Statistics and Analysis: Descriptive and Inferential Analysis

Dr. Elaina A. Hyde

Analysing data using statistics

Introduction

'Statistics is a body of methods and theory that is applied to quantitative data' (Collis and Hussey, 2014, p. 226) so you will need to quantify any qualitative data

Two main branches

Descriptive statistics are a group of statistical methods used to summarize, describe or display quantitative data (may be sufficient for an undergraduate research project

Inferential statistics are a group of statistical methods used to draw conclusions about a population from quantitative data relating to a random sample

A statistic is a number that describes a sample

A parameter is a number that describes a population

Statistics

Statistics - a set of concepts, rules, and procedures that help us to:

-organize numerical information in the form of tables, graphs, and charts;

-understand statistical techniques underlying decisions that affect our lives and well-being;

-make informed decisions.

Descriptive vs Inferential

- Descriptive statistics are used in a univariate analysis (single variable) to examine frequency distributions, measure central tendency and dispersion of the data, and to conduct normality tests
  - At higher levels, you will go beyond this exploratory analysis and conduct bivariate and multivariate analyses using inferential statistics
- Inferential statistics are 'a group of statistical methods and models used to draw conclusions about a population from quantitative data relating to a random sample' (Collis and Hussey, 2014, p. 261

Example Research design

Research question

-What are the factors that have a significant influence ?

Methodology

-Survey

Method of data collection

-Postal questionnaire with one reminder

Method of data analysis

-Descriptive and inferential statistics

The research data

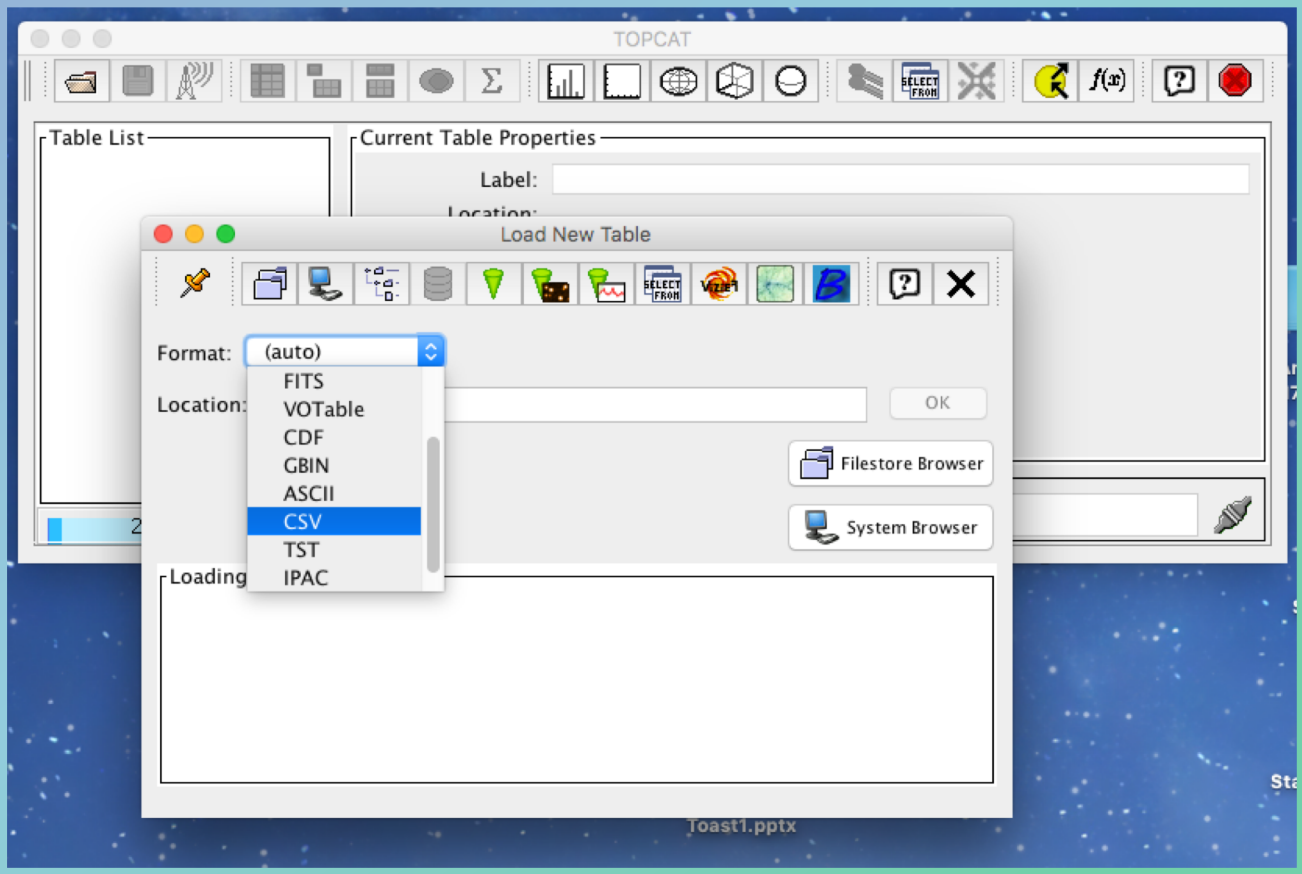
Statisticians can calculate statistics using formulae, but we can use Python, Excell, IBM SPSS or TopCat to analyse research data

TopCat:

FREE analysis software from:

http://www.star.bris.ac.uk/~mbt/topcat/

Topcat – view ASCII, CSV data



Simple Frequency Distribution

A frequency distribution shows us a summarized grouping of data divided into mutually exclusive classes and the number of occurrences in a class. It is a way of showing unorganized data e.g. to show results of an election, income of people for a certain region, sales of a product within a certain period, student loan amounts of graduates, etc. Some of the graphs that can be used with frequency distributions are histograms, line charts, bar charts and pie charts. Frequency distributions are used for both qualitative and quantitative data.

Simple Frequency Distribution

These are the numbers of newspapers sold at a local shop over the last 10 days:

22, 20, 18, 23, 20, 25, 22, 20, 18, 20

Count how many of each number :

Papers Sold	*Frequency*
18	*2*
19	*0*
20	*4*
21	*0*
22	*2*
23	*1*
24	*0*
25	*1*

Binning by 5

It is also possible to group the values. Here they are grouped in 5s:

