Slide 1

Email: jkanglim@unimelb.edu.au
Office: Room 1110 Redmond Barry Building
Website: http://jeromyanglim.googlepages.com
Appointments: For appointments regarding the course or with the application of statistics to your thesis, just send me an email

Descriptive Statistics

325-711 Research Methods

2007

Lecturer: Jeromy Anglim

"Conducting data analysis is like drinking a fine wine. It is important to swirl and sniff the wine, to unpack the complex bouquet and to appreciate the experience." (Wright, 2003)

-Wright, D.B. (2003) Making friends with your data: Improving how statistics are conducted and reported. British Journal of Educational Psychology, 73(Mar), 123- 136.

DESCRIPTION:

This seminar will explore the role that descriptive statistics play in helping the researcher get a feel for the simple and complex relationships that will come out in the data set. We will review concepts of frequencies, measures of central tendency, and dispersion. Preliminary issues of data screening, missing data, and outliers will be discussed. Data analysis will be presented as a process of reasoned argument built on research questions. Important contemporary issues in data analysis will be discussed including effect size, confidence intervals, power analysis, the accuracy in parameter estimation approach, and null hypothesis significance testing. A common theme will be analysing research within a meta-analytic framework.

Slide 2

Semester 2 Seminars

Topic	Presenter	Date
Quant analysis 1: descriptive & univariate analysis	Jeromy Anglim	27 July
Quant analysis 2: (M)ANOVA, covariates	Jeromy Anglim	3 August
Quant analysis 3: cluster & factor analysis	Jeromy Anglim	10 August
Quant analysis 4: linear/logistic regression	Jeromy Anglim	17 August
Quant analysis 5: moderators and mediators	Jeromy Anglim	24 August
Introduction to SPSS 1	Danielle Chmielewski	31 August
Introduction to SPSS 2	Danielle Chmielewski	7 Sept.
Introduction to SPSS 3	Danielle Chmielewski	14 Sept.
Semester break	No class	
Semester break	No class	
Quant analysis 6 : structural equation modeling	Jeromy Anglim	5 Oct.
International research	Anne-Wil Harzing, Ying Zhu	12 Oct.
Free		19 Oct.
Free		26 Oct.

Slide 3

Approach of the course

- A modern approach, attitude and philosophy to data analysis
  - Don't prejudge: Listen to what the data is saying
  - Take the time to get to know the data
  - Reasoned decision making
- Extensive Detail, but bringing it back to the big picture – The Research Question

Slide 4

Big Picture

- Statistics
  - Tool for **UNDERSTANDING** the world and developing theory
  - About explaining **VARIABILITY**
  - About **GENERALISING** from a set of observations to the broader population
- Tools for thinking about the world
  - Critically evaluate existing research
  - Conduct and report your own research

Overview of the Session

- Big Picture
- Overview of descriptive statistics
  - Central tendency, spread
  - Distributions
  - Graphs
- The big issues
  - NHST, Power, Effect Size, Confidence Intervals, Accuracy in Parameter Estimation, meta-analytic thinking
- Preliminary analysis issues
  - Data entry, Missing data, outliers, constructing scales
- A little more of the Big Picture

Readings

- Field, A. (2005). Discovering Statistics Using SPSS. London: Sage.
  - Chapter 1: Everything you ever wanted to Know about Statistics (Well Sort Of)
  - Chapter 3: Exploring Data
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). 4th edition. New York: Macmillion Publishing Company.
  - Chapter 1: Introduction
  - Chapter 2: Examining your data
- Additional Readings
  - Wilkinson, L. & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and expectations. *American Psychologist, 54*(8), 594-604.
  - Wright, D.B. (2003) Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology, 73*(Mar), 123- 136.
  - Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect size. *Educational Researcher, 31*(3), 25.
    - Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7(2)*, 147-177.

Field (2005): Particularly chapter 1 is really worth reading. It provides the structure for understanding what we will be doing in future weeks. Field is excellent in his use of analogies, humour, and self-deprecation to make statistical ideas clear.

Hair (2006): Chapter 1 gives a nice overview of the many multivariate procedures that exist and provides context for what we are going to cover in the coming weeks. Chapter 2 gives a thorough introduction to graphics, outliers, missing data, transformations and assumption testing.

Wright (2003): Excellent article that highlights some important current issues in data analysis as well as outlining an appropriate attitudinal and philosophical orientation to data analysis.

Wilkinson, L. & Task Force on Statistical Inference (1999) This review provides many important recommendations from the American Psychological Association on writing up your results for a thesis or journal article. It's written in a very accessible format and should be considered essential reading.

Schafer & Graham (2002): An excellent review of current issues in missing data. If you have missing data in your dataset, this provides an overview of the issues involved. However, you will probably find Chapter 2 in Hair a more accessible starting point.

Thompson (2002): Approach to data analysis that emphasises the use of confidence intervals on effect sizes. An excellent approach to data analysis that in many ways integrates the pros and cons of the two camps of on the one hand, the strong advocators of null hypothesis significance testing, and on the others those who advocate the use of effect size measures.

Slide 7

There is something good for democracy about getting good at data analysis.

Good data analysis skills combined with an open mind to the answers that empirical data will bring represent a way of moving beyond ideology.

We can look at many debates at the moment and see how they could benefit from improved data and improved data analysis.

If we ourselves want to fully participate in these debates in our area of research or in the wider public domain, it is very powerful to be able to intelligently critique the conclusions that others have reached based on analysis of empirical observations.

Of course this requires much more than knowledge of data analysis. It requires knowledge of the substantive theory, research design, and principles of measurement, just to name a few. However, it is an aim of this series to show how data analysis fits into a broader scheme.

My Background

- Jeromy Anglim
- Educational Background
  - Bachelor of Arts (Honours in Psychology)
  - Bachelor of Laws
  - Completing final year of Masters of Industrial/Organisational Psychology and PhD.
- Teaching
  - Lectured 2nd, 3rd and 4th year statistics in psychology
  - Tutored Statistics in psychology and market research
- Statistical Consulting
  - Statistical consultant in market research
    - The Rothcorp Group
  - Statistical consultant – employee opinion surveys
    - OSA; International Survey Research
  - Psychological Test Construction (e.g., selection & recruitment)
    - Psych Press
  - Statistical consultant in Psychology
  - Statistical Consultant in Commerce Faculty
    - Lea Waters; Anne-Wil Harzing; Margaret Abernathy

The right support materials – textbooks

- Textbooks
  - Hair, et al – Multivariate Data Analysis (6th Ed.)
  - Tabachnick & Fiddel – Using Multivariate Statistics
  - Palant – SPSS Survival Manual
  - Field – Discovering Statistics Using SPSS
- The internet
  - Google the technique
  - http://www.ats.ucla.edu/stat/
  - http://www2.chass.ncsu.edu/garson/pa765/statnote.htm
- A statistical friend

Hair et al: A massive and comprehensive book on multivariate statistics written for people doing research in business; includes section on SEM and Confirmatory factor analysis; justifications and references to primary statistics literature are sometimes a little light-on
Tabachnick & Fiddel: Another massive multivariate book; greater coverage of underlying matrix algebra and formulas than Hair; Examples are drawn more from psychology.
Palant: Designed specifically for people going through the process of using SPSS for their thesis; more like a cook-book; If you don't do statistics very often and want a simple easy to apply recipe – this is the book.
Field : This book is awesome. Its funny. It explains the statistical ideas clearly using everyday language without overly sacrificing sophistication of understanding. Covers regression, PCA, different types of ANOVA
Howell: This book is good if you want to get a deeper understanding of statistics including formulas and some more obs
Internet:

http://www2.chass.ncsu.edu/garson/pa765/statnote.htm
This site is comprehensive and also has a nice set of links to other key sites
Basically with a few simple searches you can usually get useful information about almost any technique you are likely to use. You can also usually get it explained in SPSS.
http://www.ats.ucla.edu/stat/
The UCLA site is excellent and comprehensive. It has many great tutorials including videos and worked examples from key textbooks. It also covers all major statistics packages and many niches ones as well.
A statistical friend:
Perhaps your office buddy; your supervisor; a lecturer who know their stats; a statistical consultant. The statistics department also runs a consulting service.
The key to making the most of such people: 1) show that you've thought about the problem and made some initial attempts at answering it by reading up; 2) be prepared, structured and specific with your questions; 3) understand that statistical consultants are there to assist you in understanding the pros and cons of different options . Many statistical decisions can not be outsourced and rely on reasoned decisions that should be made by the researcher.

Slide 10

Today: descriptive statistics and exploratory data analysis

- Your data has a story to tell
- How are you going to coax that story out?
- Will you *understand* what the data has to say? Will you be *open* to what it has to say?
- How will you contextualise your results?
- How will you deal with the fiddly things like missing data, outliers and violations of assumptions?
- How are you approach acquiring the statistical knowledge necessary to analyse your data?

<div style="border:1px solid black; padding:1em;">

Types of variables

- Binary, nominal, ordinal, interval, ratio

- Metric vs nonmetric
  - Metric = interval, ratio
  - Nonmetric = nominal, ordinal
  - Binary can be treated either way

- The decision is sometimes up to us

</div>

Overview

Variables can be categorised into different types. The type of the variable has implications for the type of analysis you perform.

Nominal

Nominal variables are discrete, unordered categories. Examples include race, favourite food, political preference, favourite television show,

Ordinal

Limited number of ordered categories where the relative distance between the categories is not necessarily equal. If you think about the order someone comes in a running race, the difference in completion times between first and second is not necessarily the same as the difference between second and third. With ordinal variables all we know is that a score is higher or lower than another, but not the relative distance between scores.

