TOPIC 1

Definitions

- A **population** consists of all the members or selected variables/measurements of a group about which you want to draw a conclusion
- A **sample** is the portion of the population selected for analysis
- A **parameter** is a numerical measure that describes a characteristic of a population
- A **statistic** is a numerical measure that describes a characteristic of a sample

Population

Measures used to describe a population are called **parameters**

Sample

Measures computed from sample data are called **statistics**

Data – Types and Collection

- Data = observed values of a variable
- Collection
    - Primary – you collect for your purpose
    - Secondary – someone else collected for their purpose

Descriptive statistics

- Collect, organise, present and summarise data.

Inferential statistics

- Analyse and interpret the data to draw conclusions (i.e., make inferences)
- Form conclusions about a population (all items of interest) on the basis of a representative sample (portion of the population).
- Sample must represent the population
- For example a simple random sample: every item has the same chance of being selected
- There is a chance of making an incorrect conclusion
- Use probability to determine the reliability of our inference

Displaying Data

- Frequency table – categorical or numerical
    - Gives the frequency of each data value
- For numerical data need to group in classes before constructing frequency table if
    - Continuous data
    - Large number of discrete values
- Classes should be
    - Mutually exclusive and exhaustive
    - Have same width

Histogram

- Frequency on vertical axis
- Alternatively relative frequency on vertical axis
- Classes on horizontal axis
- Rectangles represent class frequency/relative frequency
- Always use classes of equal width
- So area is proportional to frequency
- And height of each rectangle represents frequency
- Start vertical axis at zero so not to misrepresent data

Descriptive Measures

- **Overall objective:** to extract meaningful information from a data set – sample or population.
- Can describe general shape and distribution of data using tables and graphs.

Numerical Descriptive Measures

- Can also calculate numerical descriptive measures (numerical measures of summary)
- These measures are precise, objectively determined, easy to manipulate, interpret and compare.
- They allow careful analysis of data (especially important when using sample data to make inferences about entire population).
- Usually interested in 2 such measures
    - centre of the data
    - spread or variability of the data

Measure of Central Location

- Where is the middle of the data?
- Most common question about data
- Three measures
    - Mean – must be able to use your calculator in statistical mode to calculate this
    - Median
    - Mode

Mode

- Most frequently occurring value
- The mode is not necessarily unique (i.e., can have more than 1 mode). e.g.,
    - Bimodal (2 modes)
    - Multimodal.
- Can have no mode
- Mode is useful when interested in the most "common value" e.g., The most frequently purchased size of .....
- Doesn't take account of the magnitude of all data values

Median

- Middle value when data **ordered** from lowest to highest values.
- Half the values lie below, half above the median
- If there are n values and n is odd, the median (Md) is the middle value.
- If n is even, the Md is the average of the 2 middle values.
- For both cases, the Md is the $\frac{n+1}{2}$ ranked value
- Median is not affected by outliers (extremely large or small values).
- Sometimes more useful than other measures – eg. income
- Doesn't take account of the magnitude of all values

Mean

- Arithmetic Mean
    - Sample: $\bar{x} = \frac{\sum x}{n}$ Population: $\mu = \frac{\sum x}{N}$
- Most common, most important
- Not perfect – affected by extreme values
- Easy to calculate, usually use statistical calculator or Excel
- Easy to interpret
- Easy to manipulate mathematically
- Use sample mean to make inferences about the corresponding population mean

Measures of Variability

- A measure of central tendency by itself does not completely summarise a set of data.
- Tells us nothing about the variability or spread of the data.

DISTANCE MEASURES

Range

- Largest value less smallest value
- Often not very informative
- Very crude: provides no information about values between minimum and maximum value
- Distorted by extreme minimum or maximum values

Quartiles

- Quartiles, (also percentiles, deciles, quintiles).
  - Cut the data into quarters (equally sized bits)

- The **lower (of first) quartile** (25th percentile) is given by location of

  $Q1 = (\frac{n+1}{4})^{th}$ ranked value

  25% of values are at most and 75% are at least

- The **upper quartile (or third)** (75th percentile) is given by location of

  $Q3 = (\frac{3(n+1)}{4})^{th}$ ranked value

  75% of values are at most and 25% are at least

Note: if formula gives .25 or .75 round to nearest integer.