Papers Sold	Frequency
15-19	2
20-24	7
25-29	1

Frequency distribution

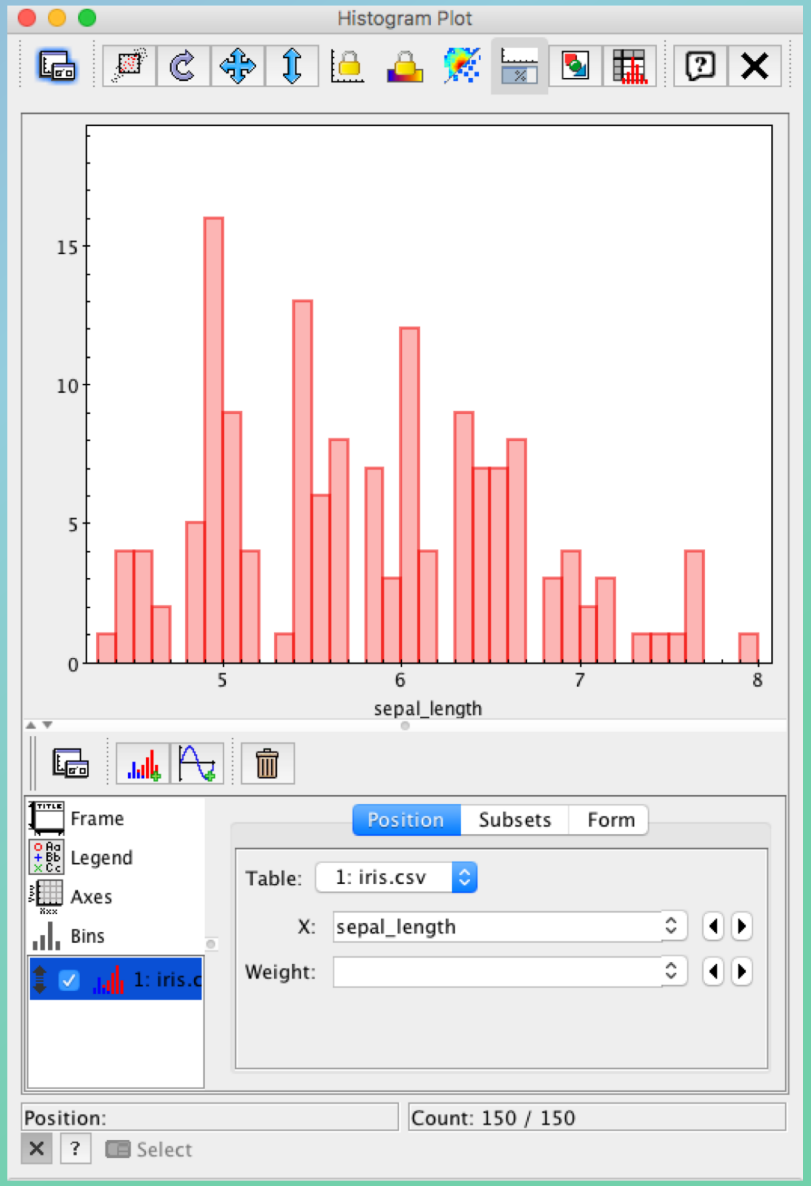
- A frequency is the number of observations for a particular data value in a variable (Collis and Hussey, 2014, p. 235)
  - A frequency distribution is an array that summarizes the frequencies for all the data values in a particular variable
  - A percentage frequency distribution is a descriptive statistic that summarizes a frequency as a proportion of 100
- Eg The survey found that 633 companies out of 790 in the sample had a turnover of less than £1m
  - Percentage frequency = $\frac{633}{790} \times 100 = 80\%$

Iris Data

- Open Source flower data, read more here: https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html

Frequency Flowers

- Generate a frequency histogram in TopCat by clicking histogram. Look at the different distributions



Measures of Center

Several statistics can be used to represent the "center" of a distribution. These statistics are commonly referred to as measures of central tendency.

Measures of central tendency

Module 1	Module 2	Module 3	Module 4	Module 5	Module 6
82%	78%	80%	64%	70%	64%

- Supposing these 6 marks were your exam results
  - The mean ($\bar{x}$) is the arithmetic average: $\frac{438}{6} = 73\%$
  - The median (M) is the mid-value of the data values arranged in size order: 64% 64% 70% 78% 80% 82% so between 70% and 80% = $\frac{70 + 78}{2} = 74\%$
  - The mode (m) is the most frequently occurring value = 64%

Mode

The mode of a distribution is simply defined as the most frequent or common score in the distribution. The mode is the point or value of X that corresponds to the highest point on the distribution.

If the highest frequency is shared by more than one value, the distribution is said to be multimodal. It is not uncommon to see distributions that are bimodal reflecting peaks in scoring at two different points in the distribution.

Median

The median is the score that divides the distribution into halves; half of the scores are above the median and half are below it when the data are arranged in numerical order. The median is also referred to as the score at the 50th percentile in the distribution.

Median

The median location of N numbers can be found by the formula (N + 1) / 2. In the distribution of numbers (3 1 5 4 9 9 8) the median location is (7 + 1) / 2 = 4.

Odd N: When applied to the ordered distribution (1 3 4 5 8 9 9), the value 5 is the median, three scores are above 5 and three are below 5.

Even N: If there were only 6 values (1 3 4 5 8 9), the median location is (6 + 1) / 2 = 3.5. In this case the median is half-way between the 3rd and 4th scores (4 and 5) or 4.5.

Mean

The mean is the most common measure of central tendency and the one that can be mathematically manipulated. It is defined as the average of a distribution is equal to the SX / N.

Or a sample mean of:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x$$

Mean is computed by summing all the scores in the distribution (SX) and dividing that sum by the total number of scores (N). The mean is the balance point in a distribution such that if you subtract each value in the distribution from the mean and sum all of these deviation scores, the result will be zero.

