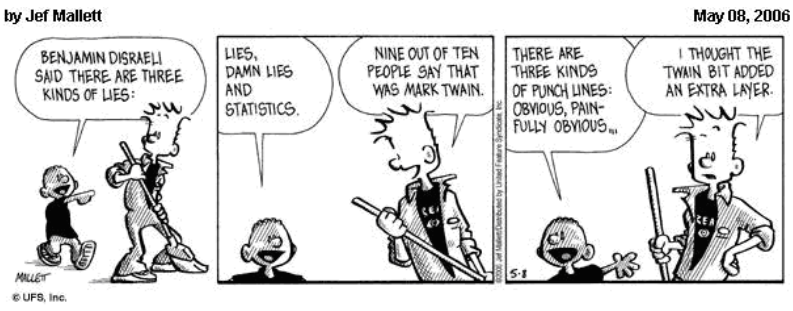
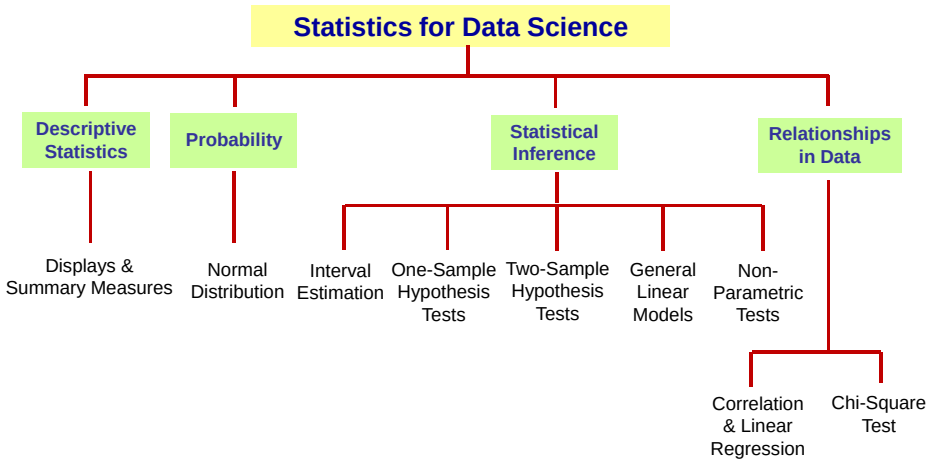
MATH 4044
Statistics for Data Science

Introduction to data

The role of Statistics …

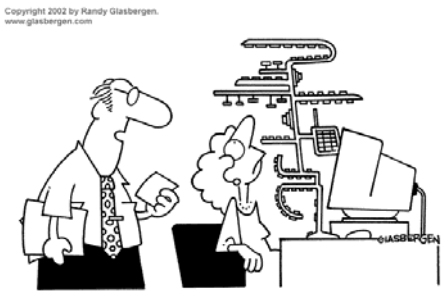


Course outline



Topics to be covered

- The scientific method
- Statistical thinking
- Collecting data
- Experiment types
- Types of Data



Field, A & Miles, J, *Discovering Statistics Using SAS*, Chapter 1

Types of data analysis

- Quantitative methods:
  - Testing theories using numbers
- Qualitative methods:
  - Testing theories using language
    - Magazine articles/interviews
    - Conversations
    - Newspapers
    - Media broadcasts

The scientific method

How, What, Why, When, Who, Which, Where … i.e. something you can measure, preferably with a number.

Ask a Question

Do Background Research

Research the literature to find the best way to do things and avoid repeating past mistakes.

An educated statement about what you believe to be true, in terms of your original question.

Construct a hypothesis

Test Hypothesis

Collect data according to a designed study (experiment or observational).

Analyse data and draw conclusions

"Weigh up the evidence" to see if hypothesis is true or false.

Report Results

What did you find? Was your hypothesis correct?

Statistical thinking

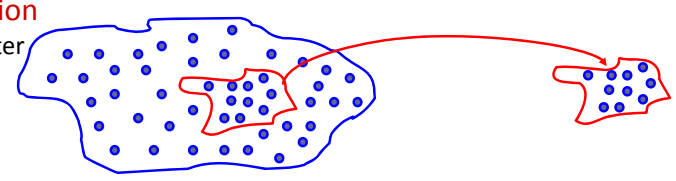
Data is *information with context*

Where do the data come from?	*Variable/Experiment types*
Always look at the data	*Numerical & graphical summaries*
Beware the lurking variable	*Correlation*
What can I conclude now/in the future?	*Inference from a sample/regression*
Variation is everywhere	*Inference, variability*

Population/Parameter *vs* Sample/Statistic

- Population: The entire group of people or objects about which information is wanted.
- Sample: The part of the population actually being examined, in order to gather information.