IQR

- The **interquartile range** is given by $IQR = Q_3 - Q_1$
- Measures the range of the middle 50% of the data values
- Discarded or **trimmed** highest and lowest 25% of the values
- The median can be thought of as the 2nd or middle quartile

Five-Number Summary

- The five statistics:
  - Minimum value
  - Lower quartile
  - Median
  - Upper quartile
  - Maximum value,

  Arranged in order of magnitude is the five-number summary
- Can illustrate the five-number summary by a box and whisker diagram (or box-plot)

Variance and Standard Deviation

- Compares each data value to the mean
- Based on squared deviations from the mean
- Variance:

  $\sigma^2 = \frac{\sum(X-\mu)^2}{N} = \frac{SSX}{N}$ $\qquad$ $s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{SSX}{n-1}$
- Standard deviation:

  $\sigma = \sqrt{\sigma^2}$ $\qquad$ $s = \sqrt{s^2}$
- Use statistical calculator or Excel to calculate

Statistics
A branch of mathematics concerned with the collection and analysis of data.

Variables
Characteristics or attributes that can be expected to differ from one individual to another.

Data
The observed values of variables.

Operational definition
Defines how a variable is to be measured.

Population
A collection of all members of a group being investigated.

Sample
The portion of the population selected for analysis.

Parameter
A numerical measure of some population characteristic.

Statistic
A numerical measure that describes a characteristic of a sample.

Descriptive statistics
The field that focuses on summarising or characterising a set of data.

Inferential statistics
Uses information from a sample to draw conclusions about a population.

Statistical packages
Computer programs designed to perform statistical analysis.

Primary sources
Provide information collected by the data analyser.

Secondary sources
Provide data collected by another person or organisation.

Focus group
A group of people who are asked about attitudes and opinions for qualitative research.

Categorical variables
Take values that fall into one or more categories.

Numerical variables
Take numbers as their observed responses.

Discrete variables
Can only take a finite or countable number of values.

Continuous variables
Can take any value between specified limits.

Nominal scale
A classification of categorical data that implies no ranking.

Ordinal scale
Scale of measurement where values are assigned by ranking.

Interval scale
A ranking of numerical data where differences are meaningful but there is no true zero point.

Ratio scale

A ranking where the differences between measurements involve a true zero point.

Summary table
Summarises categorical or numerical data; gives the frequency, proportion or percentage of data values in each category or class.

Bar chart
Graphical representation of a summary table for categorical data; the length of each bar represents the proportion, frequency or percentage of data values in a category.

Pie chart
Graphical representation of a summary table for categorical data, each category represented by a slice of a circle of which the area represents the proportion or percentage share of the category relative to the total of all categories.

Frequency distribution
Summary table for numerical data; gives the frequency of data values in each class.

Class width
Distance between upper and lower boundaries of a class.

Range
Distance measure of variation; difference between maximum and minimum data values.

Class boundaries
Upper and lower values used to define classes for numerical data.

Class mid-point
Centre of a class; representative value of class.

Relative frequency distribution
Summary table for numerical data, which gives the relative frequency of data values in each class.

Percentage distribution
Summary table for numerical data; gives the percentage of data values in each class.

Cumulative percentage distribution
Summary table for numerical data; gives the cumulative frequency of each successive class.

Histogram
Graphical representation of a frequency, relative frequency or percentage distribution; the area of each rectangle represents the class frequency, relative frequency or percentage.

Percentage polygon
Graphical representation of a percentage distribution.

Cumulative percentage polygon (ogive)
Graphical representation of a cumulative frequency distribution.

Contingency table (or cross- classification table) – descriptive statistics
Summary table for two categorical variables: each cell represents data that satisfy the given values of both variables.

Side-by-side bar chart
Graphical representation of a cross- classification table.

Time-series plot
Graphical representation of the value of a numerical variable over time.

Chartjunk
Unnecessary information and detail that reduces the clarity of a graph.

Central tendency
The extent to which data values are grouped around a central value.

Variation

The amount of scattering of values away from a central value.

Shape
The pattern of the distribution of data values.

Spread (dispersion)
The amount of scattering of values away from a central value.

Arithmetic mean (mean)
Measure of central tendency; sum of all values divided by the number of values (usually called the mean); called the arithmetic mean to distinguish it from the geometric mean.