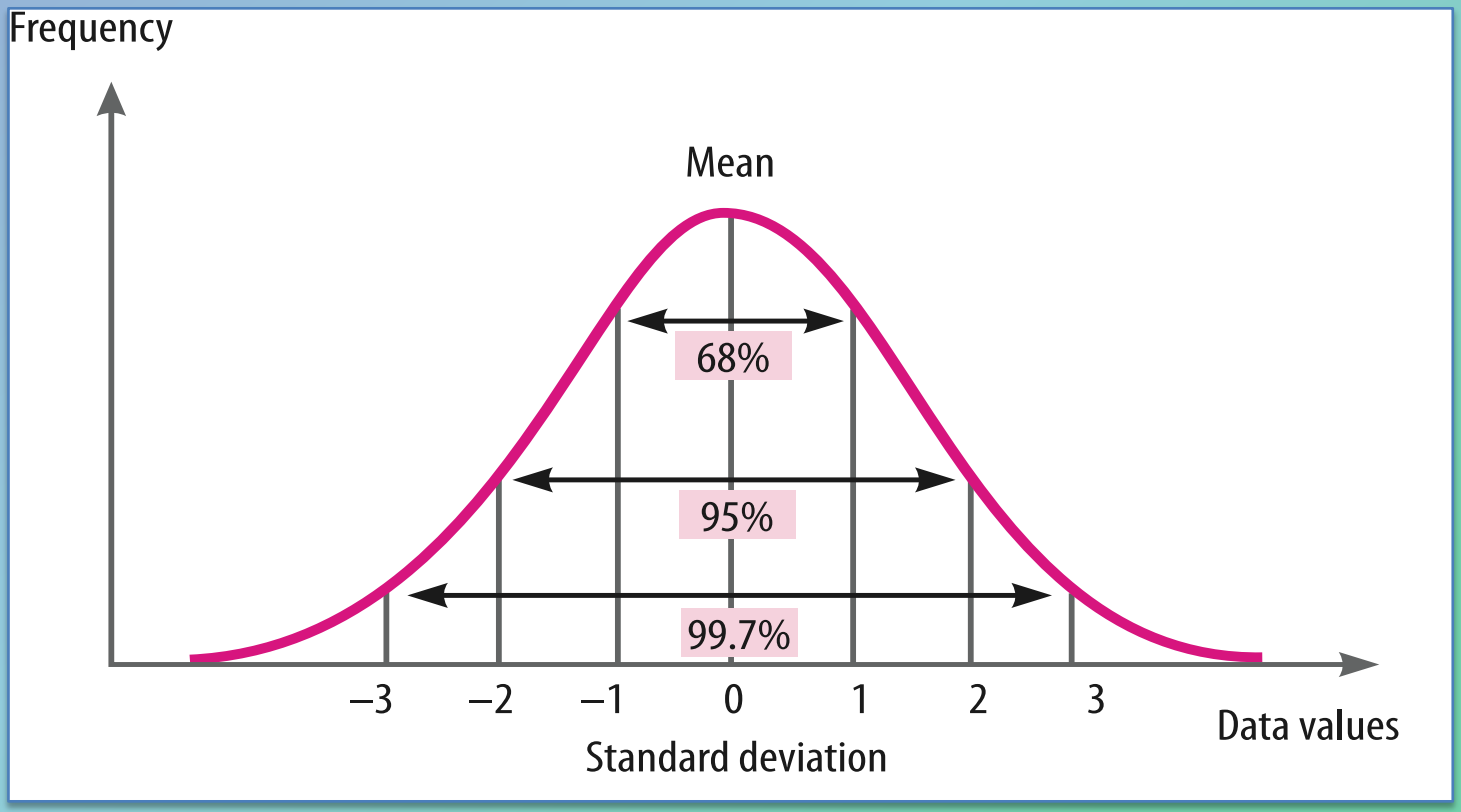
Measures of dispersion

- Measures of dispersion should only be calculated for ratio or interval variables
  - The range represents the difference between the maximum value (the upper extreme or E_u) and the minimum value (the lower extreme or E_L) in a frequency distribution arranged in size order (Range = E_u - E_L)
  - The interquartile range represents the difference between the upper quartile (Q_3) and the lower quartile (Q_1) which is the spread of the middle 50% of a frequency distribution arranged in size order (Interquartile range = $Q_3 – Q_1$)
- But neither takes account of all the data values

Measures of dispersion (continued)

- The standard deviation (sd, stdv) takes account of all the data values
  - It is based on the error and the variance, which are two statistical models used to measure how well the mean represents the data
  - In this context, the error is the difference between the mean and the data value (observation) and the variance is the average error between the mean and the data
- The standard error (se) is the standard deviation between the means of different samples
  - A large standard error relative to the sample mean suggests the sample might not be representative of the population

Normal distribution



Range

The simplest measure of variability to compute and understand is the range. The range is the difference between the highest and lowest score in a distribution. Although it is easy to compute, it is not often used as the sole measure of variability due to its instability. Because it is based solely on the most extreme scores in the distribution and does not fully reflect the pattern of variation within a distribution.

Interquartile Range (IQR)

Provides a measure of the spread of the middle 50% of the scores. The IQR is defined as the 75^{th} percentile to the 25^{th} percentile. The interquartile range plays an important role in the graphical method known as the **boxplot**.

It is easy to compute and extreme scores in the distribution have much less impact but it suffers as a measure of variability because it discards too much data. Researchers want to study variability while eliminating scores that are likely to be accidents.

Variance

The variance is a measure based on the deviations of individual scores from the mean. As noted in the definition of the mean, however, simply summing the deviations will result in a value of 0. To get around this problem the variance is based on squared deviations of scores about the mean.

The sample variance is then:

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

When the deviations are squared, the rank order and relative distance of scores in the distribution is preserved while negative values are eliminated. Then to control for the number of subjects in the distribution, the sum of the squared deviations, S(X - `X), is divided by N (population) or by N - 1 (sample). The result is the average of the sum of the squared deviations and it is called the variance.

Standard deviation

- The standard deviation (sd, stdv) is the square root of the variance
  - A large standard deviation relative to the mean suggests the mean does not represent the data well
- The standard deviation is related to a theoretical frequency distribution known as the normal distribution
  - It is bell-shaped and symmetrical and has tails extending indefinitely either side of the centre
  - The mean, median and mode coincide at the centre
  - 68% of the data will fall within 1 sd of the mean, 95% will fall within 2 sd and 99.7% will fall within 3 sd of the mean

Standard deviation

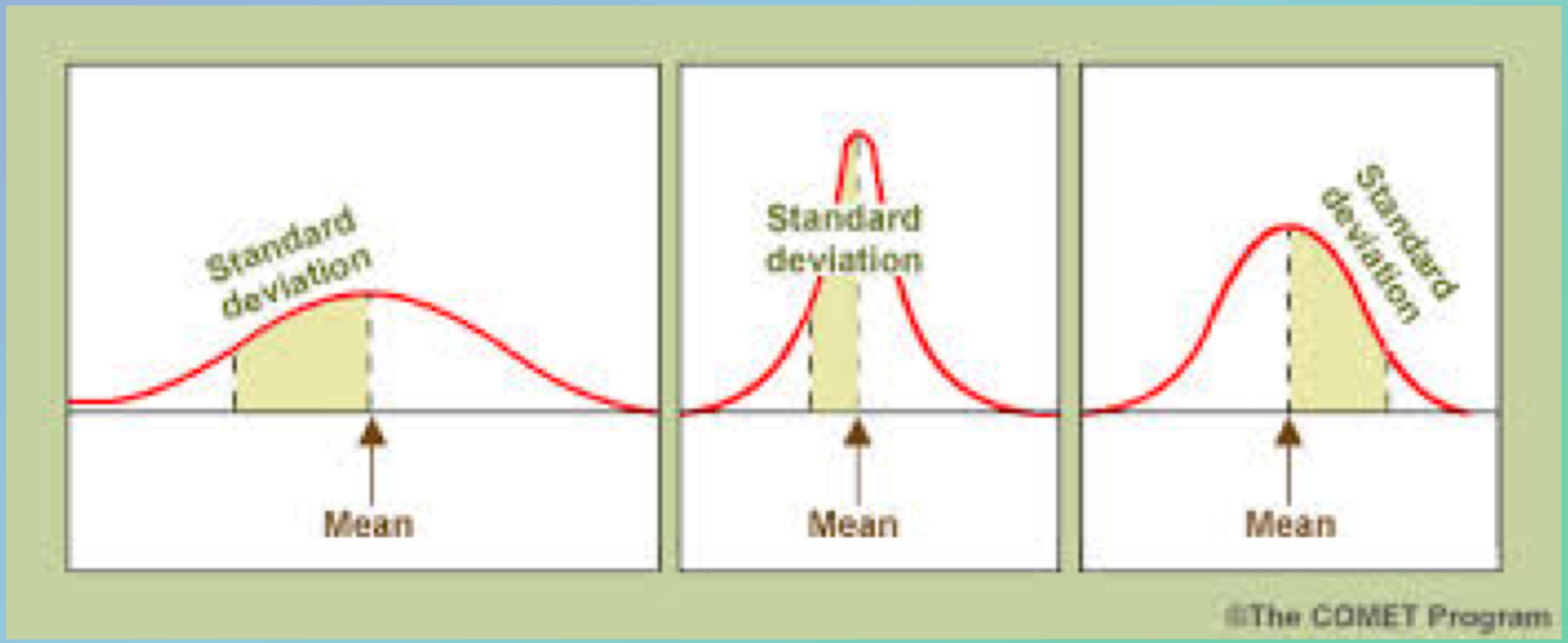
The standard deviation (**s** or **stdv**) is defined as the **positive square root** of the variance. The variance is a measure in squared units and has little meaning with respect to the data. Thus, the standard deviation is a measure of variability expressed in the same units as the data.