Examples include 5-point rating scales, rankings in a race. It should be noted that ordinal variables are frequently used in analyses which assume interval data. For example, when we use a 5-point strongly disagree to strongly agree item as the dependent variable in a t-test.
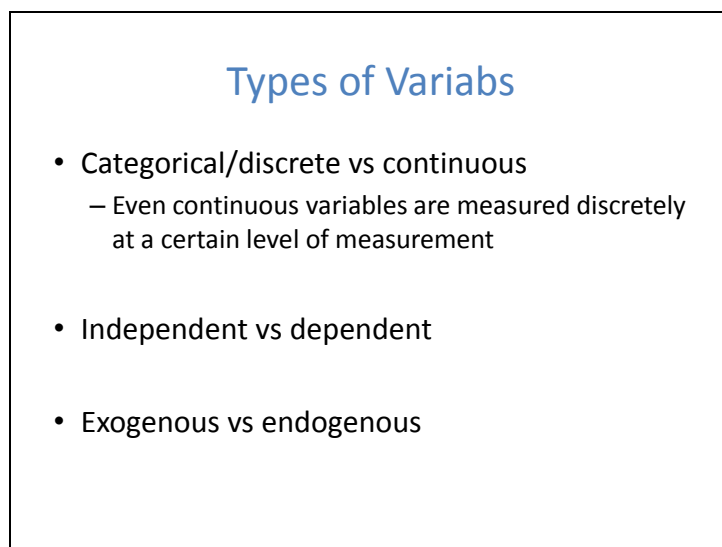
Ratio and Interval

Ratio and interval variables are ordered and the distance between two data points is assumed to be equal. The difference between interval and ratio variables is that ratio variables assume that zero is inherently meaningful. With ratio scales you can speak of someone being 20% higher on a variable than someone else. This is not possible with interval scales. Examples of interval scales include temperature in degrees. Examples of ratio scales include height, time, frequency,

Binary

Binary variables are those that take two values. These are sometimes thought of as nominal, but for many contexts they can be treated differently.

Examples include Yes/No, gender, high/low, good/bad, old/young.

Slide 12

Types of Variabs

- Categorical/discrete vs continuous
  - Even continuous variables are measured discretely at a certain level of measurement

- Independent vs dependent

- Exogenous vs endogenous

Discrete vs continuous

Another distinction made between variables is whether they are discrete or continuous. Variables are discrete when they can take on a limited set of possible values. Discrete variables are also sometimes called categorical.

Variables are continuous when there are an infinite number of possible values that can occur between two points. For example, between 1 minute and 2 minutes there are an infinite number of time points.

Independent vs Dependent
Independent / Predictor / Factor

The independent variable is the variable that we use to explain a particular outcome. Examples include whether someone has received training, the country they come from, or gender. This variable is used to explain differences on a dependent variable. The independent variable is often also referred to as a factor.

In the context of multiple regression variables used to explain a dependent variable are typically called predictor variables. While there are different ways of describing a variable that is used to predict another variable, the terms can be used interchangeably.

Dependent / Outcome

The dependent variable is what we are trying to explain.

Exogenous vs endogenous

The distinction is commonly encountered in Structural Equation Modelling. Exogenous variables are variables external to the system and are similar to independent variables. Endogenous variables are those predicted by other variables in the system as indicated by a directed arrow coming into them. Endogenous variables are similar to dependent variables, but may also be mediator variables within the larger framework of variables.

Slide 13

Measures of Central Tendency

- Mean $$\overline{X} = \frac{\sum X}{N}$$

- Median
  - The middle response

- Mode
  - The most common response

Mean

This is the most commonly reported measure of central tendency. It involves adding up all the scores and dividing by the number of scores. It is appropriate for continuous data

is the mean of all X scores

is the sum of all X scores

N is the number of scores

Median

The median is the middle score. If all scores were ordered from highest to lowest, it is the middle score. The median is the score at the 50th percentile. It is particularly useful for describing ordinal data and continuous data with skewed distributions. It is more resistant to the effect of outliers than the mean.

Mode

The mode is the most frequently occurring category. It is most appropriate for describing nominal data. If you ask people, what their favourite television channel is, the modal response would be the most frequently cited channel. To calculate the mode, you calculate

the frequencies for all response categories and identify the most frequently occurring category.

Measures of Spread

- Sums of Squares $SS = \sum(X - \bar{X})^2$
- Variance $\sigma^2 = \dfrac{\sum(X - \bar{X})^2}{N}$ $s^2 = \dfrac{\sum(X - \bar{X})^2}{N-1}$
- Standard Deviation $\sigma = \sqrt{\dfrac{\sum(X - \bar{X})^2}{N}}$ $s = \sqrt{\dfrac{\sum(X - \bar{X})^2}{N-1}}$
- Interquartile Range
  - Q1 = 25th percentile
  - Q3 = 75th percentile
  - Interequartile range = Q3 minus Q1
- Semi-Interquartile Range
  - Interquartile range divided by 2
- Range
  - Maximum minus minimum

Variance
Variance is the mean of squared deviations from the mean. A deviation from the mean is just the difference of a score from the mean. Squaring the difference from the mean has the effect of removing the sign associated with the difference (e.g., -3 squared = 9; 3 squared = 9). Explaining variance is a recurring theme in statistics.
Population Variance =
Sample Variance =
SS = Sums of Squares =
X = each score
= the mean of all *X* scores
N = number of scores
Standard deviation
The standard deviation characterises the typical deviation from the mean. It is the square root of variance. This has arguably more intuitive meaning than variance.
Population Standard deviation =
Sample Standard deviation =
SS = Sums of Squares =
N = number of scores
Interquartile Range
The interquartile range represents the width of the middle 50% of scores. It is the score for the 75th percentile minus the score for the 25th percentile.
Semi-interquartile range
The semi-interquartile range is half the interquartile range.
Range
The range of scores is the difference between the smallest and largest score.
Range = maximum - minimum

Frequencies

- Frequencies
- Percentages
  - Just the mean of a binary variable
- Cumulative Percentages

Frequencies

Frequencies describe the number of scores of a particular value. Frequency tables can be expressed in raw counts or as a percentage.

Understand your variables

- Use metrics with greatest inherent meaning
- Use Cognitive testing of self-report instruments
- Know norms or results from prior studies

Substantial value can be obtained from reflecting on the nature of our variables.

If I tell you that an organisation, after summing 10 5-point items obtained a mean sale score of 37 and a sd of 5, what does this tell us?

At first perhaps very little. But there are many things we can do to increase the meaning communicated by a scale. This can enable simple descriptive statistics such as means, standard deviations and frequencies, to tell very interesting stories. It can also make unstandardised regression coefficients more meaningful.

Use metrics with greatest inherent meaning

If I was trying to describe a person's typing speed, I could put it on a 3 point scale of poor, good and excellent. However, if you are familiar with the idea of words per minute, this may be substantially clearer.

If I am understanding temperature, it is a lot more meaningful to me to talk in Celsius than in Fahrenheit. Why? I have spent my entire life experiencing the weather and relating to points on the Celsius scale. Thus, each point on the scale has many concrete associations.

There are many scales that we have extensive experience with: E.g., height, weight, price, dollars per hour, rent per week, kilometres per hour

Other scales, some of us will be familiar with and others will note: E.g., BMI, Blood Pressure, CPI, Intelligence

General rule:

- Use scales that have the greatest meaning to yourself and your audience
- Adopt strategies that allow you to learn the meaning of a scale
- Communicate strategies to your audience which allow them to understand the meaning of a scale

Cognitive testing

Many studies include self-report instruments. These include survey items (e.g., "are you satisfied with your boss"), and established scales (e.g., life satisfaction, self-esteem, anxiety, personality). What are the thought processes that go into responding to one of these items?

Do the response options available, assuming a forced choice item, adequately match the responses of the respondent?

Even, just doing the questionnaire yourself and reflecting on your own thought processes is important in understanding what the scale may be measuring.

An example of cognitive testings: http://www.bls.gov/ore/pdf/st960120.pdf

Know Norms, Benchmarks & Reference Points

Some tests will have formal norms or benchmarks for particular populations. Examples include ability tests and personality.

Other times norms can be a lot more informal. Prior studies may have reported descriptive statistics for a particular scale. If you have longitudinal data, you may be able to compare scores to previous time points.

There may be established benchmarks.

Slide 17

Graphs

- Many kinds of graphs
  - Line Graph
  - Bar Chart
  - Histogram
  - Box Plot
  - Pie Chart
  - Scatterplot
- Graphs often tell the story best

Graphs can be a powerful and efficient way of communicating information to your audience. The graph chosen depends on a number of factors, including the kinds of variables being graphed (nominal, ordinal interval) and the questions being asked.

SPSS has a number of ways of bringing up graphs. It has a Graph menu. This allows you to select the graph you wish to run. It has two ways of running graphs. The traditional way will be shown here, but be aware that there is also a graphing module called "interactive graph". This allows you to set up your graph and change features in real-time. SPSS also has graphing procedures distributed across its main analysis modules.

Slide 18
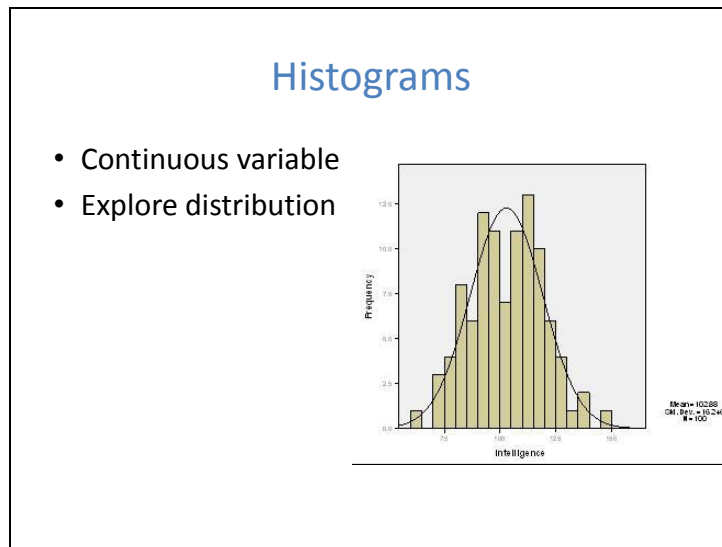
Output shows the frequency for different ranges of intelligence score. An examination of the graph shows that it is relatively normally distributed as the raw distribution matches closely to the normal curve. Mean, standard deviation and sample size ('N') are also displayed.