Population
Parameter



Sample
Statistic

- Has size N.
- A *parameter* is a numerical summary of the population.
- Has size n.
- A *statistic* is a numerical summary of the sample.

- Population Parameters are denoted with Greek letters and their equivalent Sample Statistics with English letters.

Validity

- Validity
  - Whether an instrument measures what it set out to measure

- Content (internal) validity
  - Evidence that the content of a test corresponds to the content of the construct it was designed to cover
- Ecological (external) validity
  - Evidence that the results of a study, experiment or test can be applied, and allow inferences, to real-world conditions

Reliability

- Reliability
  - The ability of the measure to produce the same results under the same conditions

- Test-retest reliability
  - The ability of a measure to produce consistent results when the same entities are tested at two different points in time

Reliability and Validity – IQ test

- The reliability of an IQ test can be judged by comparing scores for the test given on one day to scores for the same test given at another time.
- To test the validity of an IQ test, we might compare the test scores to another indicator of intelligence, such as academic performance.
- Many critics say that IQ tests are reliable, but not valid:
  - They provide consistent results, but don't really measure intelligence.

Data sets, individuals and variables

- Raw data are numbers and category labels that have been collected but not yet processed in any way.
  - A data set is a complete set of raw data in an investigation.
- Individuals are the objects described by a set of data.
  - Individuals may be people, but they may also be animals and things.
- A variable is any characteristic of an individual.
  - A variable can take different values for different individuals.

Data Sources

- Primary data → Data collector uses the data they collected for analysis
  - Advantage: obtain data you need to suit your purpose.
  - Disadvantage: could be costly and time consuming.
- Secondary data → Data collected by other organisations or individuals
  - Advantage: data already collected or collection framework set-up – the hard has work has been done for you.
  - Disadvantage: how trustworthy are the data? They may not exactly match your needs.

Data sources

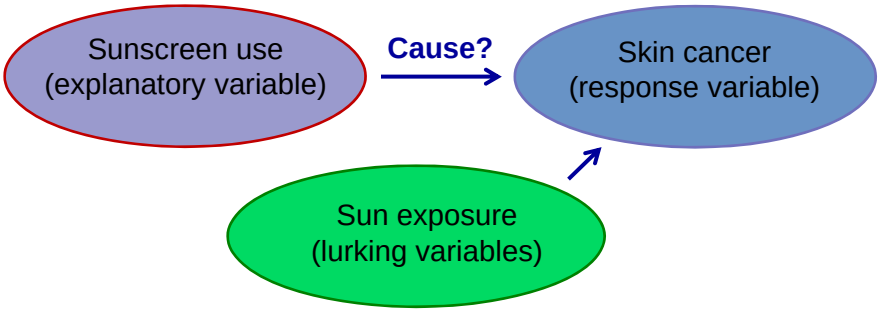
- We typically obtain data from two distinct types of investigative study designs:
  - Observational Study: In this type of study, we observe and measure specific characteristics, but we don't attempt to modify the subjects being studied.
  - Controlled Experiment: Here we randomly assign treatments to subjects and then proceed to observe its effects on the subjects.
- These studies try to identify the cause or explanation of some event or behaviour, by examining the association between a factor (explanatory variable) and an outcome (response variable).

Experimental research methods

- Cause and effect (Hume, 1748)
  - Cause and effect must occur close together in time (contiguity)
  - The cause must occur before an effect does
  - The effect should never occur without the presence of the cause
- Confounding variables: the '*Tertium Quid*'
  - A variable (that we may or may not have measured), other than the predictor variables, that potentially affects an outcome variable
  - E.g. the relationship between breast implants and suicide is confounded by self esteem
- Ruling out confounds (Mill, 1865)
  - An effect should be present when the cause is present and when the cause is absent the effect should be absent also
  - Control conditions: the cause is absent

Confounding

- Two variables (explanatory variables or lurking variables) are confounded when their effects on a response variable cannot be distinguished from each other.