The sample stdv is:

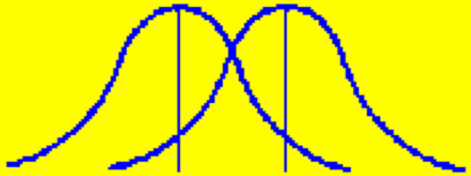
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2}$$

The standard deviation is very much like a mean or an "average" of these deviations. In a normal (symmetric and mound-shaped) distribution, about two-thirds of the scores fall between +1 and -1 standard deviations from the mean and the standard deviation is approximately 1/4 of the range in small samples (N < 30) and 1/5 to 1/6 of the range in large samples (N > 100).

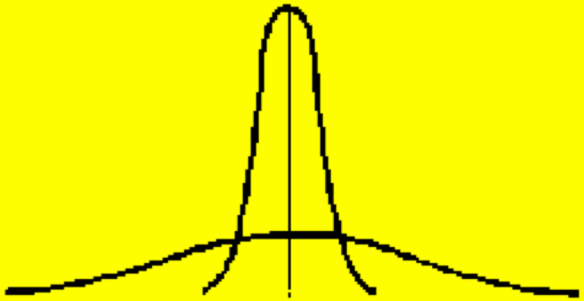
Standard deviation



Different Means
Same Standard Deviation



Same Mean
Different Standard Deviations

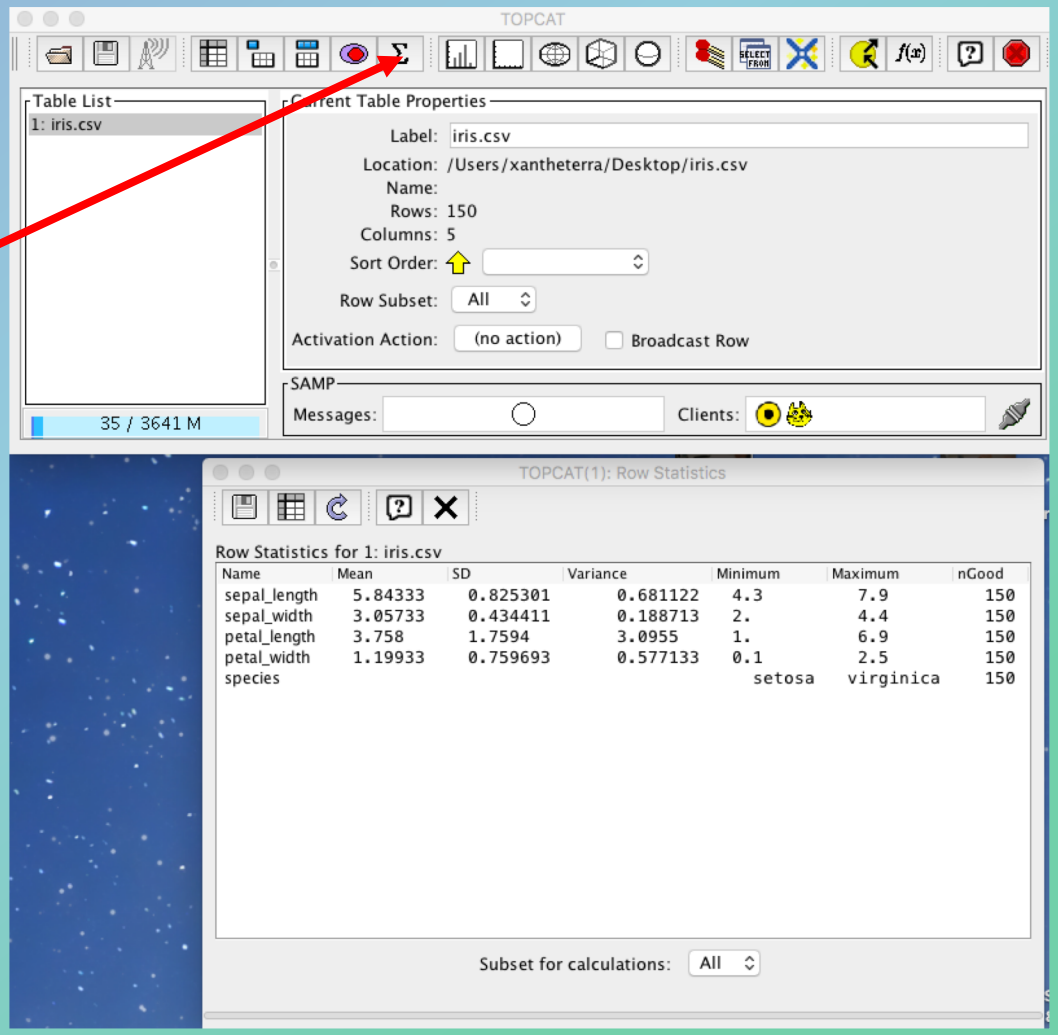


Different Means
Different Standard Deviations

TopCat Statistics

Let's look at the statistics in TopCat and compare it to the histograms of the variables we have.

In the view menu you can select different statistics.



Measures of normality

- A normal distribution is a theoretical frequency distribution – it is a mathematical model representing perfect symmetry, against which empirical data can be compared. If your data is supposed to take parametric stats you should check that the distributions are approximately normal.
- The best way to do this is to check the skew and Kurtosis measures from the frequency output from SPSS. For a relatively normal distribution:
- skew ~= 1.0
- kurtosis~=1.0
- If a distribution deviates markedly from normality then you take the risk that the statistic will be inaccurate. The safest thing to do is to use an equivalent non-parametric statistic.

