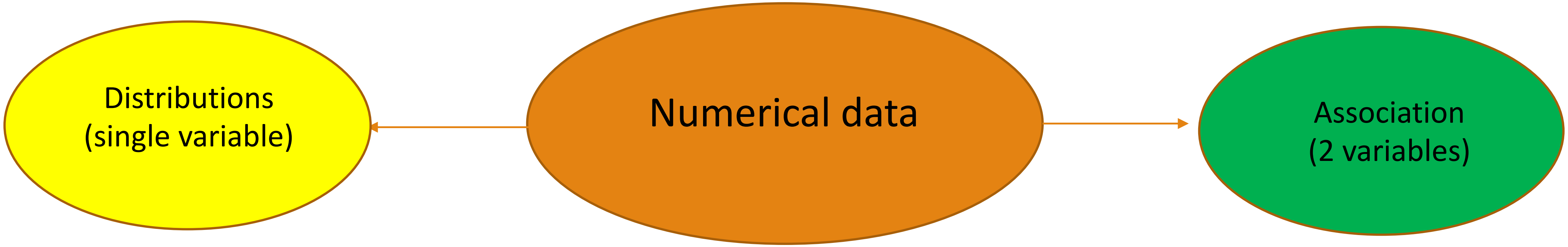
# Working with Numerical data

## KEY CONCEPTS

Kelvin Horia

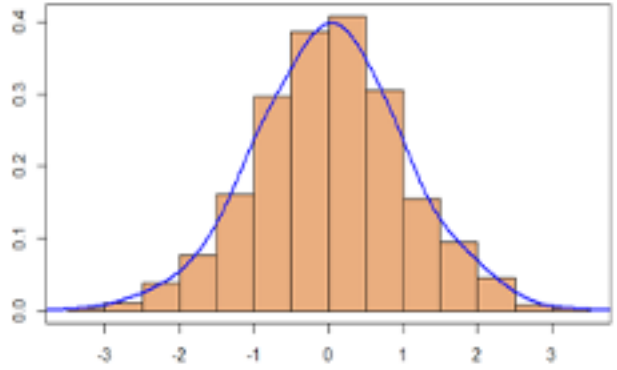
pvokh@nus.edu.sg
Teaching assistant
Provost office, QR unit

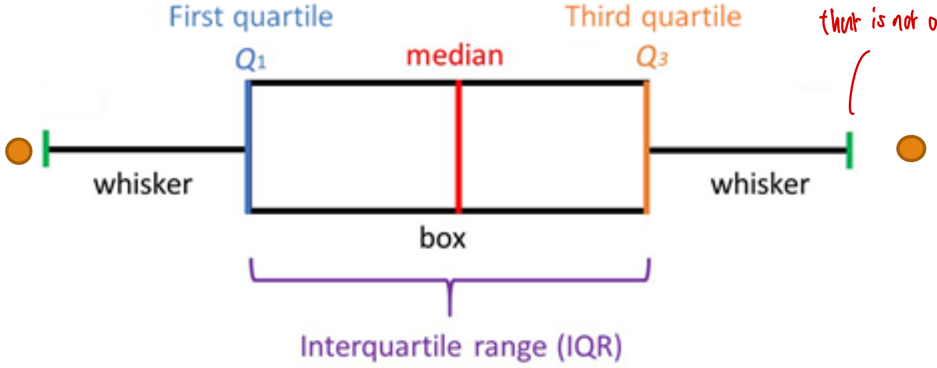
# Outline


Distributions
(single variable)

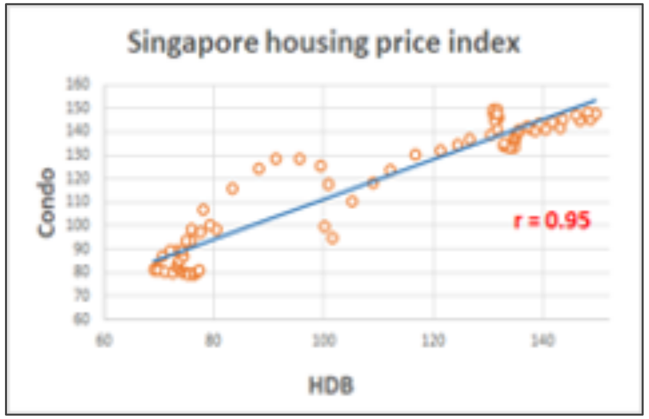
Numerical data

Association
(2 variables)