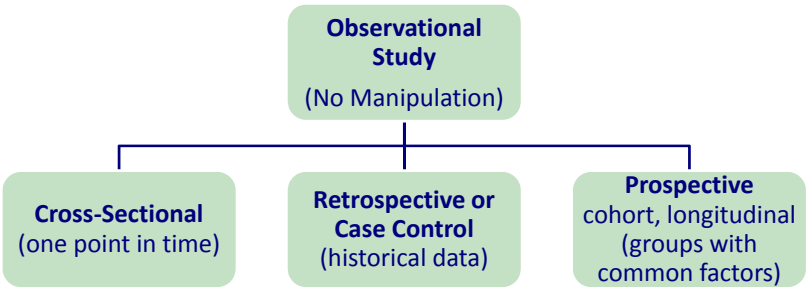


Example: Investigative study types

- Suppose we want to conduct a study to determine the effect of a Vitamin C pill on the risk of obtaining a cold.
- We compare those people who take the Vitamin C pill with those who do not.
  - Explanatory variable (Factor) = drug type
  - Treatments (Groups) = "Vitamin C pill" and "No pill".
  - Response variable (Outcome) = Obtain a cold (Yes/No)

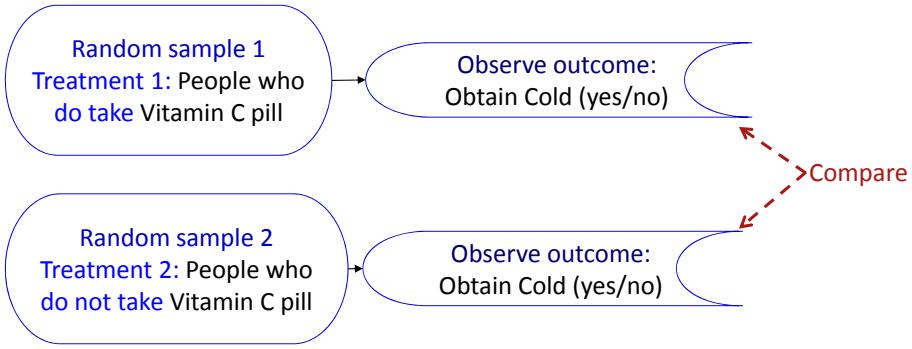
How would an observational study differ from an experiment in this example?

Observational study types



- Main challenge: balance groups for other variables, that is confounding variables, that might explain differences.

Observational study

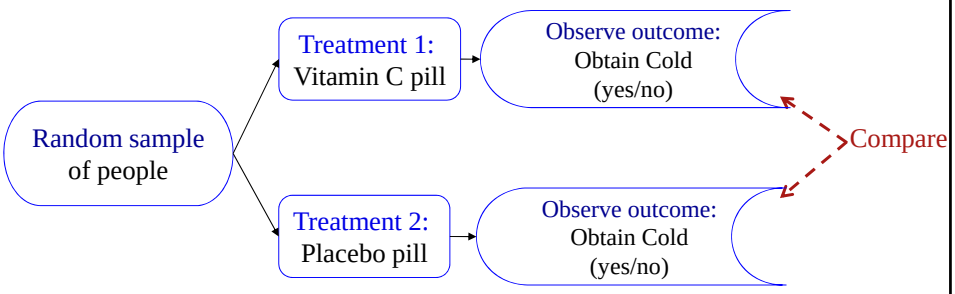


- The treatment groups arise from samples from different groups (or subpopulations).
- Difficult to establish whether the different outcomes are the result of the treatments or some confounding variable.

Experiment

- Conduct an experiment to determine the effect of a Vitamin C pill on the risk of obtaining a cold.
- Compare those people who take the pill with those who do not.
- Researcher has control over the treatments (Vitamin C or Placebo pill).
- Research randomly assigns a treatment to a person.
  - Random assignment ensures the groups are equivalent before applying the treatments.
- If we observe a difference between the two groups then we can more confidently conclude that type of treatment was the *cause* of the difference in the risk of obtaining a cold.

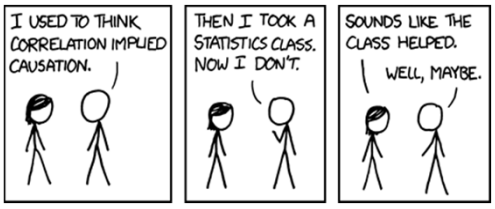
Experiment



- Randomisation minimises the effect of confounding variables.
- Expect on average the people in the Vitamin C pill sample have diets similar to the people in the Placebo pill sample.
- The effect of a glass of water will be reduced since people in both samples take a glass of water when taking the pills.