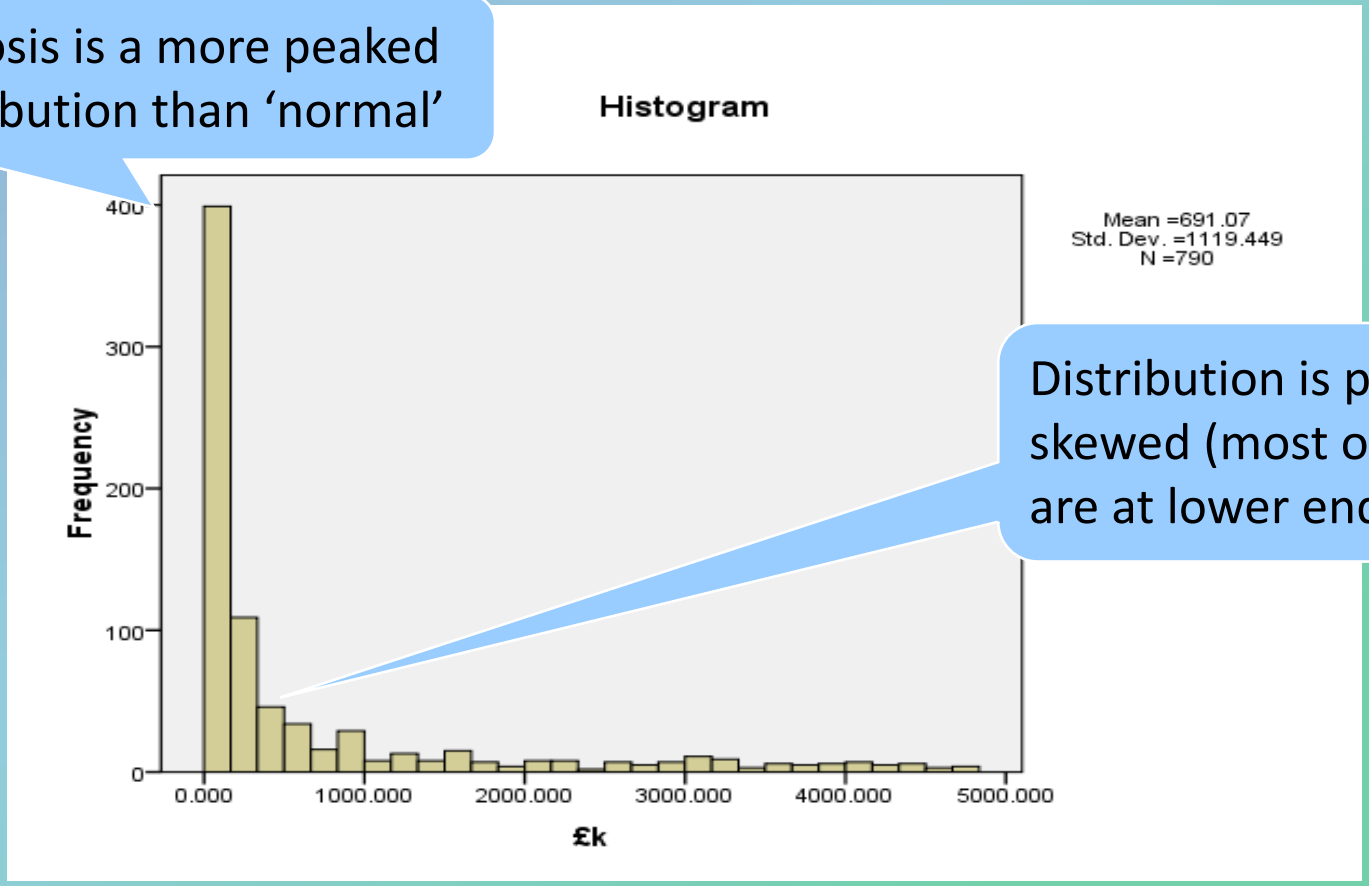
Skew & Kurtosis

A positive skew (greater than 1) indicates a distribution that has a positive tail greater than a normal distribution, i.e. the peak is more towards lower values & mean is greater than median

A positive Kurtosis (greater than 1) indicates a distribution that is more peaked than a normal distribution

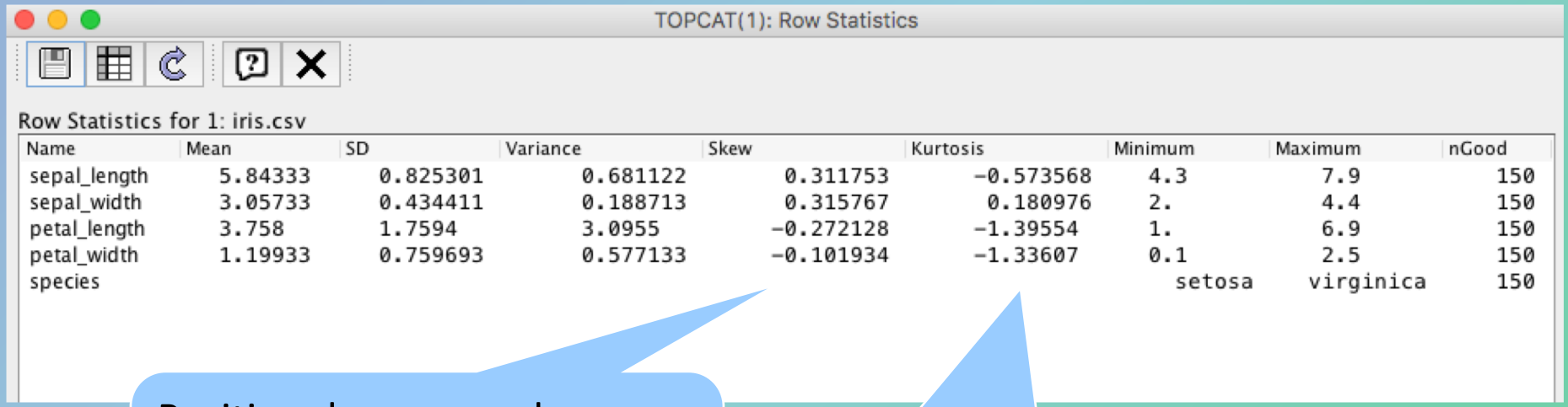
SPSS output
Histogram for turnover

Kurtosis is a more peaked distribution than 'normal'



Distribution is positively skewed (most observations are at lower end of the range)

TopCat output
Descriptive statistics



Positive skewness value indicates most observations are at lower end of the range

Positive kurtosis value indicates a more peaked distribution than ‘normal’

