INTRODUCTION TO
Quantitative Research Methods

A Guide for Research Postgraduate Students
at The University of Hong Kong

A publication of the Graduate School

Introduction to Quantitative Research Methods

Author:
Professor John Bacon-Shone

Publisher:
Graduate School
The University of Hong Kong

Feedback:
johnbs@hku.hk

Acknowledgements:
Dr Margaret Taplin contributed the section on qualitative research

ISBN:
978-988-12813-0-2

Version:
2013-10-17

This course is designed to include sufficient statistical concepts to allow students to make good sense of the statistical figures and numbers that they are exposed to in daily life. At the end of the course, students should understand the basics of quantitative research and be able to critically review simple statistical analysis.

The examples are intended to be relevant in Hong Kong for a wide range of disciplines.

Most of the following topics and questions will be covered in the course:

I: Research Methods

What is research?

‘A systematic and unbiased way of solving a problem (by answering questions or supporting hypotheses) through generating verifiable data.’

This is the fundamental definition we need, so we need to understand systematic, unbiased, hypotheses and verifiable, all of which we will examine later.

Why use research methods to solve problems?

This is the question that is so fundamental that we do not always ask it!

Other possibilities: rely on authority (parents, supervisor, police, etc.), personal experience (what happened when I tried to do this before), common sense (apply simple logic), revelation (rely on my god to tell me) or intuition (rely on my instincts or feelings).

Let us examine some problems to understand how research compares with the alternatives:

1) Should I cross the road at a specific place where there is no pedestrian crossing?
2) What should the HK government do to improve air quality in the next 20 years?
3) Who should I marry?
4) Should I become a Christian (or Buddhist)?

Conclusion: these other methods may all be useful at times, but not good ways to provide good long-term solutions to important problems.

The research process includes problem definition, information search, formulating hypotheses, choosing a research design, collecting/obtaining data, qualitative and/or quantitative analysis of data to verify, modify or reject hypotheses, writing a report of findings, all of which are important (we will look at the process in more detail later).

Key word: verifiable (testable?)

In all research, it is important that other researchers can try to replicate your findings.

Experimental scientists talk about repeatable experiments as researchers are expected to provide enough details that others can try to replicate their findings by repeating their experiment.

However, some research cannot be repeated (e.g. effect of handover on Hong Kong) because the conditions of the data collection cannot be repeated. May be able to reanalyse the data, but not collect a new set.

Hence high quality data collection is a particularly important issue for social science (but also for others where data collection is very expensive or difficult to repeat, e.g. astronomers studying creation of stars or geologists studying volcanoes). It also explains why in the US and Europe, research funded with public money must share the data and journals often require public access to any data analysed in a journal paper.

The reality is that we all make mistakes and have false preconceptions. Society (and the state of knowledge) can only advance if mistakes can be identified and corrected, which is possible with research.

Perhaps this is the key reason that democracy works better than the alternatives, because if the representative you choose is ineffective, you can try another one?

We will now discuss what is called the ‘scientific method’ of research, including both qualitative and quantitative methods, which are used in the social, biological and physical sciences.

What is the scientific method?

Defined as ‘an objective, logical and systematic method of analysis of phenomena devised to permit the accumulation of reliable knowledge’

What are phenomena?

Phenomena are what were observed (often cannot measure directly, although brain measurements have the potential to change that in some cases) – which can be either subjective (e.g. Attitudes, feelings) or objective (e.g. time, weight) measurements.

What is objective?

Objective: evaluate phenomena from a dispassionate, apolitical, atheological, nonideological viewpoint.

That sounds difficult and boring!

Can we really evaluate with No passion, No politics, No religion, No ideology?

Note: the key is objectivity when testing the ideas and reporting the findings, not when choosing the research problem (if you have no passion for your research topic, life will truly be boring!)

Note: unethical evaluation which tries to distort the evaluation, clearly causes problems, here we focus on unintentional errors.

Can a scientist truly be objective? Can the process be objective?