Slide 19
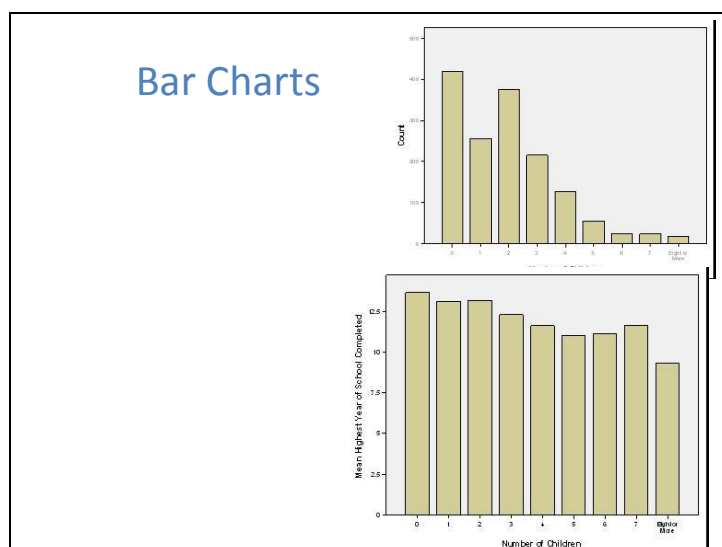
Bar charts can be used for a range of purposes. They are effective at presenting percentages and counts for data with discrete categories, such as ordinal and nominal data. Bar charts can also be used to compare means between groups.

In the output we first see the frequency counts represented in a bar chart of the frequencies in the sample of number of children. The bar chart quickly shows that most of the sample has between 0 and 4 children. 'No children' is the most common response. Of the people with children, two is the most common. The bar chart also allows the determination of the raw frequencies. We can also see that the frequency counts for each category are quite large (e.g., over 400 in the 0 category).

In the second graph we see the mean number of years of high school education for each count of number of children. It would suggest a trend whereby people with more children have had slightly less education on average.

Slide 20
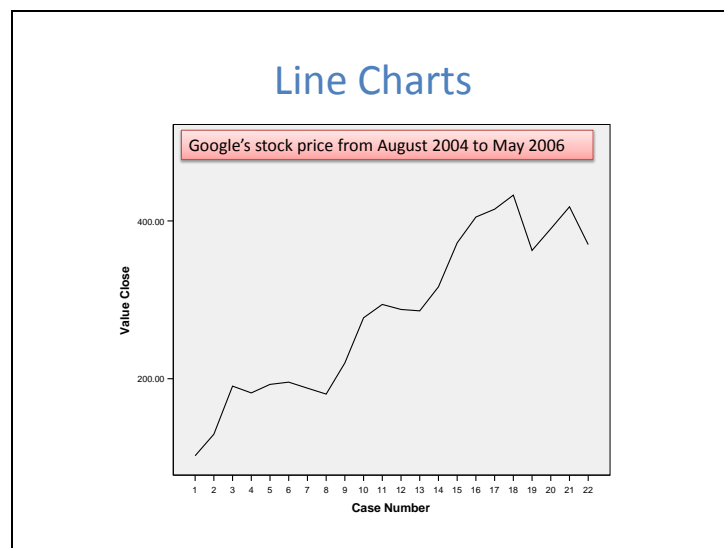
Line charts can be used for a range of different purposes. They can be used in many of the same ways as bar charts. They can be used to show frequency and count information for particular values of a categorical or ordinal variable. They can also be used to show summary statistics, such as the mean, on a second variable across the levels of another variable. They can be particularly good for showing summary statistics when there are two or more categorical grouping variables (e.g., sales revenue in different locations and in different product lines).

Line charts are also particularly good for showing changes in a variable over time. Plotting stock prices, sales, number of customers, and other variables over time can be very useful for exploring trends over time and examining seasonal cycles.

This example is based on a data file of Google's stock price from August 2004 to May 2006. Each row of the data file is a month. If we wanted to plot the stock price in a graph, we would select: Simple – Values of individual cases. Then in the next dialog box, we place the variable representing the stock price (in this case, it is called 'close') in the 'Line Represents' box.

The output shows the way Google's stock price has increased over time but has also gone through periods of stability.

Slide 21
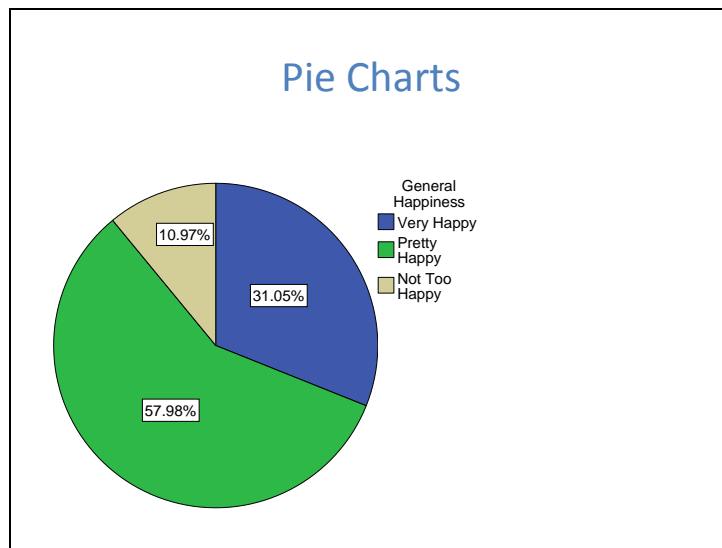
Pie charts are used for data with a limited number of categories, typically nominal or ordinal data to show the relative percentage of each category. The size of a segment of the chart reflects its percentage size.

In the U.S. Social Survey data file, participants were asked to rate their general happiness (very happy, pretty happy, not too happy).

The variable of interest (happiness) goes into the 'define slices by' box.

Slide 22
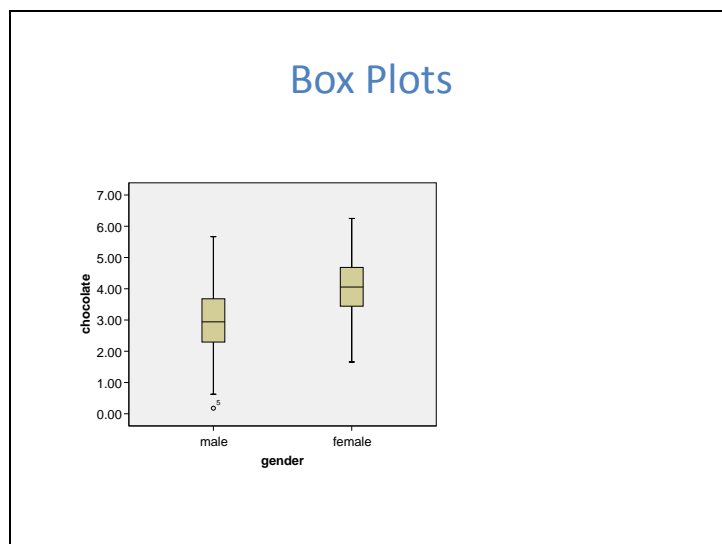
A box plot is typically used to explore the distribution of one or more continuous variables. The box plot marks a number of points on the distribution. The middle black line represents the median. The two points above and below the median, which define the box, represent the 25th and 75th percentile. The tails of the box plot which extend from the box represent the highest and lowest values within 3 semi-interquartile ranges of the median. Circles represent outliers and crosses represent extreme scores.

Box plots are useful in assess whether a variable is normally distributed and in identifying potential outliers that might be having excessive influence on analyses.

In example one, we see the distribution of chocolate liking ratings for males and females. The median liking rating is higher for females than males. Both variables look relatively normally distributed; the median is in the middle of the box and the tails extend relatively evenly either side of the median. In the male sample there was one outlier with a case number of 5.

Slide 23
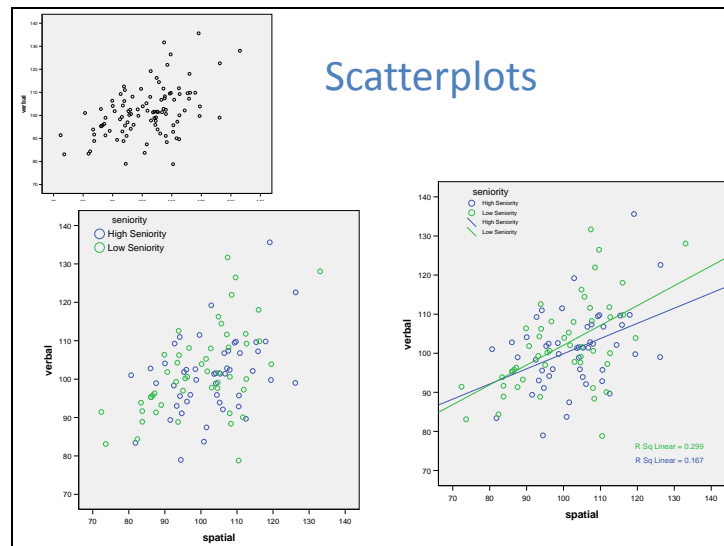
Scatter Plots are used to show the relationship between two continuous variables. They are particularly useful in the context of correlation coefficients. Examination of scatter plots can assist in determining whether a relationship is linear or not. SPSS allows you to attach data labels and colour code data points. SPSS also allows you to plot lines of best fit.