Mean Median Mode

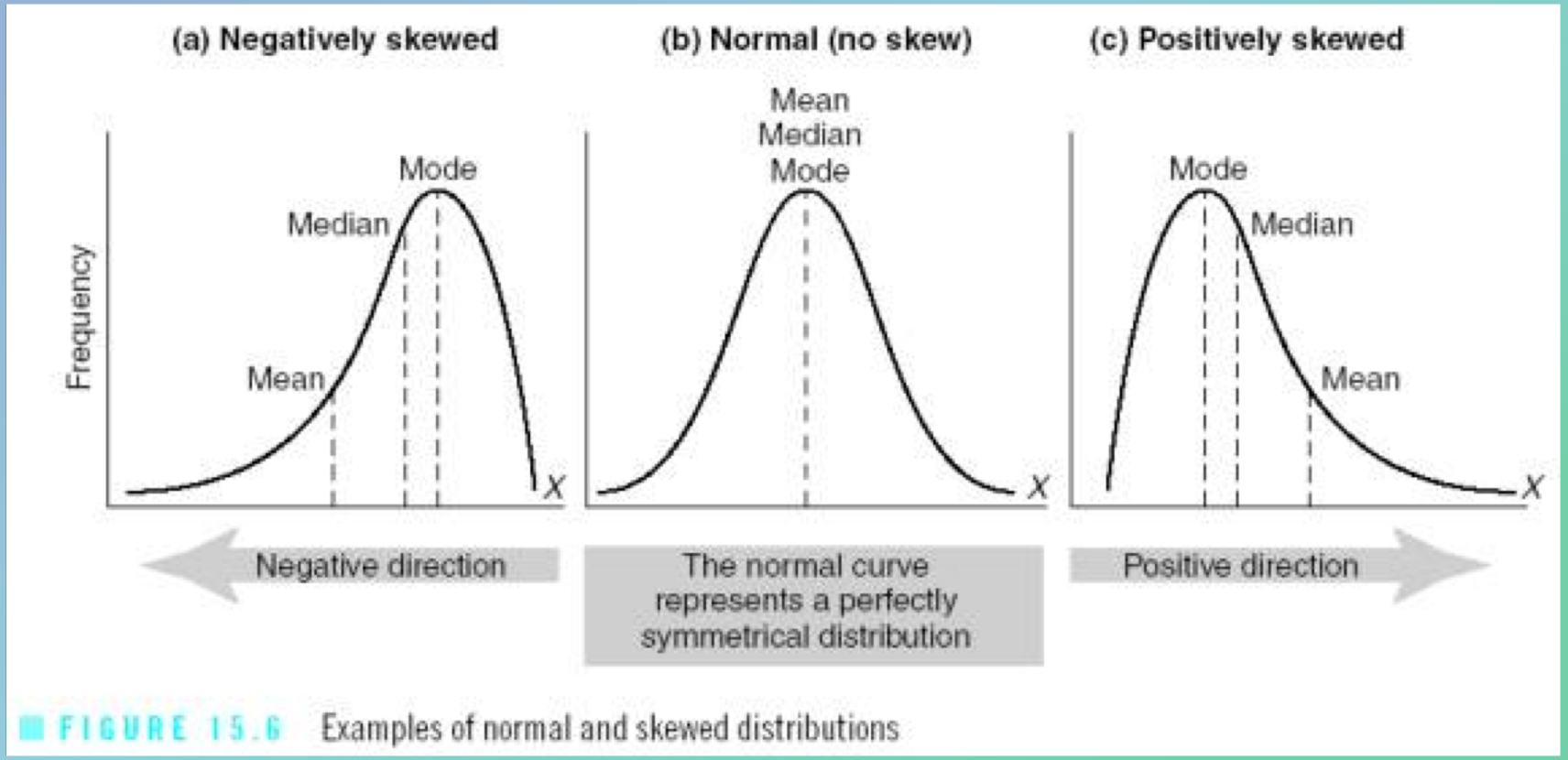
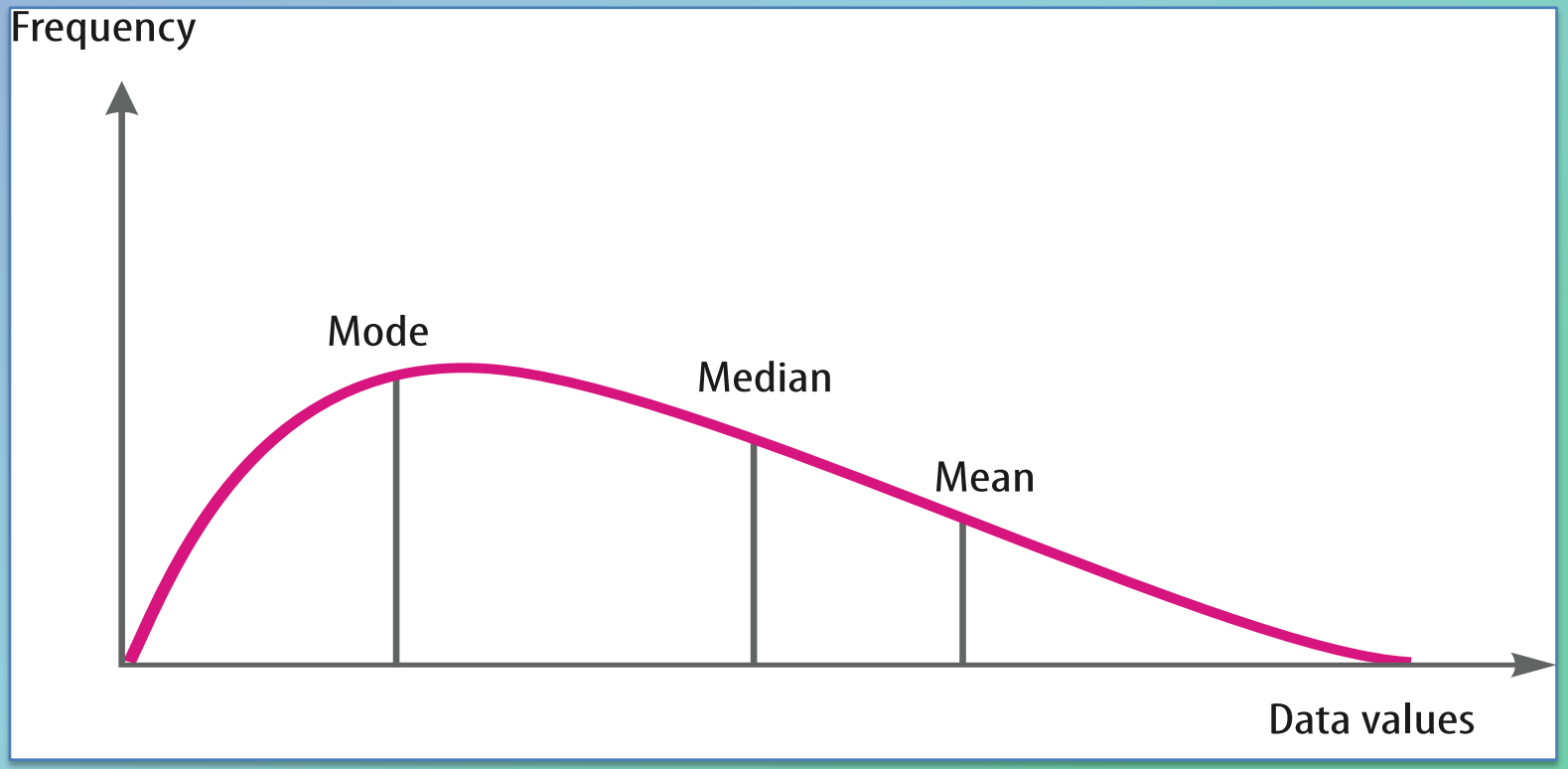
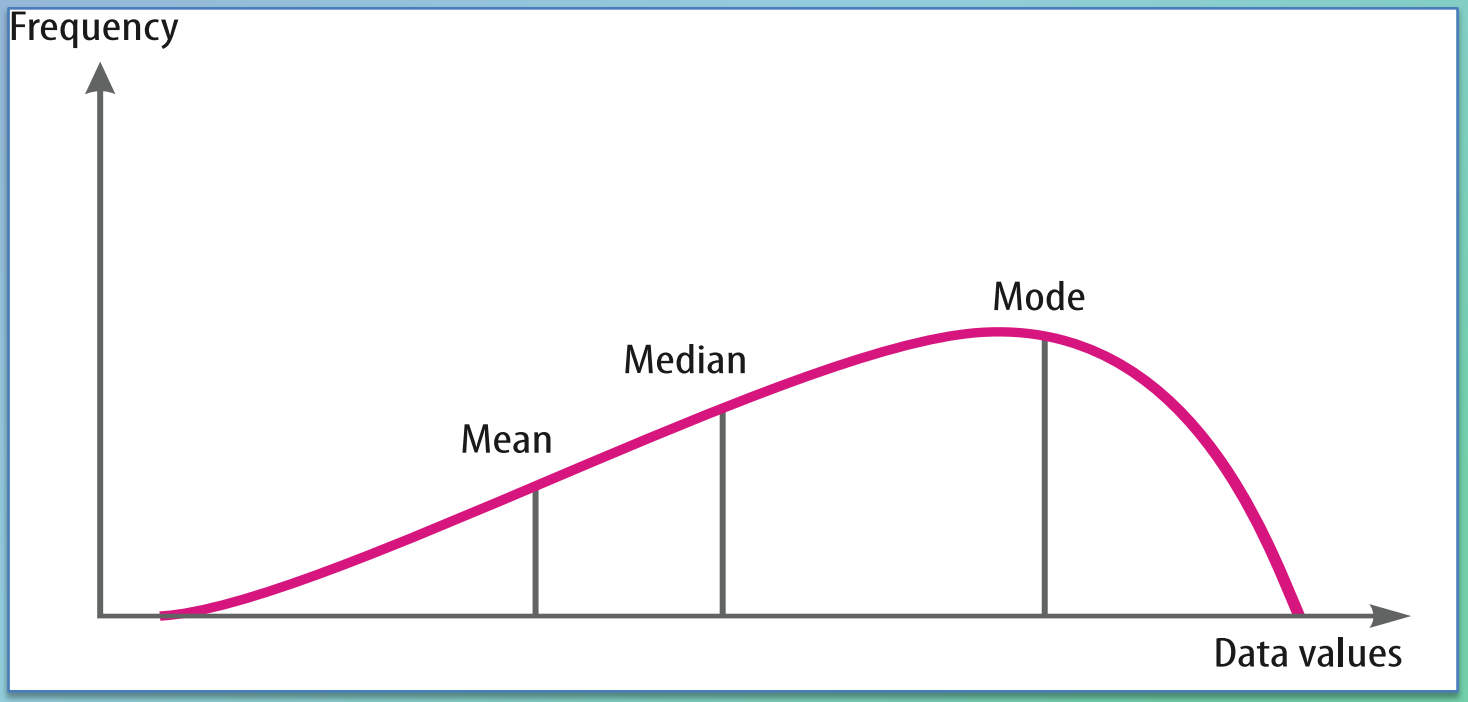