Note: asking questions can alter respondent views and even for quantum experiments, measurement affects reality.

Example: survey that asks whether elderly people have discussed with family members the possibility of retiring in the Mainland – if the answer is no, what do you think they will do after the survey is finished? This is an example of why survey design requires skill to avoid bias in the process.

Example: during SARS, we did a telephone survey for the government of people living in Amoy Gardens asking about hygiene behavior – this was intervention, not research!

Example: there is known bias in the academic process (publication bias) where evidence that findings that are different to current mainstream thought may be either selected for or against. For this reason, in many countries, clinical trials of drugs must now be registered, so the results are all made public regardless of outcome.

What is a logical and systematic method?

Logical reasoning: following the (rational) rules of induction and deduction:

Deduction: general to specific (generate ways to test theories in new situations, looking for a situation where the theory fails)

Induction: specific to general (generate theories from observations, creating a new theory, replacing any failed theories)

Amusing story: a former Physics chair professor in HK gave a public lecture just before he left HKU – he claimed that physicist are the only true scientists because they rely on deduction, not making the mistake of using induction! I wanted to ask him how he thought Einstein came up with the theory of relativity, if not using induction?

Systematic Method: a procedure for doing things that can be explained to others and is built on previously existing knowledge. If you cannot explain to others, how can they understand it or evaluate it?

What is analysis?

Analysis means both qualitative and quantitative methods (i.e. with and without numerical information) of processing and summarizing information.

Levels of quantitative analysis

Descriptive: simple statistics relate description of sample to description of population, how good is the description? This level of analysis can be of vital importance, e.g. population size and unemployment rate for the Hong Kong population, as done by the Census and Statistics Department, but most research published uses a higher level of analysis.

Explanatory: understand why things happen as they do, how reliable is that understanding, or are there other explanations? For example, if the unemployment rate has risen, we want to understand why – is it because of young people entering the labour market, is it because restaurants are sacking dishwashers and replacing them with machines? Statistical models can be invaluable in this situation.

Predictive: need a model of future outcomes, how well does it work? For example, if we are looking at admissions to universities, can we predict which students will receive admission offers before they get their public exam results, if we know their performance in school exams? This is not hard for the group of all school students, but is very hard for individual students.

What is a theory?

A theory is a generalised synthetic explanatory statement, in other words, an abstract conceptual explanation of the world. Conceptual explanations are important in order to generalise our findings across a wide range of situations, but unless they lead to predictions, we cannot test our theories, so we need models.

What is a model?

A model is a way to generate verifiable predictions based on a theory, although there may be uncertainty (randomness) involved and also unknown parameters for the model, that need to be estimated in our research. For example, if our model predicts that personal incomes increase proportional to the years of education, we still need to estimate the slope and intercept of that linear relationship before we can make predictions. We need statistical models to estimate these parameters, given that there is also uncertainty as these are not exact models. Even when there is an exact relationship, there is always uncertainty in practice from measurement error.

A theory is of little value until it is testable, in other words, we should be able to build a model that can be empirically tested. If that model fails, then we will need to modify our theory. If the model does not fail, it does not mean the theory is correct, but suggests the model may be useful, if and until we are able to find a test that the model fails.

What do we mean by “a model fails”? We mean that there is some inconsistency between what the model predicts and what happens. We will examine later how to use statistical testing to identify a model failure.

Clearly, a theory that leads to models that allow us to make good predictions is a useful theory, but we should not assume that it is necessarily a fundamental truth about the world.

For many centuries, people thought that Newton's laws were a fundamental truth, until Einstein showed that they fail under certain conditions (speeds close to the speed of light). Note that we still use Newton's laws every day, despite Einstein's findings, because they are so close to true at everyday speeds!

In short, while absolute truth may exist, we will never know whether we have found it yet using the scientific method!

A famous statistician (George EP Box) said "all models are wrong, but some are useful".

What is a hypothesis?

A hypothesis is a statement that can be empirically tested, i.e. translation of theory into a testable statement.