Slide 24
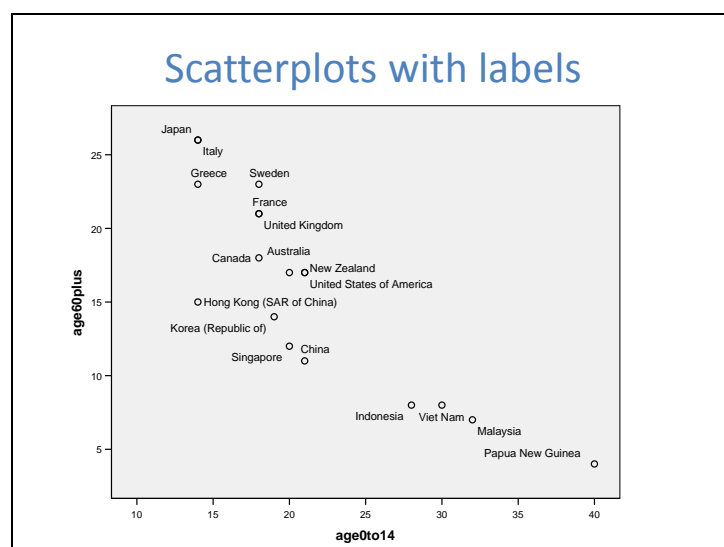
Imagine that we wanted to see the relationship between the percentage of young people and the percentage of old people in selected countries. The following data is taken from the United Nations 2005, World Population Prospects: The 2004 Revision,

In this context we are actually interested in the individual data points as well as the overall pattern. Imagine we wanted to see the relationship between 'percentage of 0 to 14 year olds' with 'percentage of people over 60' in different countries.

Slide 25

<div style="border:1px solid black; padding:1em;">

Distributions

- Types
  - Unimodal vs Bimodal
  - Symmetric vs Asymmetric
  - Rectangular
  - Normal
  - Positively Skewed
  - Negatively Skewed
- The message
  - Inherently interesting
  - Relevant to assumption testing
  - A matter of degree
  - Population distribution only estimated from sample

</div>

Unimodal vs Bimodal
A distribution is unimodal if it only has one mode or one peak. A distribution is bimodal if it has two peaks.

Symmetric vs Asymmetric
A distribution is symmetric if when you draw a line through the middle of the distribution, the left side is a mirror image of the right side. When a distribution is symmetric, its mean and median will be the same.

Rectangular
In a rectangular or uniform distribution, the distribution covers a range of values. Every value within the range of possible values is equally likely to occur.

Normal
The normal distribution is a frequently assumed distribution. It has a characteristic bell shape. It is unimodal and symmetric.

Positively Skewed
A distribution is positively skewed when its tail points to the right towards positive numbers. A common example of a positively skewed variable is income in the general population. Most people earn a little while a few earn a lot and fewer still earn a huge amount. In positively skewed distributions the mode is to the left of the median and the median is to the left of the mean. The mean gets pulled out by the extreme scores. In really skewed data, the median may be a better measure of central tendency.

Negatively Skewed
A distribution is negatively skewed when its tail points to the left towards negative numbers. The same rules apply to it as to positively skewed distribution, but in reverse.

Slide 26
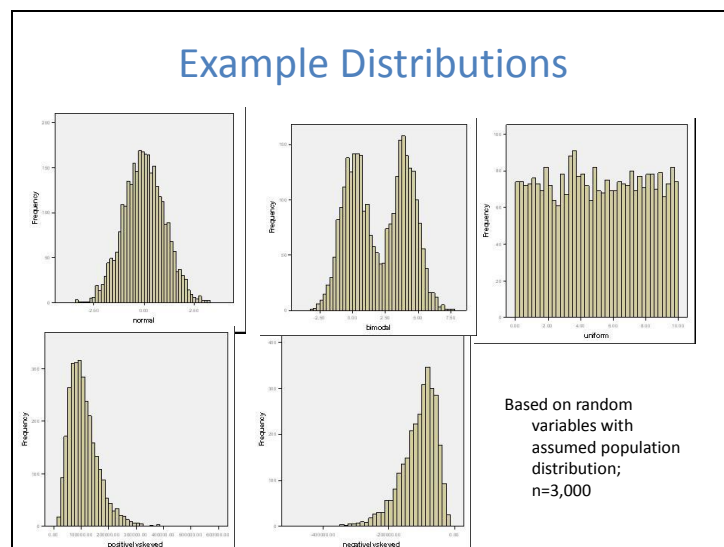
Histograms are a good way of visualising the distribution of a continuous variable.
SPSS has tools for creating random variables using TRANSFORM >> COMPUTE.
Creating random samples of variables with known distributions is a useful way to train your intuition about what the shape of certain distributions look like and how random sampling can result in differences between the sampled distribution and the population distribution.

Slide 27

Z-scores are a useful way of describing an individual's score in a standardised way. A distribution of z-scores has a mean of 0 and a standard deviation of 1. On a normal item, when someone gets a score of 7 out of 10, we do not necessarily know what this means. We need to compare this score to some frame of reference or benchmark in order to understand it. If I said that someone has driven for 30 years and never had an accident, we would agree this is good (or lucky), because we assume that the normal person is likely to have an accident or two over that period of time. Z-scores tell use where someone stands in relation to the mean. Thus, a z-score of 1 indicates that some one is one standard deviation above the mean.
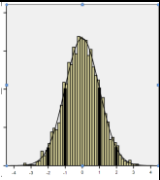
In SPSS:
Analyze >> Descriptive Statistics Descriptives
Place the variables in the variable box and select "Save standardized values as variables"
This creates two new variables in our data file that represent z-scores of the original variables.

Slide 28

Normal Distribution

- Normality (68 – 95 – 99.7 rule)
- Degree of Skew:
  – Normal – Mild – Moderate – Strong - Severe
- Assessment
  – Graphical
  – Statistics
- Consequence of violation
  – P values may not be accurate
  – Relationship between variables may not be optimal

The normal distribution is reflected by a bell shaped curve. It is assumed to arise in many settings in contexts where many random processes are operating. For example, the size of noses, the height of females, and shyness might all be assumed to exhibit a normal distribution. The normal distribution is an assumption of many statistical tests. In reality your data is often not normally distributed, and the question becomes what analyses should I perform. Often the tests that assume normality are relatively robust to violation of the assumption. It is often sufficient to show that our data is relatively symmetric and hope that the test is statistically robust.

When the normal distribution is composed of z-scores it is called the Z normal distribution. It has a mean of 0 and a standard deviation of 1. Based on knowledge of the normal distribution we can state that 68% of scores will be within 1 standard deviation of the mean, 95% within 2 standard deviations and 99.7% of scores within 3 standard deviations.

Assessing Distributions in SPSS

There are two main ways to explore distributional properties. You can assess them graphically or you can assess them with statistics. Graphical assessment of distributions is often better particularly if your sample size is above 100. This is because the statistical tests are often too sensitive in detecting violations of normality. To graphically assess the distribution, bring up a histogram of the variable of interest.

There are a number of ways to assess a distribution statistically. Two common summary measures are skewness and kurtosis. Skewness describes the degree to which a distribution's tails go off in one example (see the examples of positive and negatively skewed distributions above). Kurtosis refers to the degree to which the distribution is peaked or flat. Through SPSS Analyze >> Frequencies2 procedure you can bring up skewness and kurtosis information. When the value is greater than 3 times the standard error of the statistic, this may suggest a significant violation of normality.

Slide 29

Transformations

- Approach
  - Assess normality before transformation
  - Run a transformation and assess normality on new variable
- Arguments against transforming
  - If the untransformed metric is inherently meaningful
  - Most parametric procedures are robust to mild violations of normality
  - If a non-parametric procedure is adequate and appropriate
- Arguments for transforming
  - The skewness is strong or severe
  - Transformations are typically applied in the literature to this type of variable
  - Transformation appears to reveal the "true" relationship between the variables
- Make an integrated reasoned argument for your decision

Slide 30

Transformations

- Standard transformations based on degree of skew
  - Mild: Log10
  - Moderate: Square Root
  - Strong: Minus Inverse
  - Severe: Minus 1/x^2
- Add constant to make smallest value = 1
- For negative skew
  - Reverse the variable before transforming
  - Then reverse the transformed variable back to the normal direction

---

Inferential Statistics

- Population Parameters
- Sample Statistics
- Hypothesis Testing
  - H0: Null Hypothesis
  - H1: Alternative Hypothesis
- P-value
  - The probability of obtaining a result as large as that observed in the sample if the null hypothesis were true.
- Alpha

Imagine we wanted to draw some conclusions about the nature of employees in a particular country. On average how many hours a week do they work? How much money do they earn on average? How many weeks holiday a year do they get? If we were going to research these questions, it is rarely feasible to obtain data from every person in the population of interest. Thus, we draw a smaller sample of people and assess them on how many hours they work, how much they earn and how many weeks holiday they get each year. We then may attempt to infer the characteristics of the broader population from our sample.

Samples and population
Population (parameters)
A hypothetical or actual target population. We are trying to draw inferences about population parameters

Sample (statistics)
A selection of individuals drawn from the population that provide sample statistics to estimate population parameters

Hypothesis testing
H0: null hypothesis
H1: alternative hypothesis

p value
The probability of obtaining a result as large as that observed in the sample if the null hypothesis were true.

