Week 1

Variables:

Exploration, Familiarisation and Description.
Descriptive Statistics.

Convergent validity: the degree to which results/evidence from different tests/sources, converge on the same conclusion.

- A way you can get some confirmation that you're on the right track.

4 steps:

1. Recognise the problem.
2. Gather data to help understand and solve the problem.
3. Analyse and present the data
4. Act on the analysis.

Parameter: numerical measure that describes a characteristic of a population.
Statistic: numerical measure that describes a characteristic of a sample.

Descriptive Data:

Collecting, summarising and presenting data.

1. Collect Data.
   Eg. Survey
2. Summarise/Characterise Data.
3. Present Data.

Inferential Statistics:

Drawing conclusion about a population based on sample results.

1. Estimation
2. Hypothesis testing

Data Types

Categorical (non-numerical/Qualitative)

- Nominal (labels that do not imply order) eg. Yes/no.
- Ordinal (values that are still labels but have order) eg. HD/D/C/P/N

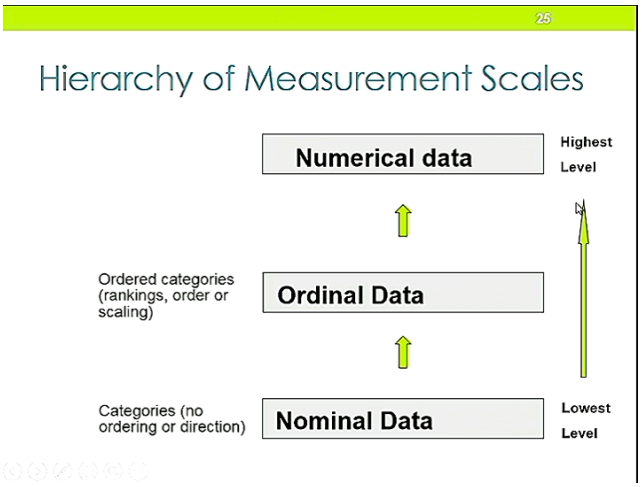
Categorical data CAN be coded numerically. Eg. Option 1,2,3,4,5.

Numerical (quantitative)

- Discrete (counting process) eg. How many children.
- Continuous (measured) eg. time

Data can be grouped or ungrouped.

- Grouped (Observations are grouped into classes eg. $30k - $50k)
- Ungrouped



Week 2

Numerical Data:

An example of numeric variable is salary.
How are salaries distributed across different people?
To answer this, ask these 5 questions.

1. What is the average salary?
2. How spread out are the salaries? (variance)
3. What are the extreme salaries at either end? (outliers)
4. Is the distribution of salary symmetric or skewed?
5. Does the distribution of salary have any other important features?

Measures of Central Tendency:

Mean = average

- Most used measure of central tendency.
- Very affected by extreme values.
- Aggregated distance of data values from the typical value is lowest if that 'typical' value is the mean.

Median = midpoint of ranked values.

- You have to rank the data first.
- Middle or middle mean of middle 2 values.
- Position of median = $\frac{n+1}{2}$

Mode = Most frequently observed value.

- You can have more than 1 mode.
- Mode can be used for nominal data.

Quartiles:

Position of quartiles:

Q1 = $\frac{(n+1)}{4}$

Q2 = $\frac{(n+1)}{2}$

Q3 = $\frac{3(n+1)}{4}$

N= number of data.

Percentiles: partition a set of data.

Week 3

Measures of Variation:

Range: simple measure of variation.

- The range is the difference between the largest and the smallest.
- Largest - Smallest.
- Ignores the distribution of the data.
- Sensitive to outliers.

Interquartile Range (IQR):

- 3rd Quartile - 1st quartile.
- Resistant to outliers.
- Range of the middle 50% o the data.

Using boxplots is a good way of describing *numerical data*.

To summarise a set of data:

1. Measure of average (mean, median, mode)
2. Measures of average.

Measures of Variation:

- Standard deviation squared.
- Each value in data contributes to it.
- It is sensitive to outliers.

Standard deviation:

- Square root of variance.
- Easier to interpret.

Notation	Sample	Population
Number of observations	n	N
Mean	$\bar{X}$	μ
Variance	s^2	σ^2
Std deviation	s	σ

Population standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Sample standard deviation:

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

Shape of Distribution:

2 good ways to examine the distribution of numerical variables:

1. Histogram
2. Boxplot

… now, how to describe distribution?

Shapes of distribution:

Symmetrical:

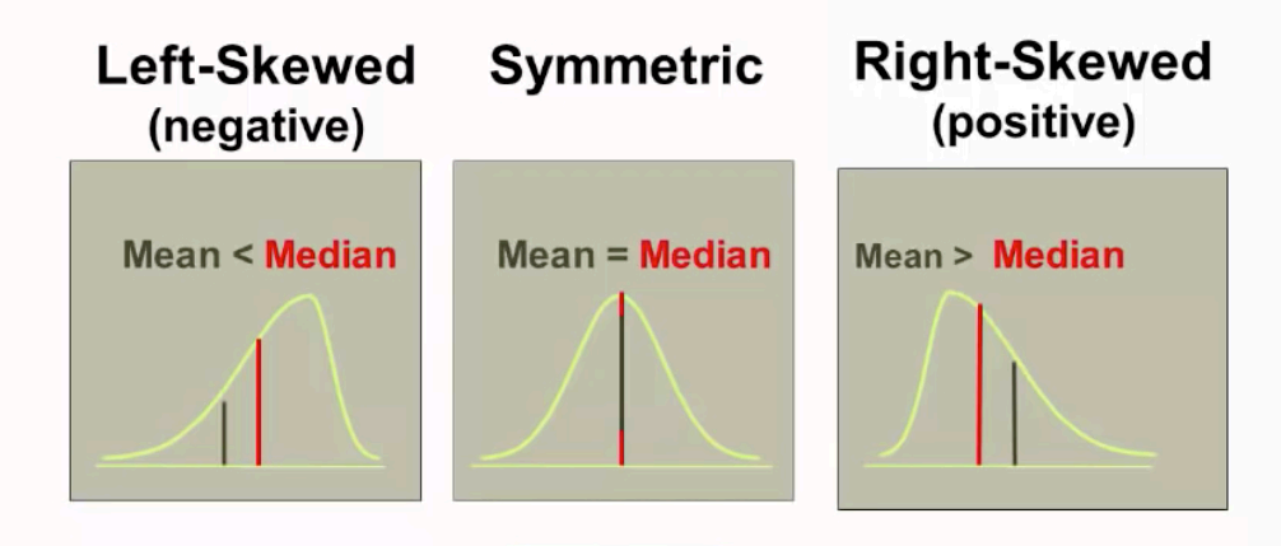
- Has a single peak.
- Looks approx. the same left and right.

Positively skewed/Right skewed:

- When the tail is toward the right, it is right skewed.

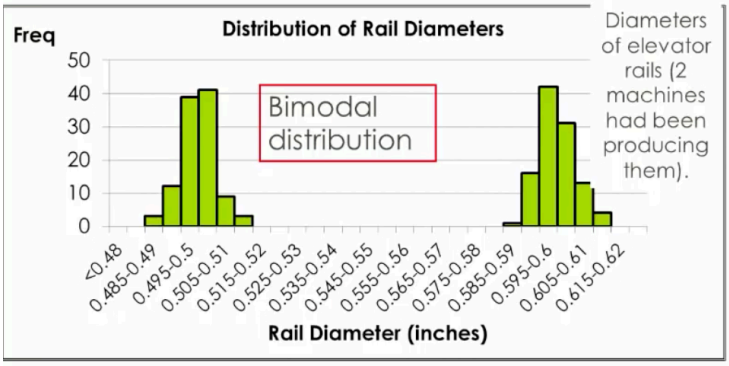
Negatively skewed/ Left skewed:

- When the tail is toward the left.



Multimodal:
Bimodal distribution.

- Has two peaks, not necessarily equal height.
- In this case, split the data into 2 sets and analyse separately.



Probability

The link between descriptive and inferential statistics.

Probability: a numerical value that represents the chance, likelihood, possibility that an event will occur.

Event: each possible outcome of a variable.

The probability that random variable X is equal to a particular value x is denoted: P(X = x)

Probabilities *p(x)* are estimated from relative frequencies.
Eg. 3/60 = 0.05

All probabilities must lie between 1 and 0. And the sum of all probabilities must equal 1.

The Binomial Distribution:

1. The experiment consists of n trials.
2. Two possible outcomes of each trial. Success/Failure.
3. The probability of success is identical at each trial.
4. Trials are independent.

Eg.
Experiment: toss a coin 3 times.

1. A trial is 1 toss of a coin. N = 3
2. We are interested in the number of heads. Head = success
3. P(success) = 0.5. P(failure) = 0.5
4. Trails are independent because the outcomes of one toss is independent of the outcome of another.

Random variable X is the number of heads.
= binomial distributed.

See table 4a and 4b in formula and statistic tables on Moodle.
4a gives point probabilities.
4b gives you cumulative probabilities.
Or.

- Use excel's statistical function BINOM.DIST:

EG. Where x is binomially distributed, n = 10, find

- $P(3<X<8) = P(X\leq 7) - P(X\leq 3)$

Week 4

The Continuous Distribution

(recall discrete random variables)

- Toss a coin 3 times and look at number of heads (x)
- X = 0,1,2 or 3.
- We can calculate P(X = a particular value)
- Eg. P(X=3) = 0.125

A continuous random variable:

- Has an uncountable infinite number of values.
- Not any exact number.
- Can assume any value in the interval (between 2 points)

Eg. Survey of women's heights.

- Height of randomly selected woman.
  A continuous random variable.
- X may take any value.

Its not useful to consider that X will equal *an exact number*
However, it is sensible to consider that X will lie within a range.

- Eg. P(161.5 < X < 162.5)

The Probability Density Function:

- Organise the data into class intervals of 5cm.
- Plot the corresponding RELATIVE FREQUENCY:
- When you reduce the class interval (make it smaller) it will make the graph smoother.

Continuous distribution has a continuum of possible values.

- Eg, X = all values between 0 and 100 or
- X = all values greater than 0.

Then, the total probability of 1 is spread over this continuum.

F(x) measure **probability density.**

- The interval X values which are more likely to occur are shown in the regions of the graph where the probability density is larger.
- The larger the density, the more probable that it will occur there.

Total area between the graph of f(x) and horizontal axis represents the total probability = 1.

The Normal Distribution

The most important probability Distribution in statistics.

- This is because many data sets have a histogram that is well described by the normal distribution.

Properties:

- Symmetrical, unimodal.
- Mean = median (approx.)
- Modal class is in the region of the mean(median)
- The curve extends to ± infinity in both directions.
- The distribution is completely defined by two parameters.
  - Mean and Variance
- Expressed as **X ~ N(mean, variance)**