FIGURE 15.6 Examples of normal and skewed distributions

A positively skewed frequency distribution



A negatively skewed frequency distribution



- The measurement level of the variable determines which descriptive statistics are appropriates
- Your choice depends on your research questions, which may also require the use of inferential statistics

Exploratory analysis	Descriptive statistics	Measurement level
Frequency distribution	Percentage frequency	Ratio, interval, ordinal, nominal
Measures of central tendency	Mean Median Mode	Ratio, interval Ratio, interval, ordinal Ratio, interval, ordinal, nominal
Measures of dispersion	Range Standard deviation	Ratio, interval Ratio, interval
Measures of normality	Skewness Kurtosis	Ratio, interval Ratio, interval

Univariate analysis using descriptive statistics

- Use descriptive statistics to conduct a univariate analysis to explore data from individual variables

Descriptive statistics	
Frequency distribution	Percentage frequency
Measures of central tendency	Mean Median Mode
Measures of dispersion	Range Standard deviation
Measures of normality	Skewness Kurtosis

Inferential Statistics – a different kind of testing

Stating the objectives

- The analysis is guided by the hypotheses you developed from the theoretical framework
  - A hypothesis is a proposition that can be tested for association or causality against empirical evidence (data based on observation or experience, eg survey data)
- Each hypothesis is formulated as a statement about a relationship between two variables and can be expressed in the null or the alternative form

The null and the alternative hypothesis

- The null hypothesis (H_0) states that the variables are independent of one another (ie there is no association)
- The alternative hypothesis (H_1) states that the variables are associated (ie there is an association)
  - H_0 is the default
  - H_1 is accepted only if the test result provides significant evidence to reject H_0
- If you predict the IV has an effect on the DV in a particular direction, it is known as a one-tailed hypothesis
  - A two-tailed hypothesis is where you cannot predict the direction
- Collis (2003) tested 9 one-tailed hypotheses

Population parameters

- A random sample is needed to obtain estimates of theoretical population parameters
  - A parameter is a number that describes a population whereas a statistic is a number that describes a sample
- Inferential statistics include parametric and non-parametric tests, and you need to examine your population to determine whether parametric tests are appropriate
  - Parametric tests make certain assumptions about the distributional characteristics of the population under investigation

Parametric tests

- To use parametric tests, four basic assumptions about the research data must be met (Field, 2000)
  1. The variable is measured on a ratio or interval scale
  2. The data are from a population with a normal distribution
  3. There is homogeneity of variance (variances are stable in a test across groups of subjects, or the variance of one variable is stable at all levels in a test against another variable)
  4. The data values in the variable are independent (they come from different cases, or the behaviour of one subject does not influence the behaviour of another)

Non-parametric tests

- The reason why these assumptions are so important is that the calculations that underpin parametric tests are based on the mean of the data values
- However, non-parametric tests do not rely on the data meeting these assumptions because the statistical software arranges the frequencies in size order and performs the calculations on the ranks rather than the data values
- Non-parametric tests must be used for
  - Variables measured on a ratio or interval scale that do not have a normal distribution (eg TURNOVER)
  - All ordinal or nominal variables (eg CHECK, QUALITY, CREDIBILITY, CREDITSCORE, FAMILY, EDUCATION)

Parametric vs Non-Parametric

- The basic distinction for parameteric versus non-parametric is:
- If your measurement scale is nominal or ordinal then you use non-parametric statistics
- If you are using interval or ratio scales you use parametric statistics.

Bivariate analysis

- A bivariate analysis tests data from two variables
- Can examine a hypothesised relationship between a measured variable and a variable as suggested by a theoretical framework

Mann-Whitney test of difference

- Establishes whether there is a difference between two independent samples
- Nonparametric test of the null hypothesis that:
  - it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample.

SPSS output file: Mann-Whitney test

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Turnover £k is the same across categories of Q3.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
2	The distribution of Q4a is the same across categories of Q3.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
3	The distribution of Q4b is the same across categories of Q3.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
4	The distribution of Q4c is the same across categories of Q3.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
5	The distribution of Q4d is the same across categories of Q3.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

- The results show that the significance values (Sig.) are ≤ 0.01 for each of the five tests
- As our hypotheses were one-tailed (they predicted the direction of the relationship), and these are shown for a two-tailed hypothesis , we need to divide them by 2
- The outcome is unchanged with a very high level of significance and we have evidence to reject the null hypothesis for this test in respect of TURNOVER, CHECK, QUALITY, CREDIBILITY and CREDITSCORE

Chi-square (χ^2) test of association

- The χ^2 test measures the association between the two groups in the DV (companies that would have a voluntary audit and those that would not) and each dichotomous nominal IV (eg the dummy variables)

SPSS chi-square test

FAMILY * Volaudit **Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	33.103[a]	1	.000		
Continuity Correction[b]	32.212	1	.000		
Likelihood Ratio	33.031	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	33.060	1	.000		
N of Valid Cases	767				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 107.12.

b. Computed only for a 2x2 table

- The notes confirm the assumptions of the chi-square test are met, but we need to divide the significance value (Sig.) by 2 for our one-tailed hypothesis and the outcome is unchanged ($p \leq 0.01$)
- Looking at all the results, we have evidence to reject the null hypotheses for FAMILY, EXOWNERS and BANK, but not EDUCATION

Bivariate and multivariate analysis

Purpose	For parametric data	For non-parametric data
Tests of difference for independent or dependent samples	*t*-test	Mann-Whitney test
Tests of association between two nominal variables	Not applicable	Chi-square test
Tests of association between two quantitative variables	Pearson's correlation	Spearman's correlation
Predicting an outcome from one or more variables	Linear regression	Logistic regression

Correlation

- Correlation is a measure of the direction and strength of association between two quantitative variables. Correlation may be linear or non-linear, positive or negative' (Collis and Hussey, 2014, p. 270)
- Most statistics try to fit straight-line models to the data and the correlation coefficient measures the linear dependency of the two variables
  - +1 represents perfect positive linear association (both variables increase together)
  - 0 represents no linear association
  - -1 represents perfect negative linear correlation (one variable increases as the other decreases)