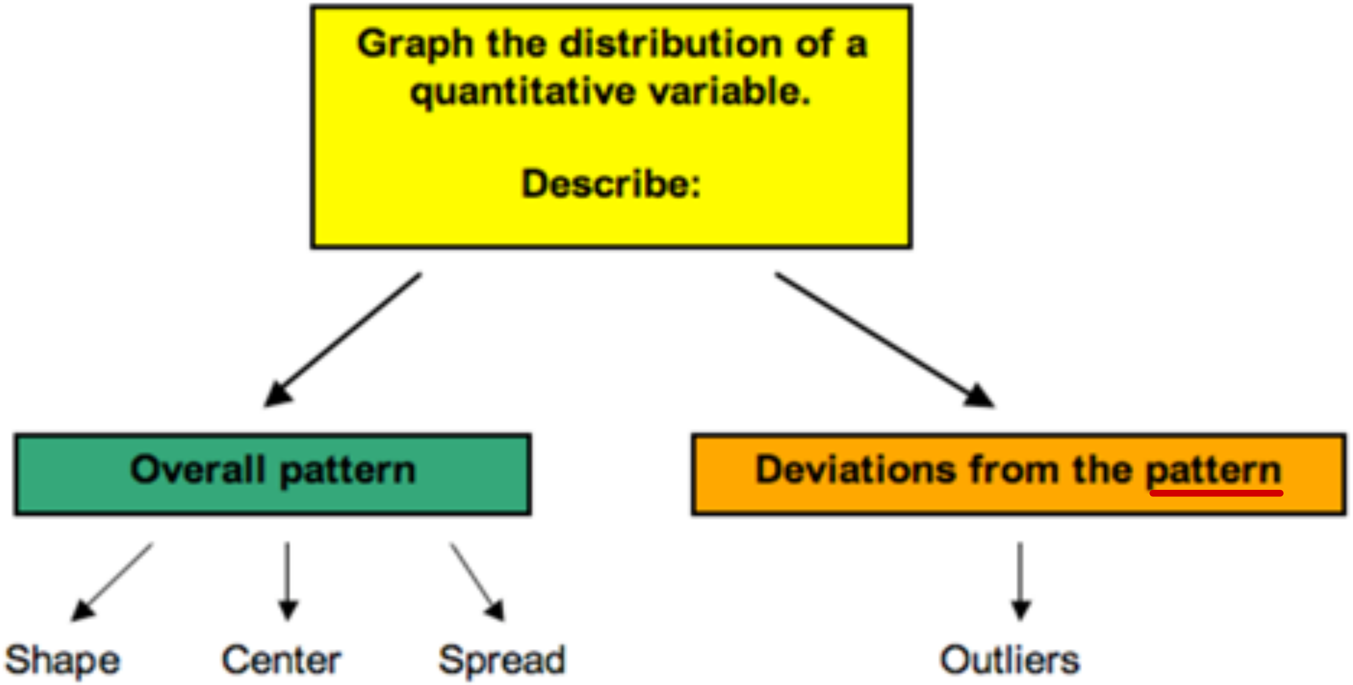


largest value
that is not outlier.



Outlier is defined as any point that lies above Q3 + 1.5 × IQR,
or below Q1 - 1.5 × IQR,

# How to describe distribution?



Graph the distribution of a quantitative variable.

Describe:

Overall pattern

Deviations from the pattern

Shape

Center

Spread

Outliers

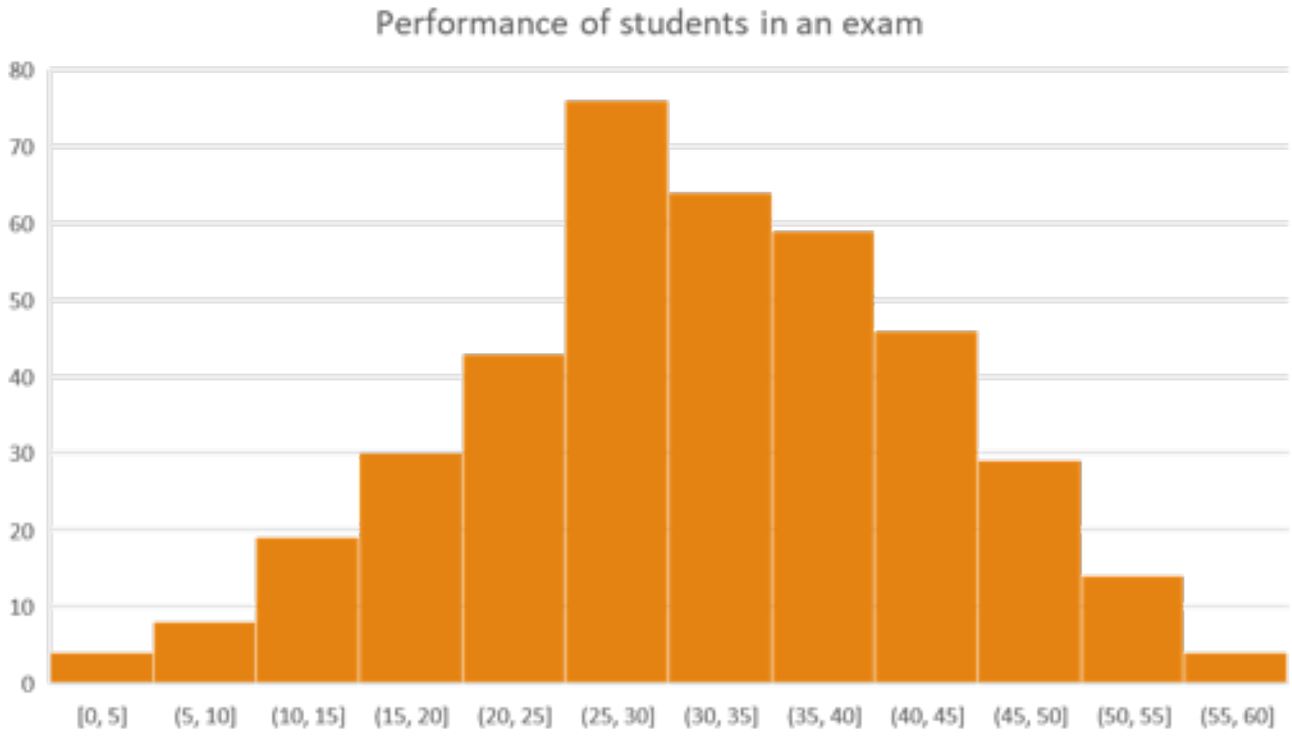
peaks
skewness

mean
median
mode

range
standard deviation
IQR

observations that fall well above or well below the overall bulk of data

# Getting information out of histograms

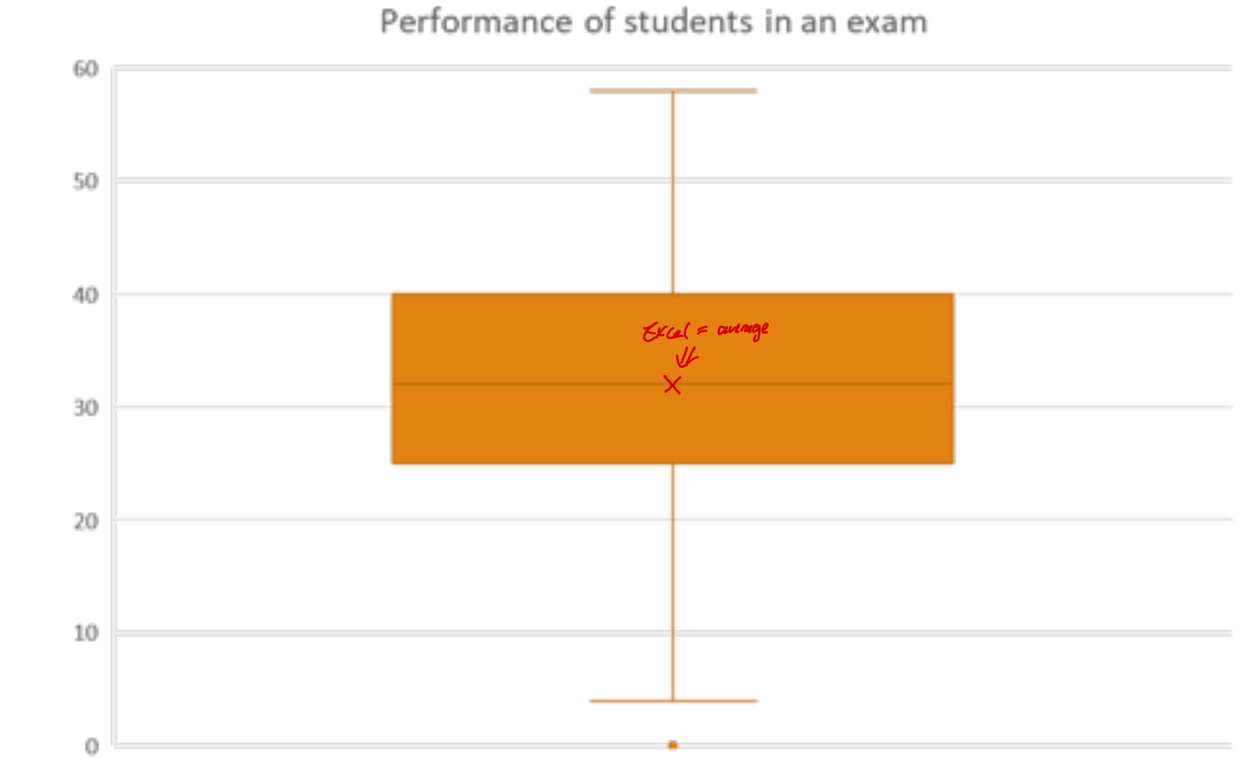


Histogram showing the distribution of scores for an exam taken by students in a school. The maximum mark for the exam is 60.

- ❑ What can you say about the performance of the students in the school based on the histogram?
  - ❑ Failure rate ✓
  - ❑ Borderline failures ✓
  - ❑ Distinction rate (those who scored 75% and above) ✓

- ❑ Is it possible to determine the following information based on the histogram?
  - ❑ Mean ✗ → score of every single student
  - ❑ Median ✗
  - ❑ Standard deviation ✗
  - ❑ Q1 and Q3 ✗
  - ❑ Outliers ✗