The **research (or alternative) hypothesis** is a positive statement about what the researcher expects to find.

The **null hypothesis** is a statement that a relationship expected in the research hypothesis does not exist, i.e. that the world is simpler than predicted by theory.

Example: Consider the proportion of males and females studying undergraduate degrees in the whole of HKU. The simplest explanation would be that there are equal proportions (50%), so this might be my null hypothesis, while I might expect the proportion to be unequal, hence this is my research hypothesis.
Note: research papers do not always explicitly state the hypotheses, but if they are testing whether there is a relationship between education and income, the implicit null hypothesis is that there is no relationship and the implicit research hypothesis is that there is a relationship.

Question: why do I think my hypothesis of equal proportions is simpler? This yields two questions – what is simple and why do I care about simplicity?

Occam’s law

If we have two explanations of the world, which are equally good, we should prefer the simpler explanation.

Question: why should we prefer simple explanations?

Karl Popper: We prefer simpler theories to more complex ones because their empirical content is greater; and because they are better testable

Richard Swinburne: Either science is irrational in the way it judges theories and predictions probable or the principle of simplicity is a fundamental synthetic a priori truth.

Ludwig Wittgenstein: Occam's razor is, of course, not an arbitrary rule, nor one justified by its practical success. It simply says that unnecessary elements in a symbolism mean nothing. Signs that serve one purpose are logically equivalent; signs that serve no purpose are logically meaningless. The procedure of induction consists in accepting as true the simplest law that can be reconciled with our experiences.

Example: the hypothesis that the world is pre-determined is unbelievably complex and of no use for predicting the future, so it will be a very low priority in our set of theories.

If statistical models are used, we expect the null hypothesis to be rejected and hence the alternative (research) hypothesis to be tenable (believable). The null and the alternative are thus usually complementary.

The null hypothesis is usually the simpler statement, such as there is no effect caused by something, while the research hypothesis would be that there is an effect, or that there is a positive effect or that there is a negative effect. If the paper uses statistical methods, there must be a null hypothesis and a research hypothesis, even if they are not clearly stated.

We can only disprove, rather than prove hypotheses, hence we look for evidence to disprove the null hypothesis and if we cannot find sufficient evidence, we accept the null hypothesis.

Strictly speaking, statistics cannot usually even disprove a hypothesis, it can at most show that the outcome we observed was (very) unlikely, if we assume that the null hypothesis is true. This is because statistics uses probability (chance) statements rather than absolute statements. Thus we are usually prepared to reject the null hypothesis as unlikely to be true if there is sufficient empirical evidence against it. We will come back to the question of what constitutes sufficient evidence.

Example: if I toss a coin 20 times and every time I get a head, would you believe this is a fair coin (has 1 head and 1 tail)? The chance is about 1 in a million of observing this outcome with a fair coin, but a certainty if my coin has 2 heads!

How do we define simplicity? Often it is defined as the sample size minus the number of parameters in our statistical model, which we call the degrees of freedom (we will revisit this).

What is proof?

Proof means that you know the truth

What is verified?

Verified means that your tests did not disprove the truth.

Example: if I compare your face to your picture in your identity card or passport and they match, I have verified your identity, but it is not proof because you might have an identical twin or have obtained a fake document.

If you test many hypotheses resulting from a theory, and none of them are shown to be false, you may think the theory is true, because of numerous verifications but you still have not proved it.

It may be that the weaknesses in the theory have not been identified yet (e.g. Theory of relativity and Newton's laws).

What is a constant?

A constant is something that (it is assumed) does not vary over the study

What is a random variable?

A random variable something that varies over time or over subjects (in other words, varies within the study), also used to mean the operational definition of a concept (how do we measure something).

Creating a good operational definition is a skill and it is important to look at previous work (hence the importance of a good literature review). We will discuss operational definitions later, including evaluation of operational definitions.

What is an explanatory variable?

Explanatory variables are random variables that are the object of research

What is an independent variable?