Alpha
The probability of falsely rejecting the null hypothesis
Typically, we talk about alpha being .05 or .01

Hypothesis testing logic
if the p value is less than alpha (e.g., .05),
the probability of the null hypothesis being true is low
we reject the null hypothesis
and accept the alternative hypothesis

<div style="border:1px solid black;">

Standard error of the mean

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

- Standard error of the mean = standard deviation divided by square root of the sample size

</div>

The standard error of the mean is the standard deviation we would expect if we did many exact replications of a study.

It is used to calculate confidence intervals around the population mean.

All else being equal, we want smaller standard errors so that we can have more accurate estimates of the population parameter.

Looking at the formula, what makes the standard error of the mean bigger?

Increasing the standard deviation will increase the standard error of the mean.

Increasing the sample size will decrease the standard error of the mean.

Note that the degree to which increasing n will reduce the standard error is not a linear relationship. Because of the square root, halving the standard error of the mean requires you to increase the sample size by a factor of 4.

<div style="border:1px solid black;">

Standard error of percentages

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} \qquad s = \sqrt{p(1-p)} \qquad s_{\bar{X}} = \sqrt{\frac{p(1-p)}{n}}$$

- *p*: Percentage is just the mean of a variable where 0 equals no and 1 equals yes
- Standard deviation of binary variable is a function of its mean

</div>

When reporting the results of surveys and other situations involving percentages, it is often useful to be able to report the margin of error. Typically margin for error reflects the 95% confidence interval.

With binary data, the standard deviation and variance is a function of the mean. Thus, the formula for the standard error of a percentage is no different from the standard error of the mean except that we can derive the standard deviation from the mean.

If you play around with the formulas, you will see that the standard deviation of a percentage is largest at 50% and declines the closer the percentage gets to 0% or 100%. Thus, the standard error is largest when the percentage is closer to 50% and/or the sample size is smaller.

If you are looking for a ballpark estimate of the 95% confidence interval multiple 2 * 50% / sqrt(n). Where 2 represents the approximate 95% confidence of the z distribution; 50% represents the standard deviation when p=.5, and sqrt(n) is the square root of the sample size.

Slide 34

Standard error of percentages

p	n	variance p	sd p	se p	CI95%
0.5	25	0.25	0.50	0.10	0.20
0.5	50	0.25	0.50	0.07	0.14
0.5	100	0.25	0.50	0.05	0.10
0.5	200	0.25	0.50	0.04	0.07
0.5	500	0.25	0.50	0.02	0.04
0.5	1000	0.25	0.50	0.02	0.03
0.2	25	0.16	0.40	0.08	0.16
0.2	50	0.16	0.40	0.06	0.11
0.2	100	0.16	0.40	0.04	0.08
0.2	200	0.16	0.40	0.03	0.06
0.2	500	0.16	0.40	0.02	0.04
0.2	1000	0.16	0.40	0.01	0.03
0.01	200	0.01	0.10	0.01	0.01
0.05	200	0.05	0.22	0.02	0.03
0.1	200	0.09	0.30	0.02	0.04
0.99	200	0.01	0.10	0.01	0.01
0.95	200	0.05	0.22	0.02	0.03

The above table highlights a way of exploring formulas. The formulas were calculated in excel showing the input variables (p – probability; n – sample size), some of the intermediate calculations (variance of p and standard deviation of p) and the resulting values (standard errors and confidence intervals.

This table can also be used to train your sense of the rules of thumb when seeing a dataset. For example, when you see a percentage reported in the newspaper that suggests that based on a representative sample of 200 people, 50% said they like chocolate, you will know that the 95% confidence interval is plus or minus 7% (i.e., between 43% and 57%). Remember that when p equals .5 this represents 50%.

Statistical Software

- Major packages
  - SPSS, SAS,STATA, S-Plus, R, and more
- Dedicated Packages
  - SEM: Amos, Lisrel, more
  - Multilevel Modelling: HLM, MLWin, more
  - More…
- Decision
  - What kind of user are you?

There are many statistics packages out there.

Most people I know in psychology, market research and management use SPSS.

Once you learn one tool, it is not too difficult to learn another.

Considering 'R'

- Open Source – i.e., Free
- http://www.r-project.org/
- Very powerful
- Encourages a better orientation to data analysis
- Takes longer to learn
  - Command line; not really GUI
- Many innovative features not found in other programs
- Improving rapidly

Open Source – i.e., Free: Being open source software, it is free to use. Thus, if after university, you work for an organisation and it does not have SPSS or SAS or any other commercial package, you will not be limited.

Takes longer to learn It runs on the command line (although GUIs do exist) and thus requires a different mindset to that of typical point-and-click programs such as SPSS.

There are extensive tutorials on the internet to get you up an running. I would think that about 100 hours of practice would be enough to get you up and running with R. But equally, once you do learn it, many tasks become a lot easier.

Who's it for? While anyone could get benefits from using R, the following people are likely to get the most benefit:
- You need to automate graphs and reports on large datasets
- You are quite good at statistics

- You see statistical analyses as an ongoing integral part of your career, academic, consulting, or otherwise
- You have done some form of computer programming in the past
- You want to take advantage of particular features that R offers
- You want to write your own custom statistical procedures or techniques and perhaps share them with others

Users: *Dabblers*: If you learn the basics of R, you can use it just for specific features that are not available in your normal commercial software. *Mainstream Users:* You can use R to do the things you do in normal statistics packages *Power Users:* You take your statistical analysis to another level of sophistication and are able to do many sophisticated tasks efficiently. You are able to produce large scale customised reports in minimal time.

Innovative features: Confidence intervals for effect sizes; cool plots; polychoric correlations; meta-analysis tools; item response theory tools; built in datasets to assist your statistical learning; many advanced regression modelling procedures

Improving rapidly: Many statisticians develop the latest techniques for R many years before they are ported to other statistics packages, if they are ever ported to other statistics packages.

Slide 37

Effect Size, Confidence Intervals, Power and Null hypothesis significance testing

- Old School
  - It's all about the p value
- In comes Effect Size, Power Analysis, and Confidence intervals
- Perhaps they can all work together as long as we remember their respective roles

> ## Effect Size
>
> - What are we trying to do?
>   - Refine our estimates of population parameters
> - Contrast with tests of statistical significance
>   - Statistical significance does not mean practical importance
> - Standardised vs Unstandardised
> - Most common
>   - Cohen's d
>   - Pearson's R
>   - Odd's ratio
>
> $$\frac{\overline{X}_1 - \overline{X}_2}{s}$$

Research enterprise:

The main aim of most empirical research is to refine our estimates of population parameters. Whether we realise it or not, many of our research questions can be reduced to parameter estimates. If our research question concerns the effectiveness of treatment intervention versus a control group, then the parameter is the difference between the two group means. If we are interested in how extraversion relates to job performance, the parameter estimate is the correlation.

Contrast with tests of statistical significance

Effect size measures show the degree of relationship or extent of relationship. With a sufficiently large sample size, even small effects can be statistically significant. When evaluating the practical significance of a research finding, whether we are concerned with group differences or the relationship between variables, it is desirable to report a measure of effect size.

Standardised vs Unstandardised:

Typically when people talk about effect size measures, they are referring to standardised effect size measures such as Cohen's d. However, there are many important unstandardised measures of effect. These include the difference between two means and the standard error of the estimate. Unstandardised effect size measure are often preferable when the metrics are inherently meaningful. For example, if you were interested in knowing the difference in income levels of males and females, a good indicator would be difference in mean annual income in dollars.

Most common standardised effect size measures:

Cohen's d is the difference between two group means divided by the standard deviation. This can be the pooled within group standard deviation or the standard deviation of one of the groups.

Pearson's r is a correlation coefficient which ranges from -1 to 1 and expresses the degree of linear association between two variables.

Odd's ratio: This is used commonly used in logistic regression, when we are predicting a binary dependent variable. Odds represent the probability of an event occurring divided by something not occurring. The odds ratio involves dividing two separate odds. This is particularly common in medical research where you might for example compare the odds of

getting a disease for those who have or have not had a particular vaccination (see the seminar on logistic regression for more detail).

Slide 39

Summary of effect size measures mentioned in this course				
Standardised or Unstandardised	**Name**	**Type**	**Type of test it is applied to**	**Interpretation**
Standardised	Cohen's d	Mean difference	Comparing two group means: T-tests, post-hoc tests, and two group ANOVAs	Difference in group means relative to within group standard deviation
Standardised	Eta-squared ($\eta2$)	Variance	ANOVA	Variance explained in DV by treatment variable in the sample
Standardised	Omega-squared (ω^2)	Variance	ANOVA	Estimate of variance explained in DV by treatment variable in the population
Standardised	Phi-prime	Variance	ANOVA	Roughly, square root of treatment variance over error variance
Standardised	Pearson's r (ie., correlation)	Variance	Typically two-continuous variables (but can be binary or continuous) Note: semi-partial, partial, and zero-order are just types of correlations	1. Squaring it allows r-squared interpretation 2. Same as standardised beta in simple regression
Standardised	R-squared (and adjusted r-squared)	Variance	Regression	Variance explained in the dependent variable by the best linear composite of predictors
Standardised	Standardised regression coefficient (Beta)	Difference	Regression	Predicted In crease in DV in terms of standard deviation when predictor is increased by 1 standard deviation holding all other predictors constant
Unstandardised	Difference between group means	Difference	T-tests, & ANOVAs	Interpretation of practical effect is clearest when the means are inherently meaningful
Unstandardised	Unstandardised regression coefficients	Difference	Regression	Predicted In crease in DV when predictor is increased by 1 holding all other predictors constant
Unstandardised	Standard error of the estimate		Regression	Standard deviation of errors around prediction
Unstandardised	Covariance	Variance	Correlation	Unstandardised correlation, typically not used as effect size measure, because meaning is often difficult to assess

The above table shows some of the common effect size measures you might encounter

Slide 40

Rules of thumb

Cohen's d	r	Cohen's convention
2	.71	
1.9	.69	
1.8	.67	
1.7	.65	
1.6	.63	
1.5	.60	
1.4	.57	
1.3	.55	
1.2	.51	
1.1	.48	
1	.45	
0.9	.41	
0.8	.37	Large
0.7	.33	
0.6	.29	
0.5	.24	Medium
0.4	.20	
0.3	.15	
0.2	.10	Small
0.1	.05	
0	.00	

- Rules of thumb
  - Starting point for building intuition
  - Deeper understanding of statistics reduces reliance on them

The following table shows the relationship between Cohen's d and r. Cohen provided some rules of thumb that might guide the practical understanding of obtained effect sizes. These are also displayed on the table. It should be noted that the actual importance of an effect size will vary across contexts.

