Autumn 2018

26134 Business Statistics

University of Technology Sydney

Learning objectives:

- apply standard statistical tools in various business decision contexts within a professionally responsible framework
- apply appropriate quantitative analytical techniques to qualify, support, select and evaluate data as information for business decision-making
- effectively interpret and communicate results of quantitative analyses for business decision-making
- effectively use a computer-based data analysis package (i.e. Excel) to critically analyse data.

Topics:

- Introduction to types of data
- Descriptive statistics
- Introduction to probability and probability distributions
- Sampling and Sampling Distributions
- Interval estimation
- Hypothesis testing
- Comparisons involving means and proportions
- Linear regression

WEEK ONE

- **Descriptive Statistics (describing data)**
- **Readings: Chapter 1, 2 & 3**

Important concepts in stats:

- Population
  - Can be widely defined
  - Collection of whole population = census
  - E.g. all small businesses
  - Descriptive measure of population = parameter (e.g population mean, population standard deviation and population variance)
- Sample
  - Cheaper and simpler
  - Descriptive measure of sample = statistic (e.g sample mean, sample standard deviation and sample variance)

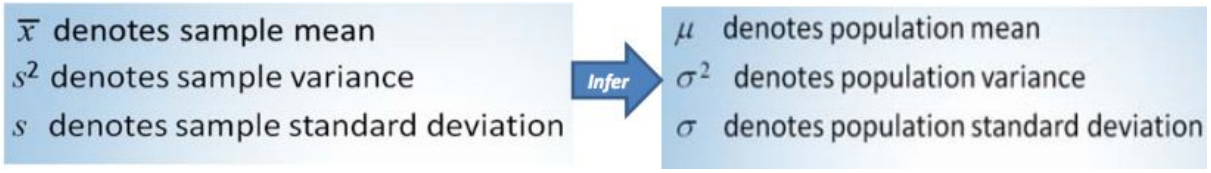
Sometimes, a researcher may want to estimate the value of a parameter, however it is not possible to complete a census due to resources of funds. In this case, the researcher may instead take a representative sample of the population and use the corresponding sample statistic to estimate the population parameter.

Analysing data from a sample can be done in two steps:

- Exploratory data analysis (EDA)
  - Where numerical, tabular and graphical summaries (means, standard deviations) of data are produced to summarise data.
- Statistical inference
  - Uses sample data to reach conclusions about the population
  - An inference based on a probability model linking the data to the population
  - No inference is required for census data, as it collects the whole population

Two branches of statistics:

- Descriptive statistics
  - Tabular, graphical, numerical methods to summarise data
- Inferential statistics
  - Data will often come from a sample, and from that sample we are able to *infer* what would occur at a population level.



Describing Data (Numerical and graphical methods)

One way of describing data is with measures of central tendency. This way provides information regarding the *centre* of a set of numbers. The most common ways of measuring central tendency are the:

- Mode
- Median
- Mean (both population and sample)

Data can be split into multiple categories. These are:

Qualitative / Categorical:

- Nominal
    - Categories that *cannot* be ordered, for example different colours.
    - For example, the numbers on basketball players jerseys are only used to differentiate between each player, with there being no order (ordinal), no meaningful differences (interval) and no meaningful ratios or absolute zero (ratio).
- Ordinal
    - Categories that *can* be ordered, for example, most to least likely.

Quantitative / Numerical:

- Interval
    - Has nominal and ordinal properties, and is a fixed unit of measure. The quantity in difference is meaningful. For example, number of pets owned, temperature.
    - Interval vs ordinal - while interval has the same difference between values, (e.g the difference between 23 degrees and 24 degrees is the same difference between 46 degrees and 47 degrees - 1 degrees) ordinal does not have the same difference between ordered values, where the difference between finishing 1st in a race and 2nd is not necessarily the difference between finishing 2nd and 3rd.
- Ratio
    - Has nominal, ordinal, and interval properties, and the ratio of two values is meaningful. For example, income in dollars or weight in kilos.
    - Ratios also have absolute zero, where zero means the absence of value.
    - Interval vs ratio - In interval, zero does not indicate an absence of the property. Also, in ratio, if you put two 2 kilo weights on a scale, you get 4 kilos. But in interval you can't, for example, add two 20 degree days together to create a 40 degree day. This means the ratio isn't meaningful for temperature.

