Basic notes on course

The course consists of two parts

Descriptive

First block Large data set, few summary measures, extract essential info from data

Inferential

Second block

Definitions

Statistic

Value calculated from a set of sample data (latin letters used) Descriptive summary measure for a sample

Parameter

Value calculated from data about the entire population (greek letters are used) Numeric description of characteristic of the population

Population

Set of all items under investigation

Target population

The entirety of the population of units which falls under the scope of the study/ population about which inferences are to be made.

Sampled population

The section of the target population for which information is (can be) collected.

Raw data

Data which is unorganised numerically and remains in the form in which it was collected.

Array

An arrangement of raw data in de/ascending order of magnitude.

Randomness

Equal chance of selecting any member of the sample space.

Range

the difference between the largest value (observation) and the smallest value (observation) of the data set

Cluster

a group of points that fall very closer together

Outlier

an observation of datum that is unusually far from the bulk of the data.

Relative Frequency

Proportion of the data having a certain property. Relative frequencies give the proportion of the observations falling in a particular group.

Exam notes

Exam is representative of the entire course.

General background

Summary of statistic process

General

Collect data, organise, summarize, present (graphs and shit), analyse, interpret/valid conclusions, make decisions

Collection

Collect data from a set of unit which are under identical conditions, but which nevertheless display natural random variation .

Sampling

A sample is taken if the whole measuring the whole population is unfeasible. A sample should be unbiased/representative, ie it should have all of the same conditions as the population as a whole.

Data

Quality

Scales of measurement

All data can be sorted into one of the four scales of measurement,and appropriate methods can validly be performed on it to give a meaningful result.

Qualitative

Relating to separate categories.

Nominal Scale

Categorical

Examples

- gender type
- Colors.

NOTE: don't confuse the fact that the categories cannot be ordered with the fact that data points within a given category can be ordered.

Ordinal

categorical data that can be ordered/ ranked.

Examples

- Grades

Quantitative.

Discrete

Countable

Examples

- humans
- atoms
- (positive)integer numbers numbers.

Continuous

non- countable, smooth variation

examples

- time
- space
- (positive) real numbers.

Interval Scale.

difference between scale points is the same anywhere along the scale, however the zero of the scale is not the true zero (complete absence) of the variable which it measure.

Examples

- Fahrenheit scale

NOTE: does this scale include discrete data?

Ratio Scale

both differences between scale points and ratio's of data values are meaningful, the scale has a true zero.

Selection method

Direct observation

interval methods

Experimentation

Data is generate though the manipulation of variables under controlled conditions data on the primary variable under study can be monitored and recorded while conscious efforts are made by the researches to control the effects of a number of influencing factors.

Examples

this method is usually used in Biology, Chemistry and other natural sciences

Graphical Methods

Background

graphical techniques are often used to convey statical information (results) to the public who may wish to use them.

advantages

- quick and easy to see what is going on
- easy to pick out trends

Definitions

Range

the difference between the largest value (observation) and the smallest value (observation) of the data set

Cluster

a group of points that fall very closer together

Outlier

an observation of datum that is unusually far from the bulk of the data.

Relative Frequency

Proportion of the data having a certain property. Relative frequencies give the proportion of the observations falling in a particular group.

Types of graphs

- Dot Plots
- Pie charts
- Line graphs
- Bar graphs.charts I stem and leaf plots
- Histograms
- Boxplots.

Dot Plot

A plot of values on a vertical or horizontal scale. each number is represented by a dot and the dots must be distinct.

Pie Chart

A circle divided into segments. the size of each segment is proportional to the importance of the data category of random variable relative to the whole. (normally used for categorical data)

Line Graph

shows the value of one variable against another (usually time), used to investigate trends between two variables.

Bar Chart

Presentation of the categorical or discrete data by means of bars/ blocks orientated side by side.

Stem and Leaf plot

numbers are split into two parts. A group of digits from the left of the number (the leaf) and the remainder of the data (the stem) leaves and stem are then arranged on either side of a vertical spacer.(|)

The graph requires a key to indicate where decimal point falls.

Good for plotting discrete of continuous data with many observations (abut 15-300)

Histogram

A histogram is constructed form a frequency table, with bar continuous on one another. The intervals are shown on the x axis, and the number of items/data points within each interval is shown if the u Y axis.

A histogram is usually used for continuous variable but it can also be used to plot discrete variables. If it is used on a discrete variable then the variable is/ must be converted into a continuous variable.

####Advantages Good for showing symmetry of data.

R coding

Background

Knowledge of R code will not be tested directly in the course, however examples of data put through r code will be given and must be interpreted.