Quantitative Methods 1 – Textbook Readings Summary

Statistics – a way to get information from data, which can be subdivided into two basic areas: descriptive statistics and inferential statistics
Descriptive statistics – methods of organising, summarising and presenting data in ways that are useful, attractive and informative to the reader
Inferential statistics – methods used to draw conclusions about a population based on information provided by a sample of the population

Descriptive Measures

Population – the entire set of observations or measurements under study
Sample – a subset of observations selected from the entire population
Statistical inference – the process of inferring information about a por
Parameter – a measurement about a population
Statistic - a measurement about a sample
Estimate – an approximate value of a parameter based on a
Point estimate – a single value given as an estimate of a ... mple mean)
Confidence interval – the proportion of times that ... correct, if the sampling procedure were repeated a very large
Significance interval – measures how frequ... in the long run
Simple random sample – one in which e... an equal chance of appearing
Sampled population – the actual ... has been drawn

Variable – any characteristic
Data – observations of th

Three types of dat
1. **Numerical (** ... bservations are real numbers
2. **Nominal** ... – observations are categorical or qualitative
3. **Ordin** ... inal data

G ... **minal data**
... of presenting data and their count in each category or class
Re ... **n** – frequency distribution giving the percentage each category or class
Bar char ... hich vertical bars represent data in different categories
Pie chart – ... ubdivided into sectors representing data in different categories
Bar charts, pi... harts and frequency distributions are used to summarise single sets of nominal (categorical) data. Because of the restrictions applied to this type of data, all that we can show is the frequency and proportion of each category. The type of chart to use in a particular situation depends on the particular information the user wants to emphasise.

Descriptive statistics is concerned with methods of summarising and presenting the essential information contained in a set of data, whether the set be a population or a sample taken from a population.

Graphical descriptive techniques – Numerical data

- A **frequency distribution** is an arrangement or table that groups data into non-overlapping intervals called classes.
- Approximate class width = (Largest value – Smallest value)/Number of classes
- A **histogram** is a graphical presentation of a frequency distribution of numerical data.
- Class relative frequency: Percentage of data in each class
  Class relative frequency = (Class frequency)/(Total number of observations)
- **Relative frequencies** permit a meaningful comparison of data sets even when the total numbers of observations in the data sets differ.
- To facilitate interpretation, it is generally best to use equal class widths whenev possible. In some cases, however, unequal class widths are called for to avoid having t sent several classes with very low frequencies
- **Stem-and-leaf display** – display of data in which the stem consists of t of a given point and the leaf the digits to the right
  - ✓ Whether to arrange the leaves in each row from smalle in order of occurrence is largely a matter of persona
  - ✓ A stem-and-leaf display is similar to a histogram t holds the advantage of retaining the original

Shapes of a histogram

- **Symmetric histogram** – a histogram in wh iddle, the two halves would be mirror images
- **Positively skewed histogram** – a his
  - ➢ Mean > Median > Mode
- **Negatively skewed histogram** – ne left
  - ➢ Mode > Median > M
- **Modal class** – the class vations
- **Unimodal histogra** ue
- **Bimodal histogr**
- **Multimodal** or more peaks
- **Bell-shar** ell (or mound-shaped)
- A **fre** ting the frequency of each class above the midpoint of nts with straight lines. The polygon is usually closed by (with zero frequency) at each end
  to focus on the *trend* in exports and imports over the years, a
  e the values of exports and imports in different years, a **bar chart** is

It is importa derstand that if two variables are linearly related it does not mean that one is causing the oth . In fact, we can never conclude that one variable causes another variable. We can express this more eloquently as **'correlation is not causation'**.

A parameter is a descriptive measurement about a population, and a statistic is a descriptive measurement about a sample.

Random variables and discrete probability distributions

Measures of central location

- Mean: the sum of a set of observations divided by the number of them
  - A serious drawback of the **mean** is that it is seriously affected by outliers, which are extreme observations. For this reason, in the existence of outliers, the **median** provides a better measure of central location.
- Median: the middle value of a set of observations when they are arranged in order of magnitude
  - When there are a relatively small number of extreme observations, the **median** usually produces a better measure of the centre of the data.
  - The calculation of the mean is **not valid** for ordinal and nominal
  - The **median** is appropriate for ordinal data.
- The **mode**, which is determined by counting the frequency of ea
  appropriate for nominal data. The mode is the most frequent

Measures of spread:

- **Deviation** – Difference between an observation and ngs to
- **Standard deviation** is the more useful meas the measure is to be used in conjunction with out a single population
- **Range:** the difference between the
- The **coefficient of variation** of d deviation of the observations divided by the
- The coefficient of variati reported as a percentage, which effectively expresses tage of the mean.
- **Outlier** - An obser less than Q1- 1.5(1QR).
- The interquart e middle 50% of the observations.