An independent variable is an explanatory variable that is a presumed cause of variation in other explanatory variable(s)

What is a dependent variable?

A dependent variable is an explanatory variable presumed to be affected by the independent variable(s).

Note: If there are independent variables, the null and research hypotheses must be describable in terms of the dependent and independent variables (in most scenarios, the research hypothesis is that the independent variables affect the dependent variables and the null hypothesis is that the independent variables do not affect the dependent variables). If there are no independent variables, the hypotheses must be describable in terms of the dependent variables only.

What are extraneous variables?

Extraneous variables are random variables that are not objects of research

Confounding: extraneous variables related to independent variables

Controlled: confounding variables manipulated so they do not affect the relationship between independent and dependent variables

Uncontrolled: confounding variables that have not been controlled

Assumed irrelevant: extraneous variables believed not to play any role in the research

Clearly, the research is at risk if there are important variables in the uncontrolled section or if the variables assumed irrelevant are confounding variables.

It is important that in stating our research problem that we give careful thought (and check the research literature) to allocate variables correctly. In practice, there are limits to how many variables we can control for, so we need to ensure that we control for those with the largest potential effect on our explanatory variables.

Controlling is an attempt to exclude the effects of variables that are not of interest.

Types of control

One category: fix all subjects to have the same value (category) of a controlling variable (e.g. only study males) Disadvantage is that the conclusions cannot be generalized to other values of the controlling variable.

Block control: divide up subjects by value of controlling variable and study separately (e.g. study males and females separately). Good for simple situations with only a few blocks but often impractical for multiple controlling variables as the block size gets too small (too few subjects per block).

Randomisation: assign subjects to groups or to study using chance (e.g. randomisation to treatments in a study of medical treatments) - assumes that assignment choice is possible. Randomisation is an invaluable approach because it means that the groups will be similar on uncontrolled variables and on variables assumed irrelevant

Paired control: pair subjects to be similar on several variables and randomly assign to groups from pair

Frequency control: match the groups on basis of distribution (e.g. choose groups to have same mean income) – crude means of control

Statistical control: try to control using a statistical model - this assumes that the statistical model is good enough to completely remove the effect.

In practice, none of these controlling mechanisms are perfect, because there may be errors in the measurement of the controlling variables. Example: early passive smoking studies looked at non-smoking women with smoking husbands, but did not account for the problem that some people lie about smoking (but not usually lie about not smoking) and that it is more likely for people to lie if you are married to a smoker, which together led to bias in estimate of passive smoking effects. This problem can be solved if we can model the measurement errors (in our example, we can check urine or hair samples to check how much people smoke).

Note that controlling methods that rely on allocation are only feasible if the researcher can allocate subjects to groups, which may be infeasible or unethical (e.g. smoking for humans).

The control variables may also interact (e.g. Effect of income within males may differ from within females).

II: Probability

What is probability?

Probability is the chance (long-run relative frequency) that something will happen. In other words, if you perform an experiment many times, what proportion of the time does a particular outcome occur?

Example: toss a coin, if the coin is fair then heads and tails are equally likely, so after a large number of tosses the relative frequencies should be close to one half, meaning long-run relative frequency should be one half for heads and for tails.

Notation: Pr(H) means Probability of the outcome H(ead)

While statistical formulae are largely irrelevant to users, the language of statistics (probability) is important and some knowledge is helpful. Anyone who gambles or invests should know the basics of probability – it is easy to show that failure to follow the rules of probability guarantees that you will lose on average when gambling with someone who does follow the rules.

What are the rules that help us work out the chances for combinations of outcomes?

Start with a set of outcomes that are mutually exclusive and exhaustive (i.e. they are unique and contain all possible outcomes) (e.g. if we toss a coin, 2 possible outcomes, Head or Tail), we call this set the sample space.

All probabilities must be in the range between zero and one (obvious in terms of relative frequency).

When do we add probabilities?