Association and causation

- Observational studies involve at least some aspects of design similar to experimental studies.
- However, because we are not controlling the conditions, we cannot investigate any causality issues.
  - Same problems with association between categorical variables as between numerical variables.



Association and causation

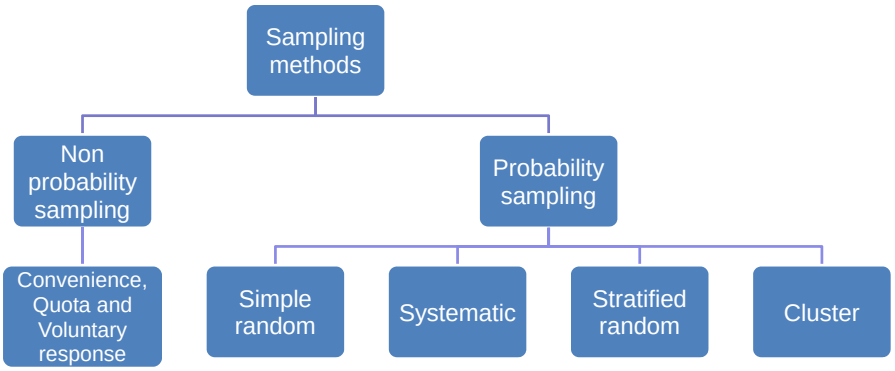
- Best method for establishing causation is to conduct a well designed experiment.
- When experiments are not possible, good evidence for causation exists when:
  - The relationship between the variables is observed in many studies of different types.
  - The association holds when the effects of plausible other variables are taken in to account.
  - A plausible scientific explanation exists for a relationship between the variables.
  - i.e. a lot of circumstantial evidence is needed.

Example: Do left-handers die young?

- Some years ago, a highly publicised study pronounced that left-handed people did not live as long as right-handed people (Cohen & Halpern, 1991).
- Letters were sent to the next of kin for a random sample of recently deceased individuals, asking which hand the deceased used for writing, drawing and throwing a ball.
  - The average age of death for those who had been left-handed was 66, while for those who were right-handed it was 75.
- A confounding factor has not been taken into account:
  - But, in the early part of the 20th century, many children were forced to write with their right hands, even if their natural inclination was to be left-handed.

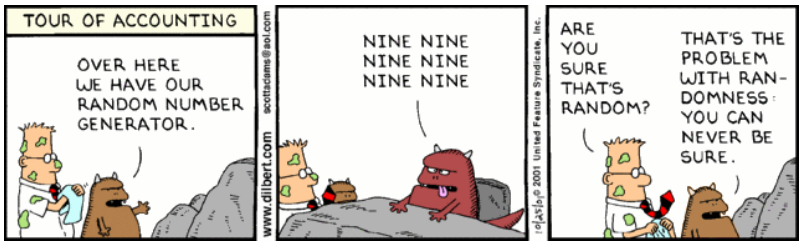
Sampling methods

- Non-probability sampling: select without knowing probabilities of selecting items/individuals – chosen on basis of availability.
- Probability sampling: select items/individuals with known probabilities.



The importance of randomisation

- An essential characteristic in statistical methods for analysing and interpreting data:
  - There is randomness in the manner in which the chosen individuals or subjects represent the general situation of interest.



Simple random sample

- A simple random sample (SRS) of size *n* consists of *n* individuals from the population chosen in such a way that every set of *n* individuals has an equal chance to be the sample actually selected.

- We can trust results from an SRS because it uses impersonal chance to avoid bias.

Errors and bias in sampling

- Sampling error: The difference between a sample result and the true population result.
  - Occurs when making a statement about a population based only on the observations in a sample taken from the population.
- Non-sampling error: Data are incorrectly collected, recorded or analyse and can occur in three ways.
  - Selection bias: The sampled population is different to the target population.
  - Defective measurement instrument
  - Recording data incorrectly
- Sampling error can be reduced by increasing the sample size.
- Non-sampling error cannot be reduced, only avoided.