Slide 41

Connections:
Power, Effect size,
& Statistical Significance

> Really understanding this slide will make you a better human being…
> Okay… Maybe not… But I think it's important

- **POPULATION EFFECT SIZE**
  - is **NOT** affected by:
  - 1) SAMPLE SIZE
  - 2) ALPHA

  Aim:
  To know the degree of relationship and the practical Importance of findings

- **POWER**
  - is a function of:
  - 1) SAMPLE SIZE
  - 2) POPULATION EFFECT SIZE
  - 3) ALPHA

  Aim:
  To have a reasonable chance of testing the research question

  Alpha: is by convention .05 or .01 unless some form of post-hoc adjustment is being applied
  Sample size: is a function of: availability, money, time, & other resources

- **STATISTICAL SIGNIFICANCE**
  - in a particular study is a function of:
  - 1) POWER
  - 2) CHANCE

  Aim:
  To rule out chance as a plausible explanation of observed relationships

Slide 42

Effect Size vs Significance Testing

Null Hypothesis Significance Testing (NHST)	Effect Size

• Key Question
 • What is the probability of obtaining the study's results if the null hypothesis were true ?
 • e.g., no difference between groups; no relationship between two variables
• Answering Question
 • Select appropriate significance test and assumptions
 • Set your alpha
 • SPSS: Compare alpha to obtained p value
 • Hand calculations: compare obtained test statistic to critical value of test statistic at the alpha you set
	• Key Questions
	 • What is the size of the effect?
	 • What does the size of the effect mean in a practical sense?
	• Answering Question
	 • Select appropriate effect size
	 • Contextualise obtained effect size:
	  • Use clinical or professional judgement
	  • Compare to obtained effect sizes in similar studies in research area
	  • Compare to typical effect sizes in studies in the broader discipline
	  • Use rules of thumb of small, medium and large effect sizes (e.g., Cohen's)

Slide 43

Confidence Intervals

- Meaning
  - 95% confident that *population parameter* lies in the specified interval
- In theory,
  - Confidence intervals are available for any standardised or unstandardised effect size measure
  - Any estimate of a population parameter from a sample statistic can have a confidence interval
- Major push in medical and social sciences
  - Think in terms of confidence intervals around effect sizes
    - Forces you to think about uncertainty
    - Focuses thinking on population estimation, which is the basis of theory
    - Encourages *meta-analytical thinking*

Slide 44

Meta Analysis

- Purpose
  - Estimate true population effect
- Composite effect size measure
  - typically d or r
- Adjustments
  - Reliability
  - Weighting by sample size
  - Publication bias / file-drawer effect
- Potential to examine moderators
- Implications for literature reviews

Purpose: Meta analysis is a method for systematically integrating the results of many studies.

> The emphasis is on getting a better estimate of the 'true' population relationship.
> While a complete understanding of the nature of meta-analyses is beyond the scope of this course. Some important elements include:

Composite effect size measure (typically d or r)

> Because different studies use different measurement scales, meta-analysis typically creates a standard effect size measure from the results from each study sampled.
> Cohen's d: difference between two groups in terms of standard deviation (e.g., effect of training on job performance)
> r: Correlation, typically used when looking at the relationship between interval level variables (e.g., commitment with job satisfaction)
> The Odds ratio is also commonly in medical settings

Adjustments

Meta analysis also adjusts for reliability problems. If measurement is less than perfectly reliable, observed correlations will be attenuated. Thus, an estimate of the true correlation can be obtained.

Frequently also samples are weighted by sample size so that studies with greater sample size have greater influence.

Issue of publication bias / file-drawer effect: This is the problem that studies with non-significant findings tend to be harder to publish. Thus, published articles may be over-optimistic about the population relationship.

Potential to examine moderators

A moderator is a variable that alters the relationship between two other variables. When there are many studies, moderators can be examined.

E.g., perhaps there is a stronger relationship between satisfaction and performance in studies where people have control over their work environment (i.e., control is a moderator of the satisfaction-performance relationship)

Implications for literature reviews: When summarising a literature, it is often useful to start with a meta-analysis where available rather than individual studies.

Slide 45
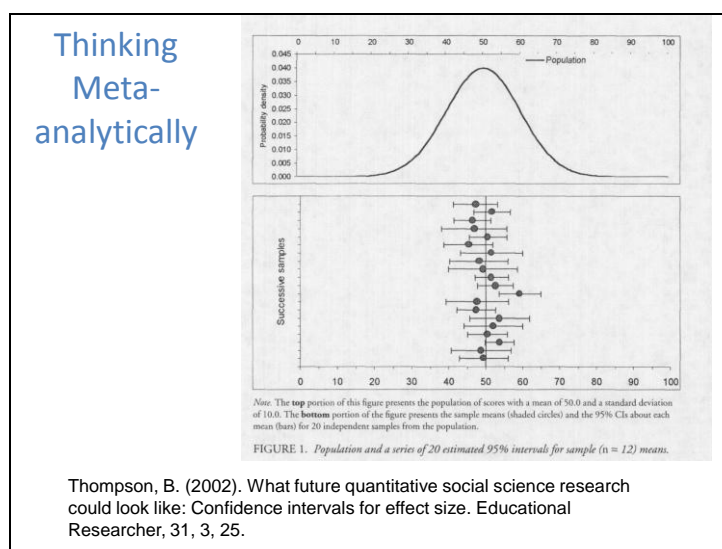
This diagram aims to show the meaning of a 95% confidence interval. The idea is that 95% of the time the true population parameter is contained with the confidence intervals.

Slide 46

Forest Plot

- Diversity of size of spread
- Diversity of mean
- Potential moderators
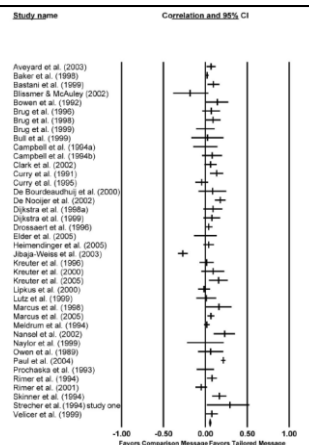
Figure 1. Forest plot displaying effect sizes and 95% confidence intervals (CIs) of studies comparing a tailored message with a comparison message. Brug et al. (1999) = Brug, Steenhuis, van Assema, Glanz, and De Vries (1999); Bull et al. (1999) = Bull, Kreuter, and Scharff (1999).

Noar, benac, harris, 2007

The Forest plot shows cohen's d and associated confidence intervals for a series of studies looking at the difference in communication effectiveness of tailored messages with comparison messages.

Confidence intervals that do not cross zero represent statistically significant differences. Imagine if one of those studies was your own. Imagine also that you were not aware of the range of results that had proceeded. How would you interpret your result in isolation? How would you interpret your study if you interpreted it within this meta-analytic context?

Given the diversity of the results, it is quite probable that different studies are estimating different population effect sizes. Even studies that are looking at the same relationship (e.g., job satisfaction and performance) differ in important ways. Samples differ (e.g., one sample might be white collar and another blue collar); measurement tools differ (e.g., performance might be measured very differently); quality of study varies. All these factors can lead to variation in results.

The idea is that some of the variation in means we see is due to principles of random sampling and some of the variation is due to systematic differences in estimated population effects.

Meta analytic thinking

- Thompson (2002, p.28) defines meta analytic thinking as:
    - "a) the **prospective formulation** of study expectations and design by explicitly invoking prior effect size measures and
    - b) the **retrospective interpretation** of new results, once they are in hand, via explicit, direct comparison with the prior effect sizes in the related literature" [italics added]

Data Sharing & Replication

- Share your data with the world
    - Advances meta-analysis
    - Allows for greater scientific scrutiny
    - Educational aid
    - Increases acceptance of your ideas, if they are good
- Issues to overcome
    - Right to first publication
    - Confidentiality
    - Copyright (in some cases)
    - Inability of others to fully understand the data
    - Fear of being caught making a mistake
    - Lack of rewards for data sharing

I have found the prospect of greater sharing of data quite exciting. Imagine a world where for any journal article you were reading, you could obtain the original data in a well documented form and quickly run the analyses that you wanted to run.

Advances in meta analysis: One of the critiques of meta-analysis is that it is based on summary data. If researchers systematically provided complete datasets as part of publication, datasets could be combined to allow for sophisticated analyses beyond the scope of a single study.

Educational Aid: Reading existing journal articles is an excellent way of learning how to report particular statistical techniques. However, imagine the benefits that could be obtained if doctoral students had the opportunity of recreating the results of the core journal articles they have read.

Increased acceptance, if the ideas are good. In many journal articles the authors decide to test a particular regression model or decide on a factor analysis with a particular number of factors. What if you think that they might not have made the right decisions. What if you

would have run the analyses differently? What if there were some analyses which they could have done based on the variables they reported, but they didn't run? In all these situations, imagine the contribution that could be made if the community of scholars were able to access this data.