# What box-plots can tell us



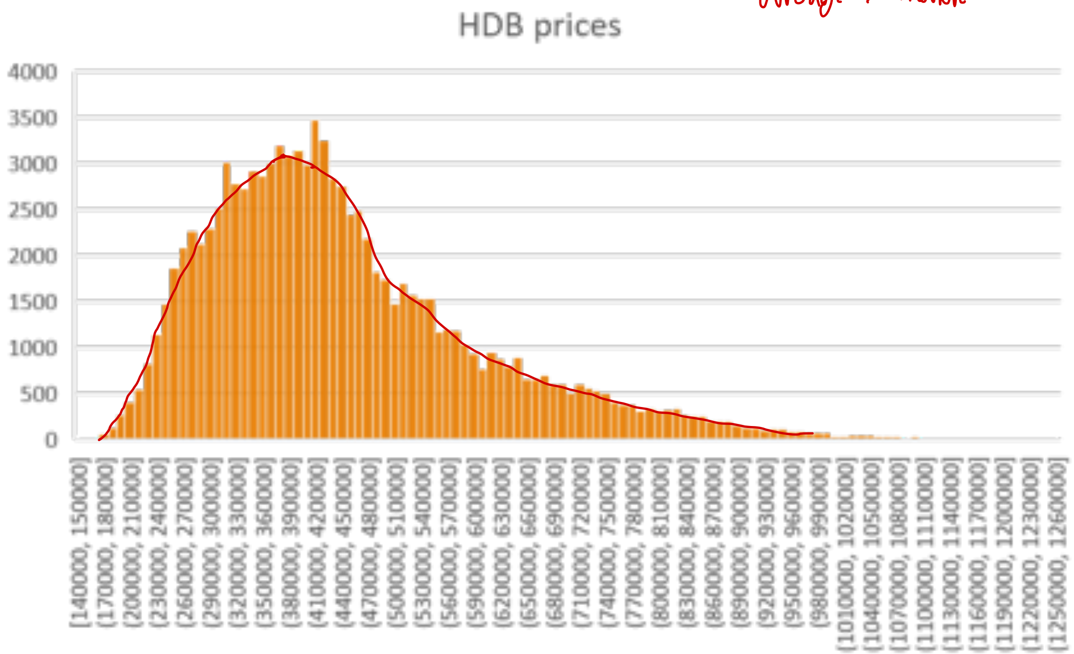
The same data shown as a histogram in the previous slide is now plotted using a boxplot.

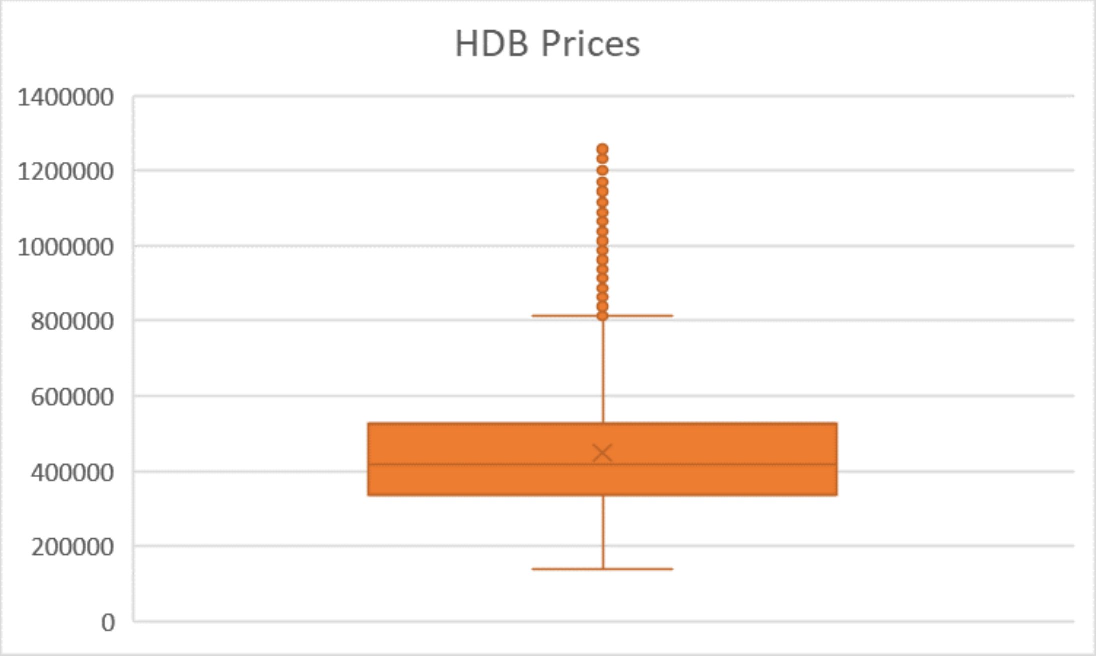
- □ What can you say about the performance of the students in the school based on the boxplot?
  - □ Failure rate ✗
  - □ Borderline failures ✗
  - □ Distinction rate (those who scored 75% and above) ✗

- □ Is it possible to determine the following information based on the boxplot?
  - □ Mean ✓
  - □ Median ✓
  - □ Standard deviation ✗
  - □ Q1 and Q3 ✓
  - □ Outliers ✓

# Outliers and skewness

average > median.