Random and systematic error

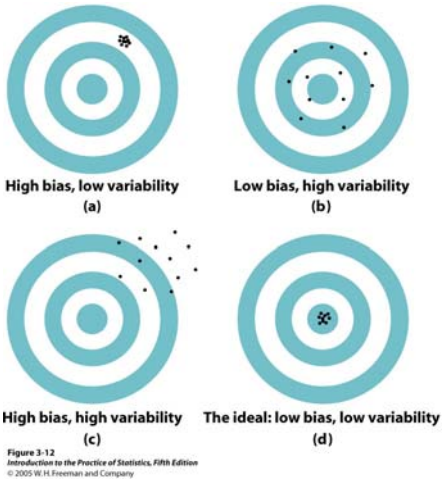


Figure 3-12
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H. Freeman and Company

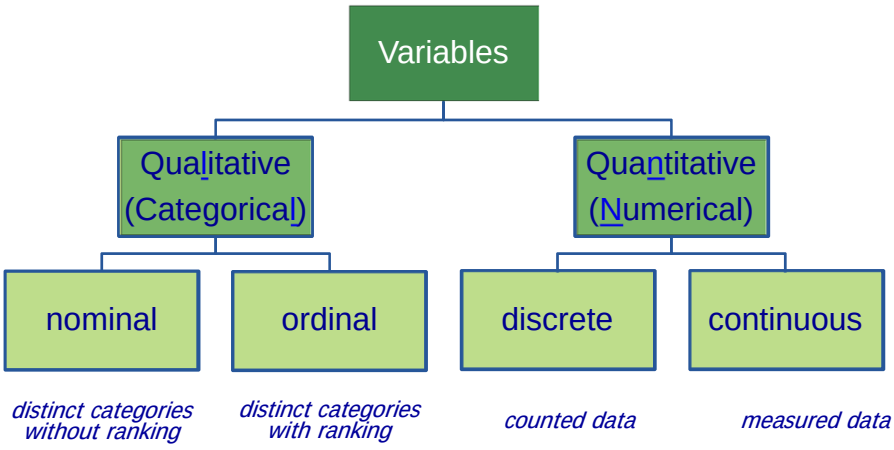
A good sampling design (like a good shooter) must have:

- High accuracy
  low bias
  validity
- High precision
  reliability

Example: Sampling bias

- We can easily access ratings for products, sellers, and companies through websites.
- These ratings are based only on those people who go out of their way to provide a rating.
- If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product?

Variable types



Variable types

- Qualitative Variable: A characteristic of interest best described/captured by non-numeric information. There are two types:
  - Nominal: Characterised by data that consists of names, labels and categories only. There is no order to the data.
  - Ordinal: Characterised by data that consists of names, labels and categories. The data can be arranged in some natural order.
- Quantitative Variable: A characteristic of interest best described/captured by numbers. There are two types:
  - Discrete: Can only take on a finite number of values that arise from a counting process.
  - Continuous: Can take on an infinite number of values and is derived from a measuring instrument (tape, scale, timepiece, etc.).

Example: Variable types

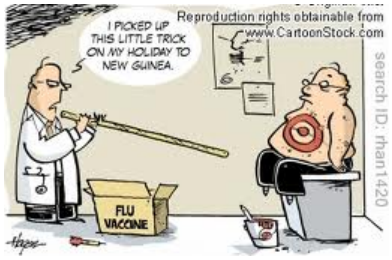
- Qualitative (Categorical)
  - Nominal:
    - Gender — 0 = Male; 1 = Female
    - Eye Colour — 1 = Brown; 2 = Blue; 3 = Hazel; 4 = Other
  - Ordinal:
    - Health status — 1 = Excellent; 2 = Good; 3 = Fair; 4 = Poor; 5 = Very Poor

Example: Variable types

- Quantitative (Numerical)
  - Discrete:
    - Number of flu vaccinations given
    - Number of patients treated during one day
    - Number of heart attacks
    - Number of people taking aspirin for heart health
    - Number of complaints about medication side-effects
  - Continuous:
    - Body Weight
    - Body Height
    - Time taken to run 100m

Exercise: Defining a statistical set-up

Do healthcare workers take their own medicine?



- Healthcare workers are widely recommended to have a flu vaccination.
- However vaccination rates amongst healthcare workers are thought to be consistently low (40%).
- Besides well-known factors such as scepticism and concerns about allergic reactions, are other factors such as *exercising regularly* causing the low uptake of the vaccination?
- A sample of 999 health care workers was studied.

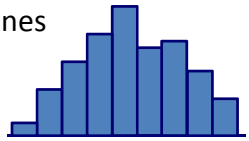
Ludwig-Beymer P, Gerc SC (2002), "An influenza prevention campaign: the employee perspective ",J Nurs Care Qual 16(3),1-12.