Sample mean
Mean calculated from sample data.

Median
Measure of central tendency; middle value in an array.

Mode
Measure of central tendency; most frequent value.

Quartiles
Measures of relative standing, partition a data set into quarters.

First (lower) quartile
Value that 25% of data values are smaller than, or equal to.

Second quartile
The median value that 50% of data values are smaller than, or equal to.

Third (upper) quartile
Value that 75% of data values are smaller than, or equal to.

Geometric mean
Average rate of change of a variable.

Range
Distance measure of variation; difference between maximum and minimum data values.

Interquartile range
Distance measure of variation; difference between third and first quartile; range of middle 50% of data.

Resistant measures
Summary measures not influenced by extreme values.

Variance
Measure of variation based on squared deviations from the mean; directly related to the standard deviation.

Standard deviation
Measure of variation based on squared deviations from the mean; directly related to the variance.

Sum of squares (SS)
Sum of the squared deviations.

Sample variance
Variance calculated from sample data.

Sample standard deviation
Standard deviation calculated from sample data.

Coefficient of variation
Relative measure of variation; the standard deviation divided by the mean.

Z scores
Measures of relative standing; number of standard deviations that given data values are from the mean.

Extreme value (outlier)
Value located far from the mean; will have a large Z score, positive or negative.

Symmetrical
Distribution of data values above and below the mean are identical.

Skewed
Non-symmetrical distribution; data values are clustered either in the lower or the upper portion of the distribution.

Population mean
Mean calculated from population data.

Population variance
Variance calculated from population data.

Population standard deviation
Standard deviation calculated from population data.

Bell-shaped
Symmetric, unimodal, mound- shaped distribution.

Empirical rule
Gives the distribution of data values in terms of standard deviations from the mean for bell-shaped distributions.

Chebyshev rule
Gives lower bounds of the distribution of data values in terms of standard deviations from the mean for any distribution.

Five-number summary
Numerical data summarised by quartiles.

Box-and-whisker plot
Graphical representation of the five- number summary.

Covariance
Measure of the strength of the linear relationship between two numerical variables.

Sample covariance
Covariance calculated from sample data.

Coefficient of correlation (or correlation coefficient)
Measure of the relative strength of the linear relationship between two numerical variables.

Sample coefficient of correlation
Coefficient of correlation calculated from sample data.

QUESTIONS

The sum of the relative frequencies for all classes in a relative frequency distribution will always equal:
One

Topic 2

What is Probability all about?

- Evaluating chance / likelihood / risk
- Deal with uncertainty in a rational manner.
- Where decisions made under uncertainty probability is used to measure the degree of uncertainty
- Crucial to inferential statistics
- Links what we know to what we infer

Probability – Statistical Inference

- Used tables, graphs and numerical measures of summary to describe samples.
- Often wish to apply conclusions reached from the sample to the population from which sample drawn.
- When use samples to make decisions about population, element of risk involved due to uncertainty.

Probability

- Probability is the link between descriptive statistics and inferential statistics.
- There are three approaches to assigning a probability to an event:
  - A priori classical probability
  - Empirical classical probability
  - Subjective probability

A Priori Classical Probability

- The probability of success is based on prior knowledge of the process involved
- Probability of occurrence = $\frac{X}{T}$

  (X = number of ways in which the event occurs and T = total number of possible outcomes)

Empirical Classical Probability

- The outcomes are based on observed data, not on prior knowledge of a process.

Subjective Probability

- A subjective probability differs from person to person.

Terminology

- **Random experiments** Any process or activity which results in an outcome that cannot be predicted with certainty, i.e., uncertain outcomes
- **Sample space S**: list of all possible outcomes.
- **Simple event:** Each possible individual (basic) outcome.
  The (basic) outcomes in S must be :
  (i) Mutually exclusive
  (ii) Exhaustive
- Event: One or more basic outcomes.
  Usually denoted by capital letters A, B, C, ...

Probability
The likelihood of an event occurring.

Impossible event
An event that cannot occur.

Certain event
An event that will occur.

A priori classical probability
Objective probability, obtained from prior knowledge of the process.

Empirical classical probability
Objective probability, obtained from the relative frequency of occurrence of an event.

Subjective probability
Probability that reflects an individual's belief that an event occurs.