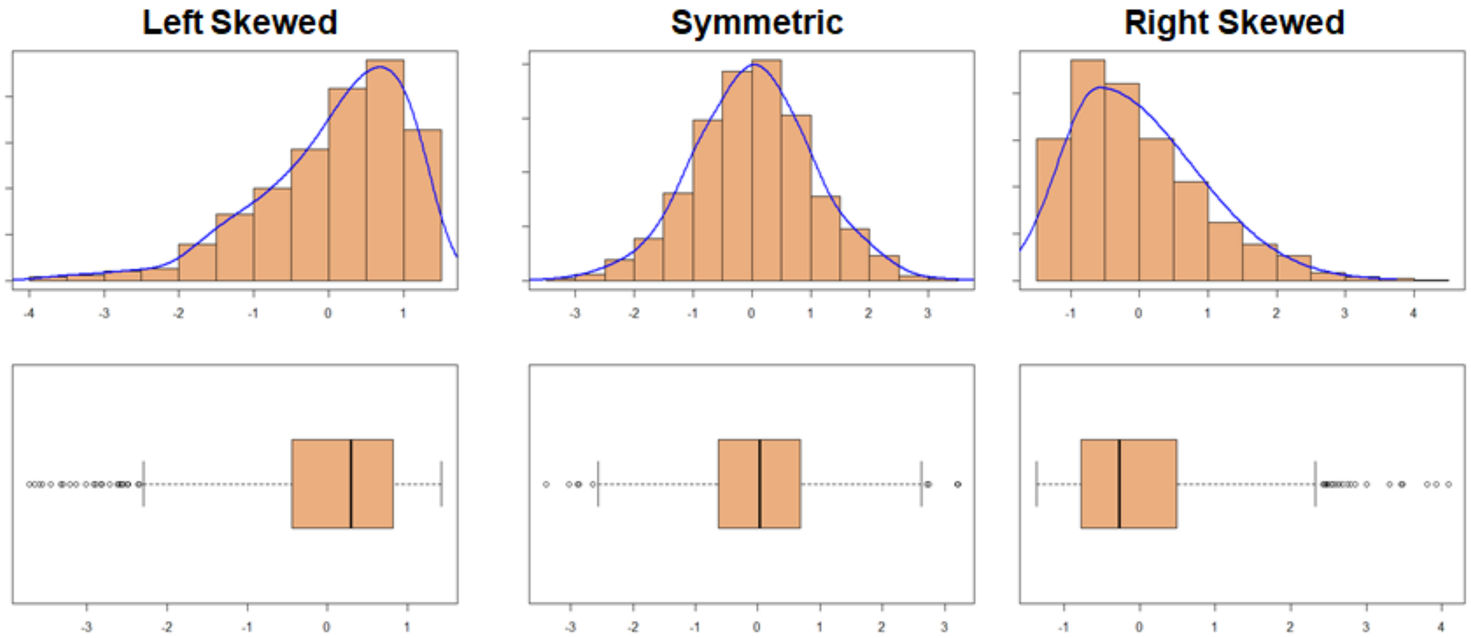


- ❑ Suppose the mean HDB price is $496 870 whilst the median HDB price is $468 000. Why do the mean and median differ significantly?

→ outliers affect average significantly, but don't really affect median.

- ❑ Why do you think they prefer to talk about median HDB price as compared to mean HDB price?

# Outliers, skewness and robust statistics



- □ How can outliers affect summary statistics such as
  - □ Mean
  - □ Median
  - □ Standard deviation
  - □ IQR

□ How do means and medians compare in the above distributions?

mean < median < mode
(in general)

mode ≈ median ≈ mean.

mode < median < mean.
(in general)

frequency/
distribution/mode.

distribution of different datasets.
to identify outliers.

# Histogram or Boxplot

- Histogram typically gives a better sense of the shape of the distribution of a variable compared to a boxplot.

- If we wish to compare the distributions of different data sets, putting the different boxplots side by side is more illustrative.

- Boxplot is better if we want to identify outliers

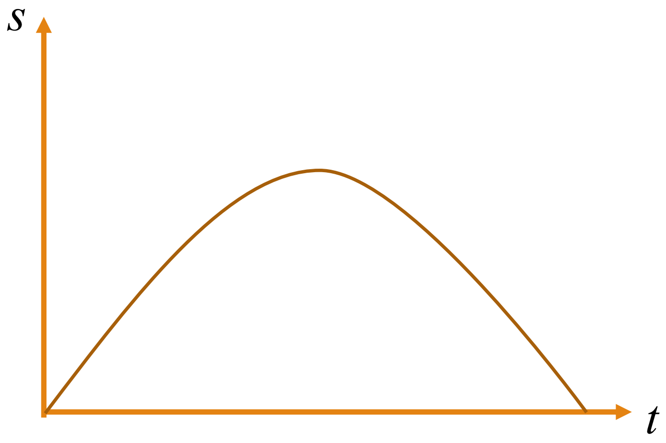
# Deterministic relationships (2 variables)

- ❑ A formula, for which given the value of one of the variables, you can calculate a **true** value for the other variable. E.g degrees Celsius to Fahrenheit.