Exercise: Defining a statistical set-up

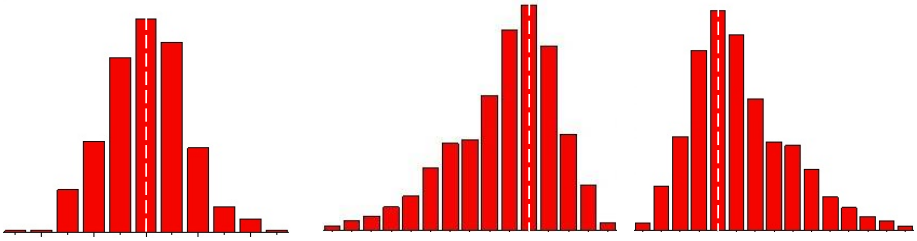
- What is the target population?
- How to find subjects?
- What are the variables of interest?
  - ☐ Categorical, nominal
  - ☐ Categorical, ordinal
  - ☐ Quantitative, discrete
  - ☐ Quantitative, continuous

Analysing data: Histograms

- A histogram is a bar graph such that:
  - The horizontal scale represents classes of data values and the vertical scale represents frequencies.
  - The heights of the bars correspond to the frequency values.
  - The bars are drawn adjacent to each other (without gaps).
- Look for the overall pattern and striking deviations.
- Can describe the overall pattern by its shape, centre and spread.
- An important kind of deviation is an outlier.
- This is an individual value that falls outside the overall pattern.
- The presence or absence of outliers also determines which numerical descriptives we can use.



Data distribution (shape & spread)

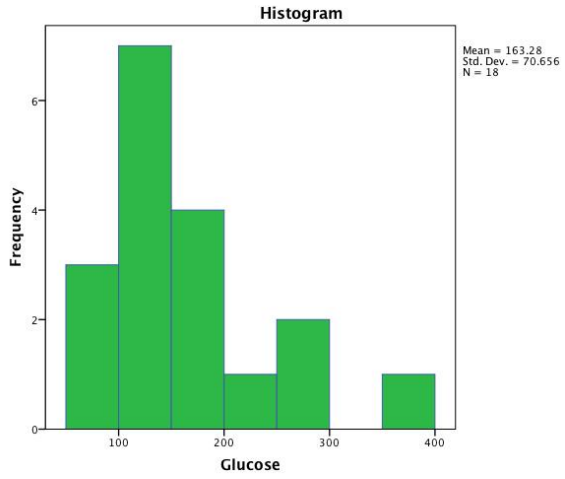


Symmetric | Skewed left | Skewed right

Observations are approximately evenly distributed about the centre of the histogram.

Observations are not symmetrically distributed. They concentrate around one end of the histogram and have a tail (right or left).

Example: Blood glucose levels



This distribution is skewed right.

It appears to have one peak between 100 and 150 mg/dl.

The spread is from 50 to 400 mg/dl.

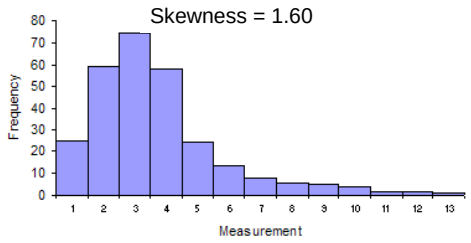
There appears to be at least one outlier, a blood glucose level of near 400 mg/dl.

What does this distribution suggest about the *typical* blood glucose level?

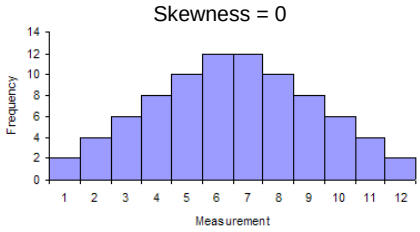
Skewness and kurtosis

- How skewed is a distribution?
  - Check the skewness statistic.
  - Characterises the degree of asymmetry of a distribution around its mean.
- How close are the data to being perfectly symmetric with a centre peak?
  - Check the kurtosis statistic.
  - Measures the flatness of the distribution.