Probabilities of mutually exclusive outcomes add (also obvious in terms of relative frequency).

The sum of the probabilities of all the mutually exclusive and exhaustive outcomes must add up to 1, because one of the outcomes must happen.

e.g. Pr(H)+Pr(T)=1

Of course, for a “fair” coin, Pr(H)=1/2, because Pr(H)=Pr(T)

When do we multiply probabilities?

For independent events (no influence on each other), probabilities multiply.

So if we toss the coin twice, independently, and equal chances of a head each time (repeatable experiment), then

$Pr(HH)=Pr(H) \times Pr(H)$

For this double experiment, there are 4 outcomes, HH, HT, TH, TT

$Pr(HH)+Pr(HT)+Pr(TH)+Pr(TT)$

$=Pr(H)Pr(H)+Pr(H)Pr(T)+Pr(T)Pr(H)+Pr(T)Pr(T)$

$=Pr(H)(Pr(H)+Pr(T))+Pr(T)(Pr(H)+Pr(T))=1$

So, this is quite simple except for we usually are interested in the number of occurrences rather than the sequence. For this double experiment, we can get 0,1 or 2 heads in 2 tosses. So if p=chance of a head (and 1-p is then chance of tail), then

Pr(2 Heads)=Pr(HH)=p^2
Pr(1 Head)=Pr(HT or TH)=2p(1-p)
Pr(0 Heads)=$(1-p)^2$

This case is easy, but in general, we must calculate the number of combinations, e.g. if we toss a coin 10 times, what is the chance of 9 heads (and 1 tail)? Need to be able to work out the number of different ways that this can happen (10, because the tail could occur on each of the 10 tosses). Harder would be the chance of 8 heads (and 2 tails). Now the number of ways is 10 x 9 /2= 45, which is written mathematically as

${}^{10}C_2=10!\div(8!\times2!)$

where $10!=10\times9\times\ldots\times2\times1$ is the number of different possible sequences of 10 items.

Pr(10 heads)= p^{10}
Pr(9 heads)=$10p^9(1-p)$
Pr(8 heads)=$45p^8(1-p)^2$

and so on, but the formula is not important as we can use tables, calculator or a computer.

Binomial distribution

This set of probabilities for how many successes we get is called the binomial distribution and is very important in statistics. The usual notation is B(n,p), where n is the number of trials (tosses) and p is the chance of a success in one trial.

We will see how to use the Binomial distribution later on to check if a coin is "fair".

What is conditional probability?

The idea is to calculate the chance of an event A, if we already know that event B occurred.

Example: if I tossed a coin 5 times and you happen to see that the last toss was a head, how should that affect your estimate of how likely it was that I got all 5 heads in 5 tosses?

We write this as Pr(HHHHH | last toss is H)

We can solve this using a rule known as Bayes's Law

What is Bayes' Law?

Bayes' Law tells us that:

Pr(A | B)=Pr(A&B)/Pr(B)=Pr(B | A)Pr(A)/Pr(B)

It is easier to remember as:

Pr(A and B)=Pr(A | B)Pr(B)=Pr(B | A)Pr(A)

In the case where A and B independent, then we get:

Pr(A and B)=Pr(A)Pr(B)

In our example, we get:

Pr(HHHHH | last is H)
= Pr(HHHHH)/Pr(last is H)

$=p^5/p=p^4$

which is obvious in this case.

The important thing for this course is to "get" the idea of conditional probability and know of Bayes' Law (even if you need to look it up!)

Note: this law is very useful if you are a card player, whether bridge, poker, blackjack or the like as it lets you calculate the probability of an event given some indirect information.

Example: if I have a pack of well-shuffled cards and the first card I deal is an Ace, what is the chance that the second card is also an Ace if use the same pack without shuffling?