Measures of location:

- Percentiles:
    - Where data is divided into percentages of the whole
    - For example the 26th percentile is a value that is 26% of the data are equal to or below to value and no more than 74% are above the value
- Quartiles:
    - Where data is divided into four parts
    - Q1, Q2, Q3, Q4

Qualitative / Categorical Data:

Nominal Data

- COUNTIF is a useful function for frequency counting of arbitrary numerical labels.
- These values can then be used in bar graphs or pie charts.
- The types of frequency are:
    - Frequency - number of items
    - Relative frequency - fraction of the total
    - Relative frequency distribution - putting info into tabular summary
    - Percent frequency - relative frequency x 100

Cross-sectional and time-series data:

- Cross-sectional = data collected at a fixed point in time.
    - For example a monthly survey providing information on consumer confidence for the given month.
- Time-series = data collected over time.
    - For example data that consists of consumer confidence over several months or years.

- Unlike cross-sectional, time-series data are time dependent

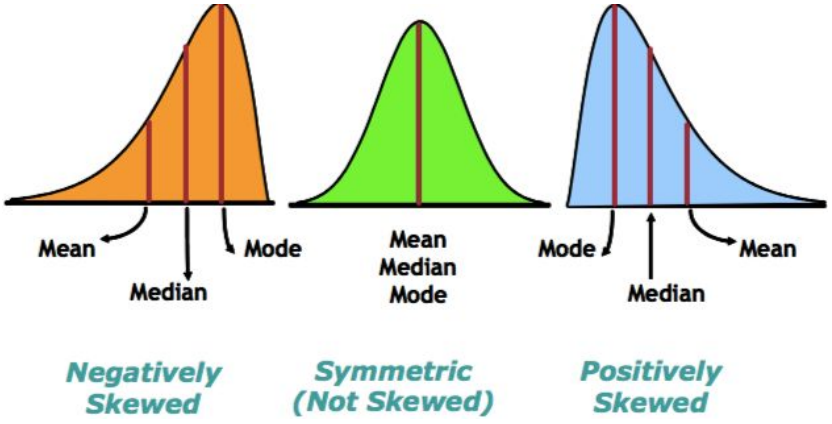
Impact of outliers to a data set:

- Outliers in a data set have the ability to change the mode, median and mean.
- The most drastically impacted thing is the mean, as an outlier will skew the data, placing the average at a point that may not accurately represent the information.

Measures of shape:

Skewness:

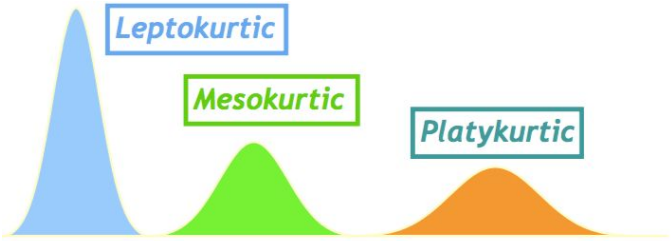
- Can be viewed in a box and whisker plot or a curve as shown below
- *Some examples of how outliers can affect the mean, and thus the skew of a data set:*



Kurtosis:

Kurtosis describes the amount of *peakedness* (tail width and fat tails) of a distribution. The types of peakedness include:

- Leptokurtic (high and thin)
- Mesokurtic (normal in shape)
- Platykurtic (flat and spread out)



Measures of variability:

- Range - the biggest number minus the smallest number
- Interquartile Range - when the data set is divided into four, the middle 50% of the set
- Variance - sigma squared equals the sum of x minus the mean squared divided by n (formula in diagram to the right)

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

- Standard Deviation - the spread of the data
- Coefficient of Variation - a descriptive summary measure that is the *ratio of the standard deviation to the mean expressed as a percentage*. (Expressed by the equation below):

CV for a population:

$$CV = \frac{\sigma}{\mu} * 100\%$$

CV for a sample:

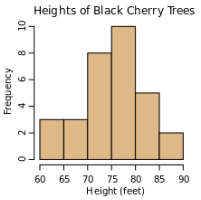
$$CV = \frac{s}{\bar{x}} * 100\%$$

Frequency distributions:

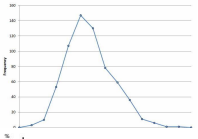
- Class midpoint / class mark = calculated by taking the midpoint or average of midpoints.
- Relative frequency = the ratio of the class interval to the total frequency
- Cumulative frequency = the running total of frequencies through the class of a frequency distribution.