$$T^0F = \frac{9}{5}(T^0C) + 32$$

- ❑ Formula for calculating height of a stone thrown in the air (assuming ideal conditions)

$$s = ut + \frac{1}{2}at^2$$

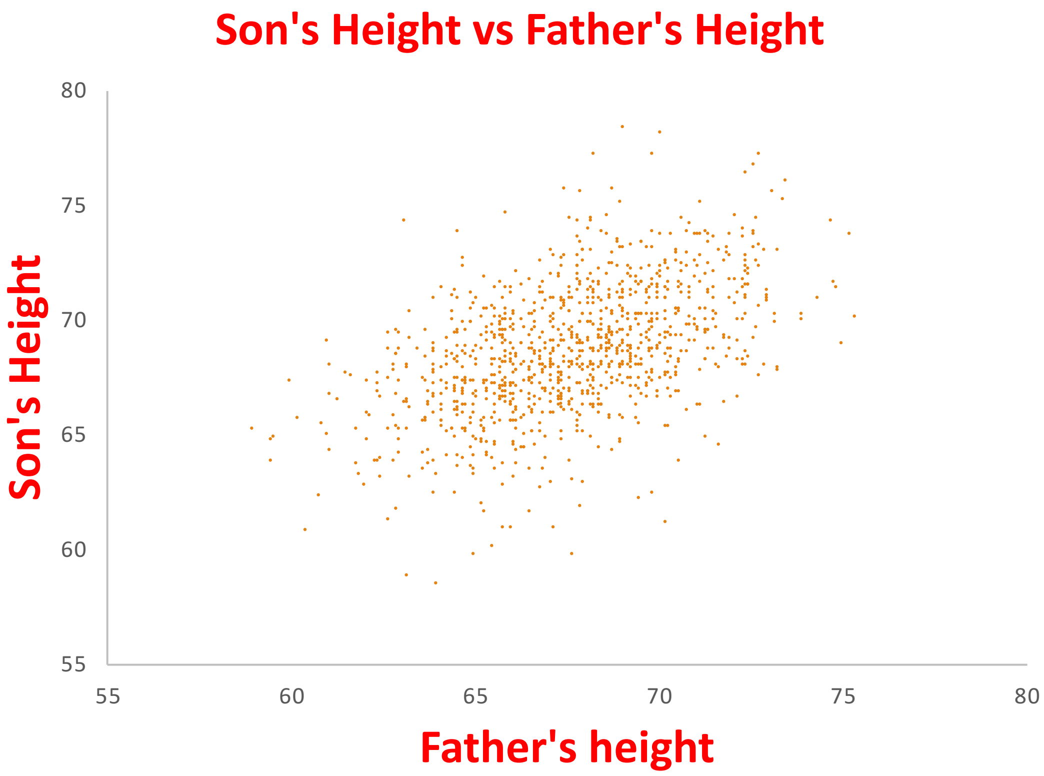


# Non-Deterministic relationships (2 variables)

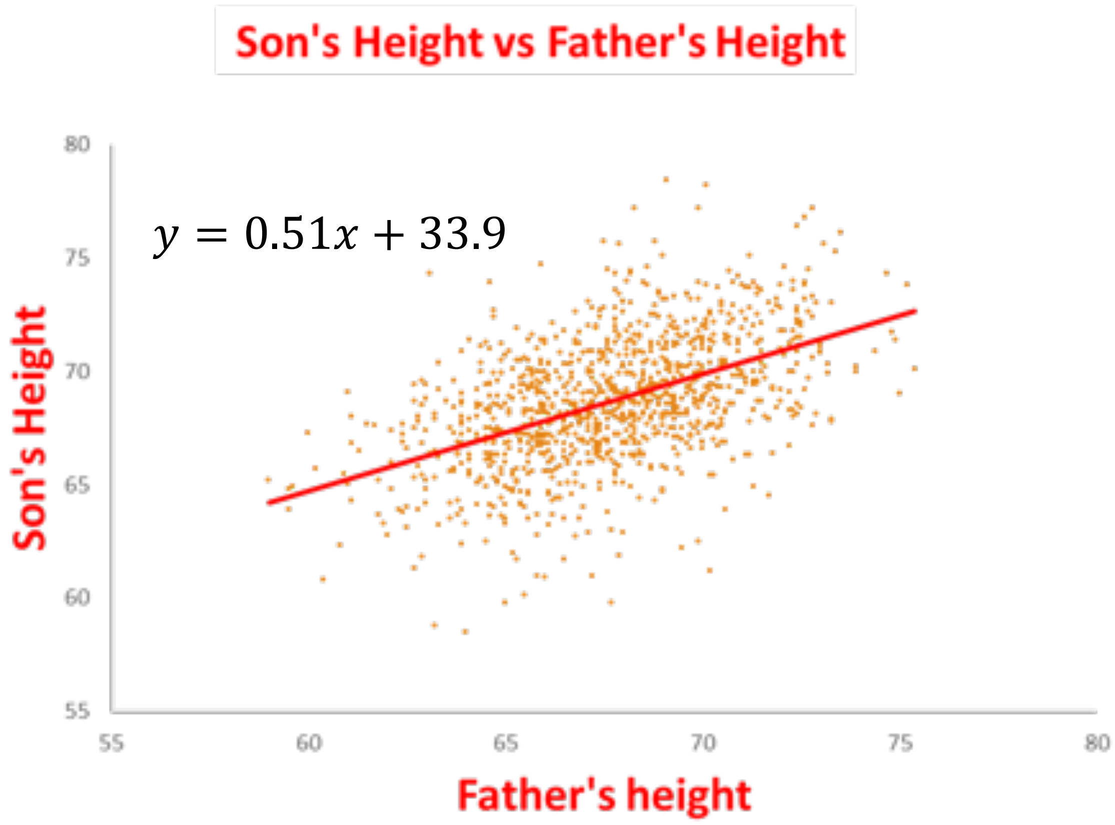
- ❑ The relationship between numerical variables **cannot** be codified into a formula which gives us **true** values.