Percentile - T or which p% of observations are less than that are greater than that value.

ercentile

$$L_p = (n+1)\frac{p}{100}$$

re L_p is the location of the pth percentile.

Quadrants

- **Q1**: where the first 25% of data lies
- **Q2**: where the first/last 50% of data lies, equivalent to the median
- **Q3**: where the last 25% of data lies

Relationship between two variables

- **Covariance** - A measure of how two variables are linearly related.

- Measured in the **same unit as data but squared** (unit of X times unit of Y = unit of X squared = unit of Y squared)

- **Coefficient of correlation** - A measurement of the strength and direction of linear relationship between two numerical variables
  - **Unit-less** – takes values from **-1 to +1**
  - A value of 0 means no association between the two variables
  - A value between 0 and 1 means a positive association
  - A value between -1 and 0 means a negative association
  - The closer to 1 or -1, the stronger the association
  - The closer to 0, the weaker the association
  - The correlation is a better measure of the linear association between two ... than covariance because correlation is unit free, and therefore measures the ... the linear relationship, whereas covariance does not.
  - The correlation is based on the covariance
- **Least squares method** - A method of deriving an estimated line which best fits the data.
  - $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$
  - $\beta_0$and β_1 are scaled in terms of the number of ...
  - The coefficients $\beta_0$and β_1 are derived using ... sum of squared errors (SSE): $\sum_{i=1}^{n}(y_i - \widehat{y_1})^2$
  - When using least squares to estimate ...
    - ✓ **The sum of residuals = 0**
    - ✓ **The mean of residual ...**
    - ✓ **The sum of square ... will be minimised**
- **Coefficient of determination** ... variation in the dependent variable that is ***explained*** ...
  - Equivalent to th ... of correlation

Knowing that an e ... us that its expected value equals the parame ... e estimator is to the parameter.

Law of Itera...

- ***L...***
- ***... [var(Y /X)] + var[E(Y /X)]***

Expect... + Y] = E[X] + E[Y]
Variance: ... + Y) = V(X) + V(Y) + 2COV(X,Y)
V(X - Y) = V(X) + V(Y) - 2COV(X,Y)
For independent X and Y then: V(X + Y) = V(X - Y) = V(X) + V(Y)

Hypothesis Testing

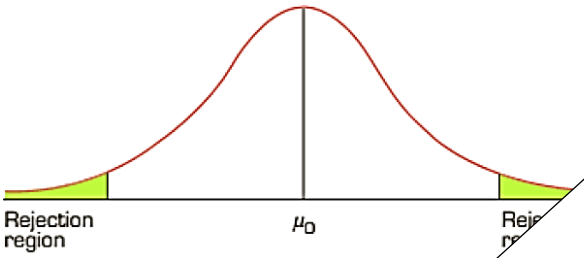
Hypothesis - A proposition or conjecture that the statistician will test by a means coiled hypothesis testing.

One-tail and two-tail tests

- **Two-tail test:** a test with the rejection region in both tails of the distribution, typically split evenly

$H_0: \mu = \mu_0$

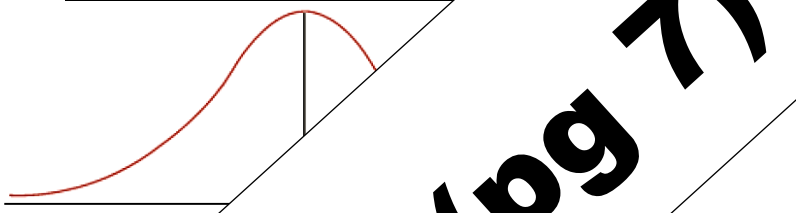
$H_A: \mu \neq \mu_0$



- **One-tail test:** a test with the rejection region in only one tail

$H_0: \mu = \mu_0$

$H_A: \mu > \mu_0$



ight-tail test

$H_0: \mu = \mu_0$

$H_A: \mu < \mu_0$

Left-tail test

Power of a te

- **Th… …when it is false = 1-b = 1-Pr(Type 2 error)**
- …ons from the null hypothesis in a single direction, then a … power (less probability of type II error) than a two tail test at …e. As a result, positive deviations of μ from m have a higher …cted by the upper tail test. So a two tail test is not invalid in this … it is not as good as the upper tail test in terms of power

Type 1 and 2

- **Type 1 e…or: H0 is true but is rejected : α**
- **Type 2 error: H0 is false but is not rejected : b**
- *As alpha increases, the probability of a type 1 error increases, the probability of a type 2 error decreases, and therefore the power of the test increases.*
- As α decreases, the level of confidence for the interval will increase
- Increasing the level of confidence comes by decreasing α, the probability of a type I error. That is, increasing the confidence level is equivalent to saying we want to decrease the probability of rejecting the true value for the parameter under test.