A1=First card is Ace
A2=Second card is Ace

Pr(A2 | A1)=Pr(A2&A1)/Pr(A1)

$=(4/52)\times(3/51)\div(4/52)$

=3/51, versus 4/52 if the pack was reshuffled

Bayes's Law is also the key idea behind many spam filters. They look at the probability of an email having these characteristics if it is or is not spam and then calculate the probability of it being spam given these characteristics. They update the probabilities as new emails come in (this is called the training process). That is why they are often called Bayesian filters.

Bayesian updating of evidence

Bayes’s Law can also be very useful in understanding how to update your evidence when collecting statistical information.

When there are two possible outcomes (Head/Tails, Win/Lose etc.), it is often easier to think of odds instead of probabilities, where

Odds (Heads/Tails)= Pr(Heads)/Pr(Tails)

We will see that Bayes’ law is easier in terms of odds when there are 2 possible outcomes.

So if we are looking at A or not A given B

Odds(A/not A | B) =Pr(A | B)/Pr(not A | B)
=Pr(A&B)/Pr(not A&B)

as the Pr(B) terms cancel out.

Example: you are evaluating whether it is a good idea to implement urine drug screening on all secondary school students in Hong Kong (about 400k) (this is loosely based on the real situation in Hong Kong, although the HKSARG seemed unaware of the need for a quantitative assessment like this and was unaware of the need for obtaining performance data for the selected screening tool, which was a relatively cheap drug test that could be done in the school, rather than in a laboratory):

Assume that before screening, the odds that someone has taken a specific drug are 1 in 1,000.

If the person has taken the drug, the probability of a positive screen is 0.98 (false negatives of 2%) and if they have not taken the drug, the probability of a positive screen is 0.02 (false positives of 2%) (these conditional probabilities were apparently not known by the HKSARG, so they are taken from data I found in an Australian study)

What are the odds an individual student has taken the drug after a positive or a negative screen?

$Pr(D)/Pr(\text{not } D)=10^{-3}$

$Pr(S \mid D)=0.98$
$Pr(\text{not } S \mid D)=0.02$
$Pr(S \mid \text{not } D)=0.02$
$Pr(\text{not } S \mid \text{not } D)=0.98$

What we want is:

$Odds(D/\text{not } D \mid S) \quad =Pr(D \mid S)/Pr(\text{not } D \mid S)$ and
$Odds(D/\text{not } D \mid \text{not } S) \quad =Pr(D \mid \text{not } S)/Pr(\text{not } D \mid \text{not } S)$

Bayes' law gives us:
$Pr(D \mid S)=Pr(S \mid D)Pr(D)/Pr(S)$
$Pr(\text{not } D \mid S)=Pr(S \mid \text{not } D)Pr(\text{not } D)/Pr(S)$

so
$Pr(D \mid S)/Pr(\text{not } D \mid S) \quad =Pr(S \mid D)/(Pr(S \mid \text{not } D) \text{ x } Pr(D)/Pr(\text{not } D)$

(statistical jargon is posterior odds = likelihood ratio x prior odds)

$=0.98/0.02 \text{ x } 10^{-3}$

which is about 0.05 or 1 in 20

$Pr(D \mid \text{not } S)/Pr(\text{not } D \mid \text{not } S)$
$=Pr(\text{not } S \mid D)/Pr(\text{not } S \mid \text{not } D)$ x prior odds

$=0.02/0.98 \text{ x } 10^{-3}$

which is about $2 \text{ x } 10^{-5}$ or 1 in 50,000

Is this a useful screen? Need to consider costs of wrong decisions and costs of data collection (i.e. each screening test), but unlikely to be useful here as the odds after a positive screen are still low (because of the high false positive rate).

Note: The government defended their position on the basis that all positive would be double checked in the laboratory using much better tests. However, the testing scheme would all know that a student tested positive on the screen, which would do great damage to the school if a false positive was leaked to the media.

III: Association and causation

What is association?

Association is observed linkage, i.e. two outcomes X and Y tend to occur together (positive association) or tend to occur separately (negative association). This is separate from the question of whether X causes Y, Y causes X or W causes both X and Y.

What is causation?

We say that X causes Y if when X occurs, Y must occur.