Skewness

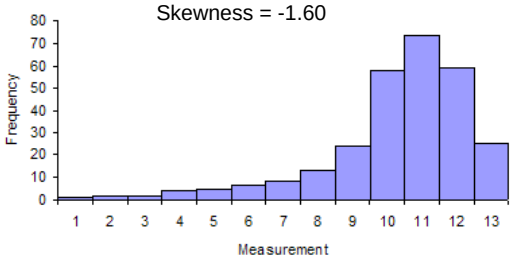


- If Skewness > 0; the distribution is right-skewed



- If Skewness = 0; the distribution is symmetric

Skewness



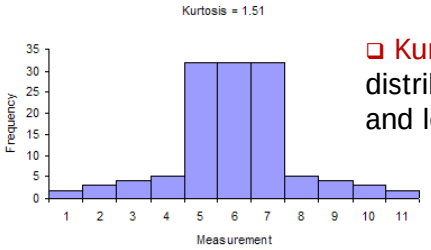
- If Skewness < 0; the distribution is left-skewed

- There are many tests for skewness, however some basic rules:

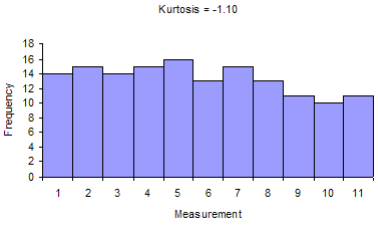
If Skewness < -2, the distribution is significantly left-skewed.

If Skewness > 2, the distribution is significantly right-skewed.

Kurtosis



- Kurtosis > 0: indicates a leptokurtic distribution that has a tall skinny peak and long, fat tails (like a kangaroo).



- Kurtosis < 0: indicates a platykurtic distribution that is flat with a short, rounded peak and short, thin tails (like a platypus).

Kurtosis

- The size and sign of Kurtosis indicates flatness of a distribution.
- Many tests for Kurtosis, however some basic rules:

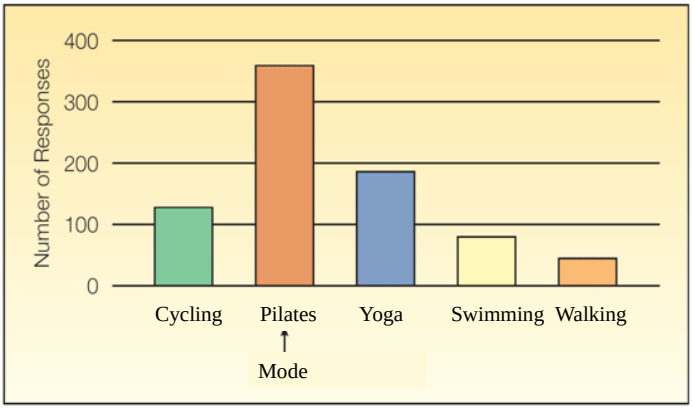
If Kurtosis > 2 the distribution is significantly peaked (leptokurtic).

If Kurtosis < -2 the distribution is significantly flat (platykurtic).

Central tendency and dispersion

- Central tendency
  - What is 'typical' for the data set under consideration?
  - Possible measures: Mode, Median, Mean
- Dispersion
  - How much can observations differ from what is 'typical' for the data set under consideration?
  - Possible measures: Range, Interquartile range

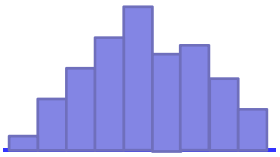
The mode



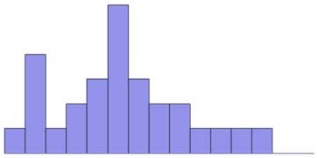
MODE: The value of the observation that appears most frequently.

Problems with the mode

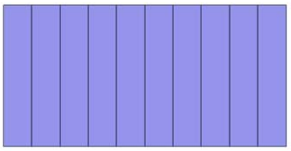
- How many peaks?
  - *Unimodal* if single (major) peak.



  - *Bimodal* if two (major) peaks.



Modality is not always guaranteed to exist and as such is the least preferred measure of central tendency



The mean

- The mean provides is what we most commonly call the average value:

$$\text{mean} = \frac{\text{sum of all values}}{\text{total number of values}}$$

- Also written as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

"sum" → $\sum_{i=1}^{n}$; "all values" → x_i; "mean" → $\bar{x}$; "total number of values" → n

The median

- The middle value in a data set arranged in order of magnitude.
  - 50% of the data have value less (or greater) than the median

If the number *n* of observations is *odd*:

Median = value of the $\frac{n+1}{2}th$ observation

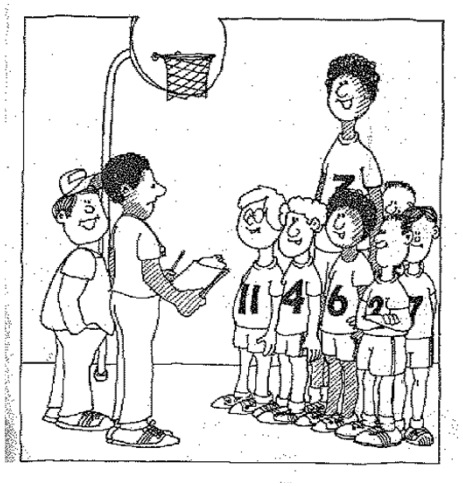
If the number *n* of observations is *even*:

Median = average of the $\frac{n}{2}th$ and $\left(\frac{n}{2}+1\right)th$ observations

Summary

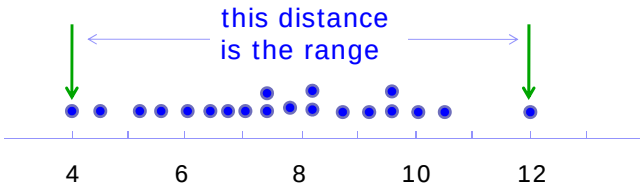
Location Measure	Advantages	Disadvantages
Mean	1. Can always calculate. 2. Useful mathematical properties. 3. Easy to calculate.	1. Not robust to outliers. 2. Sensitive to data distribution.
Median	1. Can always calculate. 2. Natural measure of centrality. 3. Robust.	1. Need to rank the data. 2. Lacks the useful mathematical properties of the mean.
Mode	1. Easy to understand. 2. Used for quantitative and qualitative data.	1. Not all data sets have a mode. 2. Mode not always unique. 3. No useful mathematical properties.

Mean or median?



Should we scare the opposition by announcing our mean height or lull them by announcing our median height?

Range

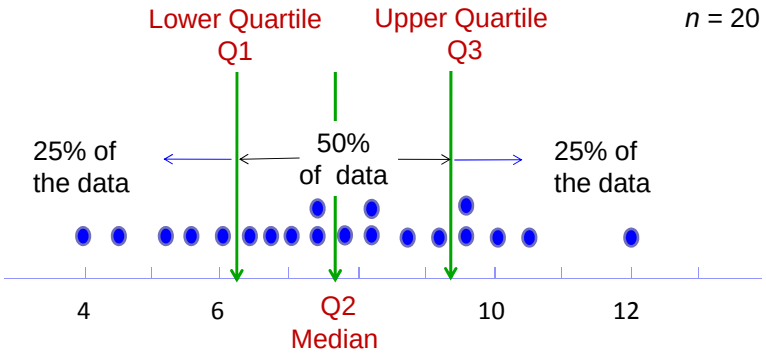


Range = Largest Value – Smallest Value

- Here Range = 12 – 4 = 8.
- Considers only the extreme values which may not be useful indicators of the bulk of the population.
  - These values can be errors in measurement or outliers.
- Rarely used as a measure of dispersion.

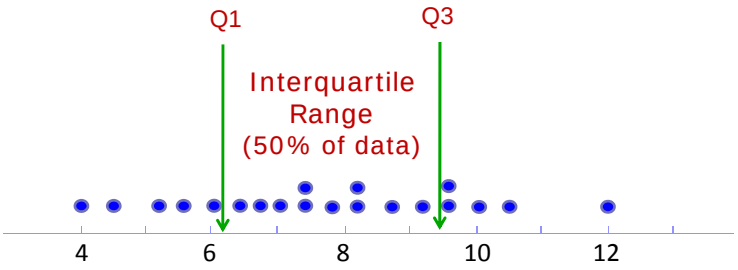
Quartiles

- Quartiles divide data into four equal parts:



Interquartile range (IQR)

- A better measure of dispersion.
- Measures the range of the middle 50% of the values.
- Calculated as the difference between the upper and lower quartiles:



$$\text{IQR} = Q3 - Q1$$

Exercise: Weekly rainfall

- Data were collected to compare weekly February rainfall at two small Australian towns, Eaglehawk and Bloomsbury.
- Over a period of four years, total weekly rainfall (in mm) was recorded for each of the four weeks in February, giving 16 readings for each of the two towns:

Eaglehawk				Bloomsbury			
13	18	5	10	58	29	66	42
0	0	1	2	14	20	29	31
96	17	38	0	83	79	73	66
7	1	130	11	52	38	36	34

- Compute the mode, median, mean, upper and lower quartile, range and interquartile range. How do they compare for the two locations?