Graphical display of data:

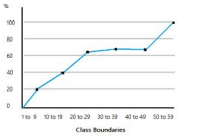
- Histograms
  - Useful for displaying continuous data
  - Can show the shape of the distribution, spread of variability, central location of data and unusual observations like outliers)



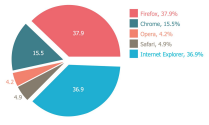
- Frequency polygons
  - Begins with creating a histogram, and then drawing a line between the midpoint of each bar.



- Ogives
  - Cumulative frequency polygon
  - When a researcher wants to view a running total



- Pie chart
  - Used to display categorical data
  - Constructed by finding the angle of each slice - finding the percentage of a category and multiplying it by 360

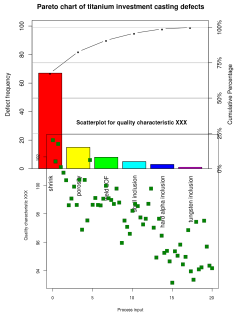


- Stem and leaf plot
  - Original data is preserved on the plot this way
  - Good for continuous data

stem	leaf
0	1, 1, 2, 2, 3, 4, 4, 4, 4, 5, 8
1	0, 0, 0, 1, 1, 3, 7, 9
2	5, 5, 7, 7, 8, 8, 9, 9
3	0, 1, 1, 1, 2, 2, 2, 4, 5
4	0, 4, 8, 9
5	2, 6, 7, 7, 8
6	3, 6

Key: 6|3 = 63 years old

- Pareto charts
  - A graphical way of displaying causes of problems
  - A quantitative tally of the number and types of defects that occur

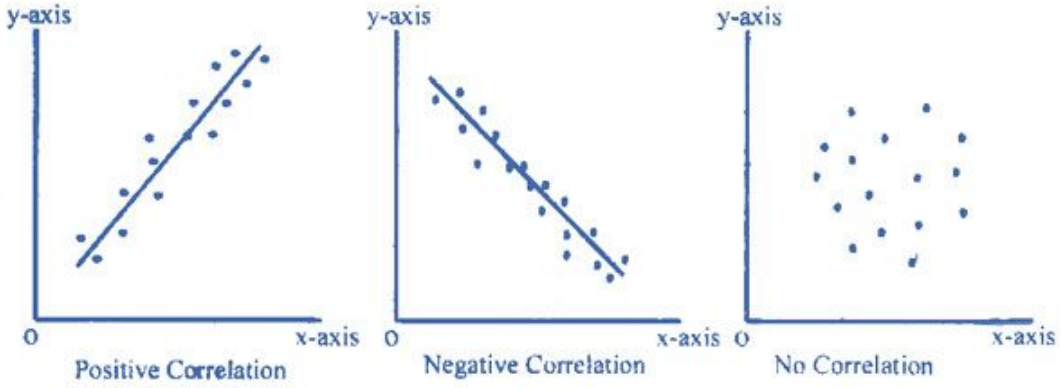


- Scatter plots
  - Quantified display of the relationship between two or more continuous variables
  - Often achieved by regression analysis

Measures of association:

- Correlation
  - A measure of the degree f the relatedness of variables
  - Researchers often want to calculate the population coefficient of correlation

- The term r is a measure of the linear association between two variables, and is a numerical value between -1 and 1.



Key Terms:

- Categorical data
- Census
- Cross-sectional data
- Exploratory data analysis
- Numerical data
- Parameter
- Population
- Primary data
- Sample
- Secondary data
- Statistic
- Statistical inference
- Time-series data
- Class mark
- Class midpoint
- Cumulative frequency
- Frequency distribution
- Frequency polygon
- Grouped data
- Histogram
- Ogive
- Outlier
- Pareto chart
- Pie chart
- Range
- Raw data
- Relative frequency
- Scatter plot
- Stem and leaf plot
- Ungrouped data