Factor analysis

- Factor analysis is used to examine the correlation between pairs of variables measured on a rating scale (for example a Likert scale)
  - The analysis identifies sets of interrelated variables on the basis that each variable in the set could be measuring a different aspect of some underlying factor (Field, 2000)
  - The resulting factor scores represent the relative importance of the variables to each factor and can be used in a subsequent linear regression analysis

Multicollinearity

- Multicollinearity occurs when there is very high correlation between the IVs, which presents a problem because the 'overlap' in their predictive power makes it hard to identify their separate effects on the DV
- If none of the correlation coefficients is higher than 0.7, the strength of correlation is not a problem
- If you find evidence of multicollinearity, exclude the variable with less theoretical importance to the research

(Kervin, 1992)

Linear vs Logistic regression

- Linear regression is a measure of the ability of an IV to predict an outcome in a DV where there is a linear relationship between them

- Logistic regression is used where the DV is a dummy variable and one or more of the IVs are continuous quantitative variables (others can be ordinal or dummy variables)

- Reminder: A dummy variable is typically a 0,1 Boolean

Time series analysis

- Time series analysis is a statistical technique for forecasting future events from time series data
  - A time series is a sequence of measurements of a variable taken at regular intervals over time
- The purpose of time series analysis is to examine the trend and any seasonal variation, both of which can be further analysed using linear regression
  - A trend is a consistently upward or downward movement in time series data
  - Seasonal variation is where a pattern in the movement of time series data repeats itself at regular intervals

Setting the significance level

- When using a statistical test, we want to be sure that the effect genuinely exists, but there are two cases when a test result leads to an incorrect result (an error)
  - H0 is true, but the test leads to its rejection (a Type I error)
  - H1 is true, but the test leads to acceptance of H0 (a Type II error)
- We specify the critical region that determines whether a test result is significant by setting the significance level
  - If the default is 0.05 we are accepting a 5% probability that the test will lead to a Type I or Type II error, and we can be 95% certain that the effect exists

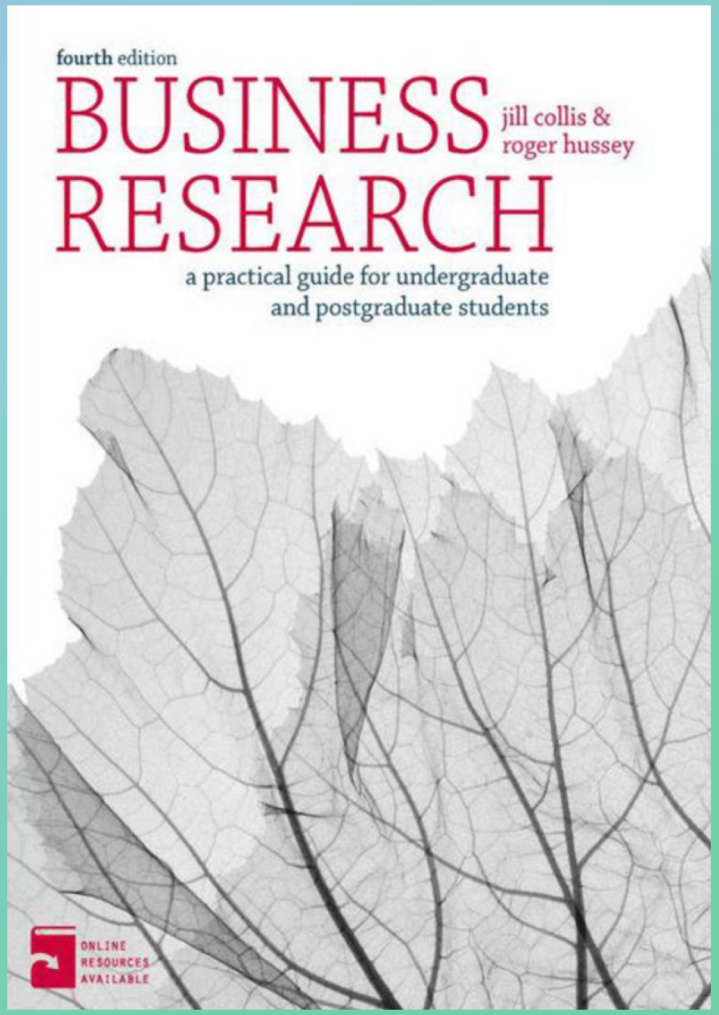
Non-Parametric Summary

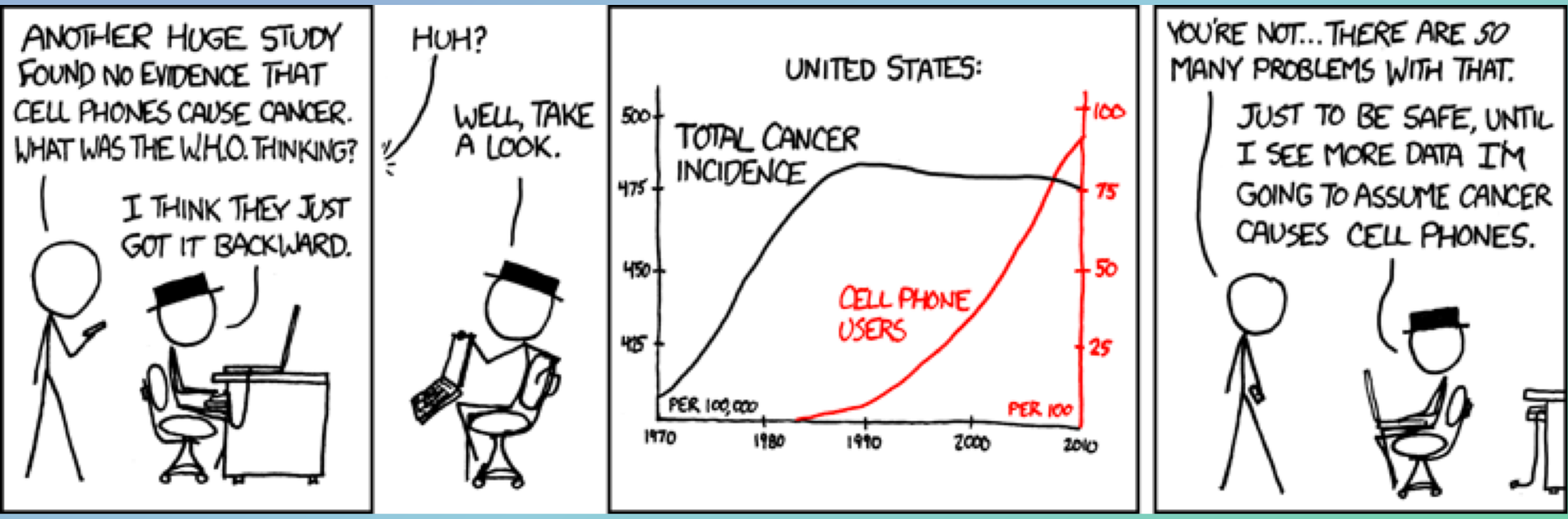
- The non-parametric tests described in this session were used to test hypotheses relating to a particular research question with a view to generalization
  - Drawing conclusions about a population from data collected from a random sample
  - If your research data meets the four basis assumptions we have discussed, you can use parametric tests
- If you have longitudinal research data, you can use time series analysis to examine the trend and evaluate any cyclical or irregular variation

Further Reading

Business Research: A Practical Guide for Undergraduate and Postgraduate Students

Jill Collis
Roger Hussey





Credit: https://xkcd.com/925/