Note: if X occurs after Y, X cannot be the cause of Y - unless you believe in time travel ;)

In practice, we can usually only make probabilistic statements such as, when X occurs then Y is more likely to occur (than when X does not occur). Clearly, this means that something else must also be determining whether Y occurs, so our model is incomplete.

What we seek in practice is a model that enables us to predict (with high probability) individual outcomes.

Example: to predict which horse wins a race - this depends on the horse's ability, the track, the weather, the jockey's ability, the trainer's ability, and on which horses and jockeys it is competing with

Example: to predict which juvenile offender will reoffend depends on his background, family support and offending history

Example: to predict which patients will respond positively to a specific treatment depends on the medical history of the patient, genetic profile.

What is a necessary condition?

If Y only occurs if X occurs, then X is a necessary condition, so drinking alcohol is necessary to know if you are an alcoholic (unless we can provide a biomarker test in future!).

What is a sufficient condition?

Sufficient condition: If Y occurs every time X occurs, then X is a sufficient condition, so breaking your spinal cord is sufficient to cause paralysis.

What is the true cause?

If X is necessary and sufficient for Y, then X is the one and only true cause of Y. These are stringent conditions that are not often met!

Strictly speaking, we cannot move from association to causation, but only work the other way, that is, causation implies association, but not necessarily the reverse. This should be obvious, as association does not indicate the direction of causation.

In short, causation can be tricky!

Example: visibility and health outcomes – poor visibility is not a cause of poor health, but poor air quality is a cause of both poor visibility and poor health.

Example: exercise and physical disability – newspaper article claiming that lack of physical exercise was a cause of physical disability – true in theory, but the association is mainly a consequence that exercise opportunities are restricted for the physically disabled

Example: juvenile crime and reading violent comics – social workers believed that reading violent comics were a cause of juvenile crime, but life histories showed the major linkage was that juvenile crime led to detention, which led to expulsion from school, which led to hanging out on the street reading comics with similar youth!

Magnitude and consistency of association

If the magnitude of association is high, then we are getting closer to showing causation, but there are usually other causal factors and it does not address the direction of causation. In general, though, stronger associations are more likely to reflect causal relationships.

If the consistency is high, this means that the association appears under a variety of different conditions. This makes causation more plausible.

Example: the original study of smoking and lung cancer in doctors by Sir Richard Doll was very persuasive because he showed that the risk increased with the amount smoked and later showed that doctors who stopped smoking had cancer risk closer to non-smokers.

Experiment versus observation

In an experiment, we can use a wider range of control mechanisms (including randomisation, and paired controls), however, often we can only observe and select subjects, not control allocation of subjects to groups. Observation does not allow us to prove causation, but we can look for consistent patterns of association and identify lack of association. Designing good experiments is discussed in a later course.

Research plan

A research plan usually contains most of the following elements:

Literature Review – what is already known?
Research Design - overview of methodological decisions taken
Sample and Population - selection process and from what population
Operationalisation - description of the measuring instruments and pre-test, testing of reliability and validity
Data Collection - how done
Data Analysis - how to process using quantitative methods
Presentation of Results - how to display data and results of analysis.
Interpretation of Results - how to make sense of the findings
Work Schedule - timetable and allocation of tasks for study
Dissemination of Results - how to publish
Draft Instrument
Cover/Consent Letters - needed for ethical approval
(NOTE: Ethical approval is REQUIRED for research using human or vertebrate animal subjects) – see these URLs:
http://www.hku.hk/gradsch/web/student/ethics.htm
http://www.hku.hk/rss/HREC.htm
Bibliography – references to prior knowledge

We will focus on research design, sample and population, operationalisation and the basics of data analysis and presentation of results in this course (practical data analysis is a separate course).

IV: Research design

What is a population?

Population is the potential respondents of interest

Example: adults aged 18 or above, resident in Hong Kong

Example: air temperatures at the Hong Kong Observatory

What is a sample?