- ❑ We can only collect data and try to model the relationship and **we use the scatterplot to describe if there's any association between the 2 variables**.

- ❑ The scatter plot/model can comes with a "formula", but we need to be careful as to what we can and cannot do with that formula.



# Simple linear regression



❑ Suppose we wish to investigate the relationship between a father and his son's height. Let $x$ denote the father's height (independent variable) and $y$ denote the son's height (dependent variable).

❑ Based on the scatter plot, is there an association between the 2 variables?

❑ What is the predicted son's height for a father whose height is 67 inches? What about a father whose height is 80 inches?

❑ What are the limitations of our "formula" as compared to a deterministic relationship?

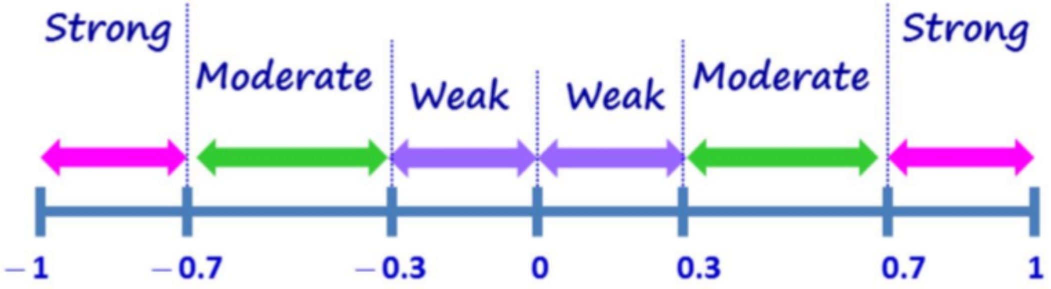
# Correlation coefficient *r*.

The correlation coefficient is a way to quantify the degree of **linear relationship** (which is one type of association) between numerical variables.

- measures ***linear*** association between 2 variables (NOT causation!)
- ranges between **-1 and 1** (no units)
- $r > 0 \rightarrow$ **positive** *linear* association

  $r < 0 \rightarrow$ **negative** *linear* association

  $r = 0 \rightarrow$ **no** *linear* association

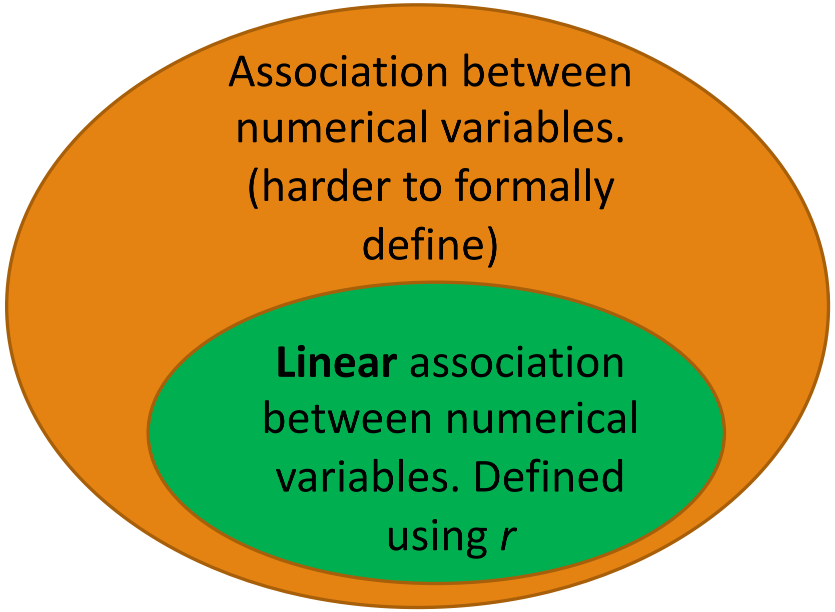


# Correlation coefficient and gradient

- $r$ is not affected by the following
  - Adding and subtracting constants to either variable
  - Multiplying and dividing *positive* constants to either variable.
  - Interchanging the *x* and *y* axis.

- $r \neq m$ in general.
  - $\underbrace{r = m \times \frac{s_x}{s_y}}$

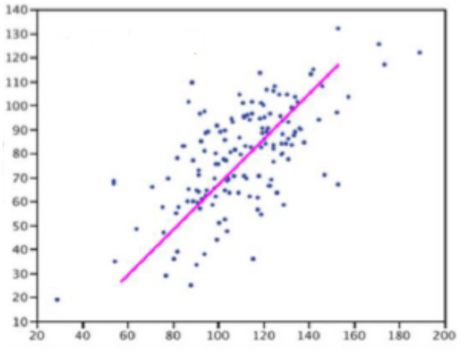
Using this you can also figure out when is *r* going to be the same as *m*



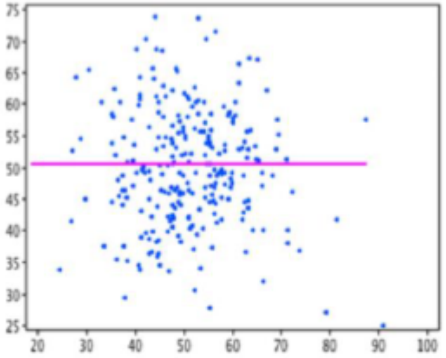
Association between numerical variables. (harder to formally define)