The following authors have wrote passionately on this topic:

David H. Johnson: http://www.psychologicalscience.org/observer/0102/databases.html

Garry King: http://gking.harvard.edu/replication.shtml

Right to first publication: There is concern about sharing data that others may publish articles based on your research before you have had opportunity to publish yourself. Of course time limits can be placed on access.

Confidentiality: Data may be stripped of identifying information.

Fear of being caught making a mistake: While this is may be a rational fear particularly for those less skilled in data analysis, it hardly seems grounded in our desire to advance the profession. If anything, accountability tends to improve quality.

It is unfortunate that governments, granting bodies, and universities do not seem to determine contributions especially in terms of contributing data to the research community.

Slide 49

Power Analysis

- Power:
  - Probability of correctly rejecting H_0 when H_0 is false
- Type I error rate (alpha)
- Type II Error rate (Beta)
- Power (1- Beta)

- **Increasing Power**
  - Increase Effect Size
  - Increase Sample size
  - Reduce Alpha
- **Power is a property of a hypothesis not a study**

	Decision	
	Retain H0	Reject Ho
H0 True	$(1-\alpha)$	Type 1 error (α)
H0 False	Type 2 error (β)	Power $(1-\beta)$

Objective Status

In the logic of hypothesis testing, there are two possible states in the world and we can draw two possible conclusions. If the null hypothesis is true and conclude that this is the case, then we have made a correct decision.

Power is the probability of correctly rejecting the null hypothesis when it is false in the population.

To state the case strongly, a study that with insufficient power is not worth doing.

From a practical perspective, it is important to know what power is and what increases it.

Power increases with bigger samples.

Power increases with bigger effect sizes.

Power is a property of a hypothesis not a study.

Typical opening question regarding power analysis with a statistics consultant:

Student: "My supervisor/committee/grant form/ ethics form requires me to do a power analysis to work out what sample size I need to have reasonable power. What sample size do I need for my study to have 80% power?"

What's wrong with this question? Power analysis is not a property of a study, it is a property of a hypothesis.
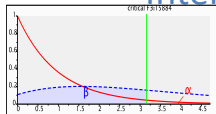
Slide 50

Slide 51

One of the nice aspects of G-power is that you can play around with different results and train your intuition about the relationship between power, sample size, effect size and alpha.

Slide 52

Slide 53

A-priori power analysis is typically used at the design phase to determine what sample size is required to have a reasonable chance of obtaining statistically significant results.

To work out the required sample size, we need to specify the population effect size, alpha and power.

Alpha is typically set at .05 by convention.

Population effect size is usually not known exactly. Meta analyses and large scale studies are typically the best indicators of what to expect. Any prior research in the area or in related areas may be useful. You might also fall back on Cohen's rules of thumbs combined with you intuition about what the effect size is hypothesised to be.

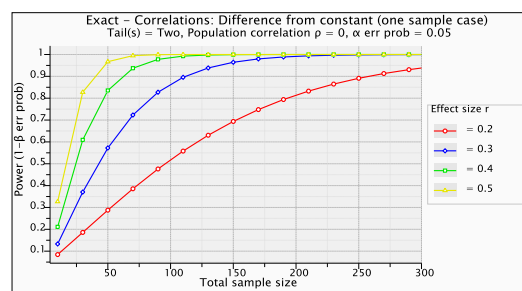*Desired power* is often set at .80 as a reasonable minimum. However, this is only a convention.

G-Power – Post hoc power analysis

- Did I have a hope of finding a statistically significant result?
  - Effect size
  - Alpha
  - Sample size

Post hoc power analysis aims to determine the power of the study on the basis of alpha, the sample size and the obtained sample effect size.

Power Curves

G Power 3 lets you produce some very nice power curves that show the relationship between the variables.
This graph shows that if you are expecting a small correlation (i.e., r=.2) you will need 180 participants in order to have 80% power.

Accuracy in Parameter Estimation (AIPE)
Going beyond Power Analysis

- Better to think about sufficiently small confidence intervals on parameter estimates
- Tasks
  - **PLANNING:** What sample size is required to have a defined confidence interval around an effect size?
  - **INTERPRETATION:** What confidence interval is around the obtained effect size

The aim of a priori power analysis is typically to determine what sample size is required to infer with reasonable probability (e.g., 80%) that a population parameter is significantly different from zero. While this has the potential to improve study design, in many areas of research, the interest is not about whether there is or is not an effect, but rather the size of the effect.

If my interest is on the relationship between intelligence and job performance, I am already pretty confident that there is a relationship from the thousands of studies and meta-analyses that have established this in the past. A conclusion that the correlation is greater than 0 would be of little interest. My interest is centred on what the strength of the correlation. Is it small (r=.1), medium (r=.3), Large (r=.5) or very large (r=.7)? How can I make sure that the confidence interval of the correlation that I obtain is small enough so that I can get a reasonable estimate of the true population correlation?

On a practical level, this tends to mean that even larger sample sizes are required that would be suggested based on a sample size.

If you are looking for practical rules of thumb, I tend to find N=100 to be reasonable and N=200 to be quite good. At these levels 95% confidence intervals around correlations become reasonably accurate in the sense that we have a reasonable approximation of the population effect size relative to rules of thumb.

I have seen otherwise good textbooks recommend not having too large a sample size as this may lead unimportant relationships becoming statistically significant. I find this advice misguided. If you are trained in an appreciation of effect sizes, you know that statistical significance does not mean practical significance. All else being equal bigger samples will always provide better population estimates.

For articles about the approach:

http://www.indiana.edu/~kenkel/publications.shtml

http://cran.r-project.org/src/contrib/Descriptions/MBESS.html

Slide 57

The Core Message:

– Sample as many people as you can
– Appreciate that in most studies there are fairly large 95% confidence intervals around results
– Don't replace p-value fundamentalism with effect size or power analysis fundamentalism
– See results in terms of a meta-analytic context
– Sample size estimation is more than just power analysis

Slide 58

Structuring Data Files

• Label your variables well
• Give appropriate variable and value labels
• Be systematic

A clear and logical set of variable names will make life easier in terms of finding variables. It also make using syntax a lot easier.

Documentation of the data file in terms of value labels and variable labels is also important. This is particularly the case if you have a break between analyses or if you want anyone else to understand what you have done.

Variable Database

- Maintain an Excel spreadsheet of all your variables
- Start with SPSS File >> Display Data File information
- Format in Excel as List (v. 2000, 2003) or Table (v. 2007)
- Add columns with any relevant information
- Uses filters and sorting tools to select variables
- Use SPSS syntax and this database to make variable selection easier

SPSS is a powerful tool, but its facilities for searching and selecting variables are quite poor. Any reasonable sized study typically can have several hundred or even thousands of variables.

It then becomes challenging to efficiently select the desired variables for analyses.
http://www.psychologicalscience.org/observer/1201/databases.html

Data Checking

- Overview
- Always include an ID number
- Basic checking of data entry
- General Approach
  – Stay close to the data
  – Compare expectation of means, category frequencies to the data
  – Check minimum and maximum values are within range
  – If you use a value to indicate missing data, make sure it is recorded as a missing value code in the software

Overview
Quality control is critical in data analysis. If data becomes severely corrupted, all the time and money spent on collecting, analysing data can amount to nothing. In fact, reporting on corrupted data is often worse than nothing, because it can strengthen beliefs in baseless claims. For these reasons it is critical to have a data validation strategy. The traditional computer science idea of 'garbage in, garbage out' is very relevant here.
Strategies
Basic checking of data entry
If the data has been entered from paper based questionnaires and tests, it is worth taking a subsample of the tests and verifying that the values in the data file are correct. If critical decisions in

relation to individual cases, such as when entering exam grades, or entering data that will effect promotions, hiring decisions and similar things, it may be worth considering a double data-entry methodology. In this case, all data is entered twice and checks are performed to verify that have been entered in the same way.

Frequency and range checks

After entering data and setting up your data file in SPSS, it is important to verify that the data is correct. There are many reasons why errors can arise. When doing data entry, errors can be made (e.g., someone type 44 instead of 4).

It is important to examine the minimum and maximum column to verify that all values are within the range of valid values. For example, if you have a 5 point scale and you have a score of '45', you can presume that there was a data entry error. Number of valid responses should be assessed to see that the number of cases with missing data is not smaller or larger than you expect.

Frequency counts should be assessed to see that the frequencies correspond with theoretical expectations. This is particularly the case for nominal data and ordered categorical data with small number of categories (e.g., less than 15).

Additional checks for continuous data

For continuous data means, standard deviations and histograms should be examined. Compare the means to what you would expect. For example if you are sampling people from the adult population, you might expect the mean age to be somewhere between 30 and 45. If the mean age was outside this range, you might want to think why that might be.

Slide 61

Error Checking

- Think > Act > Review (TAR)
- Think
  - What are your expectations?
- Act
  - Enter your data
  - OR Run an analysis
- Review
  - Run an analysis which allows you to test your expectations

The most important rule to follow for maintaining integrity and quality in data integrity is: Compare theoretical expectations to obtained results. Sometimes your expectations will not be correct. But other times such differences may reflect problems with the data. It is important to always be vigilant to this possibility.

An additional benefit of the data validation procedure is that you get to know your data. You get familiar with the basic descriptive statistics and distributional properties of the data. It also forces you to think about what the data would be expected to look like.

Constructing scales

- Sum
- Mean
- Reversing
- Scoring
- Weighted Composite

Converting data from the form it is originally entered into a form that is more appropriate for analysis is a common task in statistics. SPSS provides a range of tools to assist this process. This section outlines some of the most important of these tools with examples of how they are applied.

Some other commands that you might consider running could include

Calculate the average of a set of items

COMPUTE meanchoc = mean(choc1, choc2, choc3).

EXECUTE.

Calculate the total of a set of items

COMPUTE totchoc = sum(choc1, choc2, choc3).