**Linear** association between numerical variables. Defined using *r*

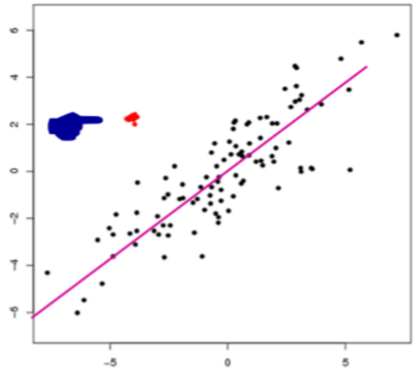
# *r* value alone doesn't tell you the whole story



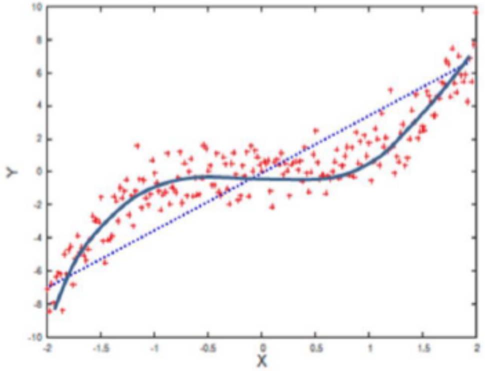
$r$ = 0.85 → might not be a straight line.

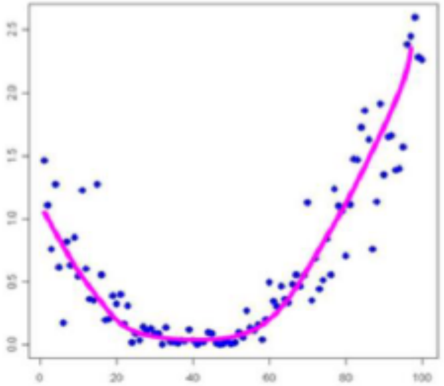


$r$ = 0 → no linear association.



Check scatter plot!

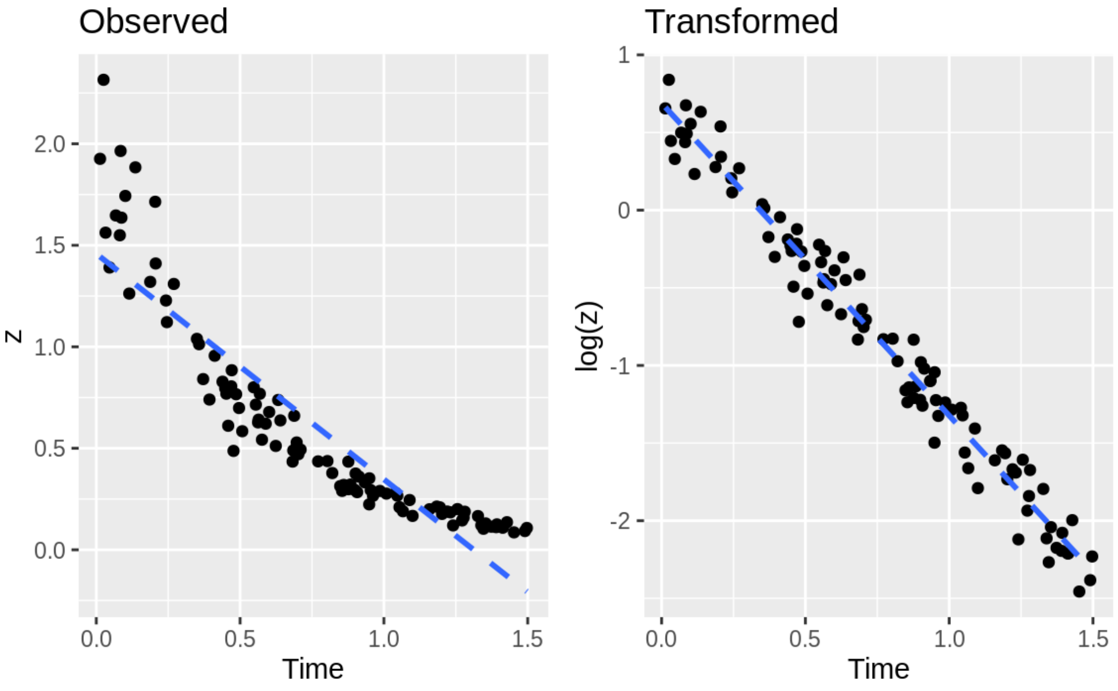




**Note:** When an outlier is removed, the $r$ value can increase, decrease or remain unchanged!

# Non-linear regression

- The goal is to use our understanding of **linear regression** to help us understand non-linear ones.



Observed

Transformed

**Exponential decay of a population of some organism:**

1. Model using

$$y = cb^t$$

2. Take log on both sides

$$\underbrace{\ln y = \ln c + t \ln b}$$

Convert to linear form

$$Y = mX + C$$

$Y = \ln y$, $m = \ln b$, $X = t$ and $C = \ln c$