EXECUTE.

Reverse an item on a 5 point scale so that 1 becomes 5 and 5 becomes 1. The rule for doing this is that the new score should equal: minimum + maximum – original score.

COMPUTE reversedchoc1 = 1 + 5 – choc1.

EXECUTE.

<div style="border:1px solid">

Reliability Analysis

- Internal Consistency reliability
- Rules of thumb for interpretation (chronbach's alpha):
    - >.8 excellent
    - >.7 good
    - >.6 mediocre
    - <.5 poor

</div>

Theory

An important property of a measurement instrument is that it is reliable.

The most commonly reported measure of reliability is chronbach's alpha. It is a measure of internal consistency. Thus, if you have six items that are all meant to measure the same thing, chronbach's alpha will give an estimate.

Rules of thumb for interpretation:

>.8 excellent

>.7 good

>.6 mediocre

<.5 poor

It is an estimate of the correlation between the observed score and the true score.

Strictly speaking, reliability is not a property of a test. It is a property of a test applied to a particular sample in a particular context.

Assumptions

Sample Size

In order to get a reasonable estimate of the reliability, you need a reasonable sample size. As a rough rule of thumb you might desire at least 80 to 100 people before calculating reliability.

Item reversal

Any reverse scored items have been reversed.

For example, if you have three items measuring 'happiness': 1) Are you happy; 2) Do you like your life; 3) Do you sometimes feel unhappy. The first two aim to measure happiness and the third attempts to measure the opposite of happiness. Thus, prior to adding the items up to form a total or running reliability analysis, you would need to reverse the negatively worded items so that high scores on item 3 now reflect happiness.

Items are continuous or binary

To perform chronbach's alpha the items need to be ordered or binary. While it is possible to use other reliability tools for categorical data, these can not be analysed with the main reliability tool in SPSS.

Missing Data
AVOID>ASSESS>ESTIMATE>JUSTIFY
(see Hair et al Chapter 2)

- Avoid it using good design
- Assess
  - Type of missing Data
  - Extent of missing data
  - Pattern of missing data
    - Missing at random (MAR)
    - Missing completely at random (MCAR)
    - Not missing at random

Overview:

Missing data is a common occurrence in data collection. Participants drop out of studies. Machines break down. Pages from questionnaires get lost. Participants forget or refuse to answer particular questions. For these reasons and many more, the final data file can often contain a large amount of missing data. Missing data can present problems for analysis. How should missing data be treated in data analysis?

Avoidance:

None of the solutions are ideal and the most important rule to follow is to try to minimise the occurrence of missing data by using good research design. Online tests can force participants to respond to every item. Questionnaires can be pilot tested to verify that all questions have valid responses. Drop-outs in longitudinal data can be minimised using a range of strategies. The point is that it is difficult to completely resolve missing data issues after a study has been completed. It is critical to think about how to minimise missing data at the design phase.

Type: There are many different reasons for missing data at it is important to consider what these reasons might be. It is important to think about the reasons for missing data.

Extent: What percentage of data is missing for each variable? What percentage of data is missing for each case?

MAR: Missing data is predicted by other variables present in the

MCAR: The most desirable scenario, missing data is completely random and non-systematic

Not Missing at Random: Missing data is predicted by variables not present in the study

---

Missing Data

- Estimation Methods
  - Listwise deletion
  - Pairwise deletion
  - Estimate missing data
  - Mean Substitution (Variable, Series)
  - Regression
  - Expectation Maximisation
  - Other approaches

Listwise deletion: Listwise deletion is the standard missing values procedure in SPSS. Listwise deletion removes any case that is missing data on any variable used in a particular analysis.

Pairwise deletion: Pairwise deletion is a missing values procedure which is often an alternative option in such analyses as regression, factor analysis and correlations. It only deletes a case with missing data for the elements of the analysis that rely on that variable. For example, in a correlation matrix, if a case has data on variable X and Y, but not on variable Z, it will be included in correlations between X and Y, but not in correlations between X and Z or Y and Z.

Replace with mean: This procedure replaces the missing data with the mean value for the variable. This procedure is not a particularly respected procedure for missing data replacement as it tends to reduce the variance of the variable. It can be done automatically in certain SPSS procedures such as factor analysis. It can also be used by going to: Transform >> Replace missing values; and only placing one variable in the 'new variable' box.

Replace with series mean: This is a more sophisticated technique than replace with mean. It is appropriate when you have a number of items in a questionnaire that are all measuring the same thing and are on the same scale. This procedure will replace a missing item with the value of the data that is present. This can be performed in SPSS by going to: Transform >> Replace missing values. Then, add the variables that form the set of similar items making sure to specify the method as 'series mean'.

Replace with best guess: This is not a particular sophisticated technique, but sometimes we have enough information about our data to estimate or know what the particular case would have received. This technique is not always appropriate and it depends on the knowledge of the analyst. There is potential for bias to enter into the data, if this is performed without due care.

Advanced Techniques

Several more advanced techniques exist for replacing missing values. These include regression, EM, and imputation. Regression attempts to make a prediction for the missing data for a particular case based on what values the case has for other variables. EM is similar to regression in that it uses information from other variables to make a prediction, but does this in an arguably more sophisticated way. One form of imputation involves finding a case that closely matches the existing case and replacing the missing value with the value for this closely matching case. These methods are not available in the base module of SPSS. They are available with the SPSS missing data analysis module.

Missing Data

- Reporting
- Answer the questions
  - What is the extent of missing data?
  - What are the reasons for missing data based on theory, observation and missing data patterns?
  - What is the best way of dealing with the missing data so that our inferences about the population are least biased?

OUTLIERS: ASSESS > UNDERSTAND > RESOLVE

Outliers - Assessment

- Definition
  - a value or an entire case that is substantially different from the rest of the values or cases
- Types of Outliers
  - Univariate Outliers
    - Metric: large absolute z-score (>2.5, 3 or 3.5)
    - Nonmetric: category with very low or high percentage relative others
  - Bivariate Outliers
  - Multivariate Outliers
- Compare numeric with graphical techniques

Univariate Outliers

A univariate outlier is a case that is particularly high or low on a continuous variable. It is often defined as a score that is more than a certain number of standard deviations away from the mean. Values larger than 2.5 in a small sample or 3.0 in a large sample may be considered as potential outliers. When outliers are encountered, it is worth considering whether they have been produced by a data entry error or some other error.

Bivariate checks

Further checks can be performed by looking at crosstabulations of categorical data and correlations for continuous data. If your variables are expected to be related in particular ways, it is worth seeing whether the particular relationship is present. Relationships may be a theoretical question, but a failure to find a particular relationship may indicate data entry problems. Sometimes you will find that where there should be a positive correlation there is a negative correlation. In this case, the variable may need to be reversed. Some questionnaires have conditional questions such that only participants who answer a

particular way to question 1 are asked about question 2. By examining the crosstabulations between question 1 and 2 you can verify that all those and only those people who were meant to respond to question 2 did in fact respond to question 2.

Slide 68

Understanding the reasons for an outlier helps determine what to do with it.

Slide 69

Remove the value
One option is just to delete the value from the dataset. This should be based on a reasoned assessment that the case is not meant to be modelled.
TRANSFORM: When the distribution is quite skewed, it is common to get cases that come out as outliers. We can think of income levels and the billionnaires. Often adopting a transformation will make the distribution more normal and will in turn resolve the issues with outliers.
Adopt a non-parametric test or other robust procedures

It may be that a parametric test may be asking the wrong question. Perhaps, instead of doing a parametric t-test comparing mean income levels between males and females, it would be preferable to perform a non-parametric test that treats income as ordinal data and compares the ranks or compares the median for the two groups.

Other robust procedures include the trimmed mean. There is also a whole class of robust regression procedures that are designed to be less influenced by outliers.

Back up:

Backing up is often best achieved by maintaining two versions of the variable in the same data file. The first version contains the original data and the second version is the one with outlier values replaced, deleted or transformed. At other times it may be more convenient to have completely separate datafiles, one with outliers

Slide 70

I have noticed a tendency at least within psychology to make statistical decisions based on simple black and white rules. Examples of such rules: You need 100 participants to do multiple regression; p must be less than .05 to be statistically significant; if normality is violated, you must do a non-parametric test.

I am not saying that any of these rules are necessarily right or wrong. The point is that they are based on reasons. The reasons often get lost when following the rule. The idea about sample size selection relates to power and accuracy in parameter estimation. A sample size of 110 is not much different to a sample size of 90 when it comes to statistical power. The idea about normality relates to the accuracy of resulting p-values. The point is that good decisions have reasons. To make good statistical decisions theoretical understanding of the phenomena under study must be combined with the pros and cons of the options present.

• Know the options
• Understand at least some of the pros and cons of each
• Getting similar results from both approaches relaxes the burden of the decision a little
• Make a decision

Slide 71

Statistics as reasoned argument

- Decisions requiring reasoned argument:
  - What the research questions are
  - Which rule of thumb to apply
  - Which analysis to run
  - Within each analysis which options to choose
  - Interpretation and generalisation of results
- Some sources of reasons
  - Data
  - Theory
  - Statistical literature: Textbooks and primary journals
  - Purpose of research
  - Convention

What reasoned decisions do you foresee having to make?

Slide 72

Core Messages

- Learning Strategies
  - Practice with Feedback
  - Explore formulas
  - Try to see commonalities across topics
  - Consolidate the basics if necessary
  - Integrate statistical understanding with everyday awareness of the world
- The Research Question
  - It's all about the research question
  - Know how to choose the right statistical tests given the research question
  - Know how to link statistical tools and statistical output back to the research question and your intuition
- Becoming a Sophisticated Data Analyst
  - Statistics involves subtle decisions and is based on reasoned argument