A sample is the respondents selected from population for study

Example: 500 mobile phone users in Hong Kong, selected randomly using their phone number

Example: 1000 average daily ozone levels recorded sequentially at the street level monitoring station in Causeway Bay

What are units of analysis?

Units of analysis are the sampling elements, often people or households (need to be clear which), but can be rats, words, songs – any countable objects.

What is representativeness?

Representativeness is the extent to which sample is similar to the population on characteristics of interest for research. This is essential if we wish to draw conclusions about the population based on a sample.

Question: how can we ensure representativeness? Cannot unless we use probability samples (see below).

What is a probability sample?

A probability sample is a sample where all sampling units have a known non-zero probability of selection. Probability samples are a requirement for all hypothesis testing and statistical models.

What is a simple random sample?

A simple random sample means that all combinations of sampling units with the specified sample size have an equal chance of selection. This also means that all sampling units have an equal chance of selection.

Example: a simple random sample of size 2 of students in a class with 10 students means that all 45 (10×9÷2) possible pairs are equally likely to be selected.

What is a cluster sample?

Cluster sampling means sampling at two or more levels (e.g. school, class, student), usually require that each individual has the same chance of selection overall.

Example: a sample of 500 secondary students is hard to draw without a list. However, we can easily obtain a list of secondary schools. We sample schools, classes within schools and students within classes. In practice, we often sample all the students in a class as the marginal cost of sampling is low once we have disturbed a class. Cluster sampling requires disturbing many less schools than a simple random sample – which would require obtaining consent from nearly all schools.

Disadvantage: students in the same class are likely to be more similar than students drawn at random (similar school philosophy, recruitment, teachers, social interaction). Intra-cluster correlation can be used to adjust the sample size.

What is stratified sampling?

Stratified sampling means sampling within subgroups of known size that are relatively homogenous

Example: Sample children separately in year groups because we usually know the number of children in each year group and children within year group are often quite homogeneous (as opposed to across year groups).

What is systematic sampling?

When the population can be placed in an ordered list, a systematic sample involves selecting a random start point and then selecting every kth individual where k=population size/sample size. This usually works well if the list order is associated with an important variable, but poorly if there are periodic patterns in the list. It is very efficient for sampling from large databases.

What is network sampling?

Network sampling is a sampling strategy that draws a sample through a random selection of links in a network. It makes assumptions about how people are linked in networks, which may not be valid, but has the advantage of sometimes being feasible when a simple random sample is impossible.

Example: it is very hard to sample drug addicts in general because they are not all easily identifiable, but they are usually connected in drug supply networks. A network sample can be drawn by asking for introductions to their drug-using friends and then asking those friends etcetera.

What is distance sampling?

Distance sampling involves randomly placing points or lines (transects) on a map and then measuring distances to objects.

Note: In this course the analysis will assume simple random sampling.

Telephone versus face-to-face interviews:

Most interviews now use either telephone or face-to-face. Telephone loses visual contact, gains on lower cost, better supervision, faster collection, ease of computer assistance (CATI stands for Computer Aided Telephone Interviewing, which automates question display, telephone number selection etc.). In Hong Kong, fixed line telephone penetration is still very high (85%+) and it can be difficult to gain entrance to private housing estates, although the compact geography makes face-to-face interviews in households possible. For children, schools provide a context for face-to-face or self-report.

Mobile versus fixed-line telephone surveys:

Mobiles provide an alternative to fixed lines as coverage of fixed lines is starting to drop and be replaced by mobile. However mobile is linked to individuals, while fixed is linked to households, so the sampling unit is different. For mobile, coverage is still quite low for the elderly, while for fixed lines, the coverage is quite low for young households and the unemployed.

Primary versus secondary data:

Secondary data analysis means using data previously collected. Data collection is very expensive and usually not repeatable for social science data, so use of previously collected data or materials is a good idea, if feasible (e.g. use data archive). Thee are data archives in UK, US, Norway etc. that mainly cover socio/economic/political data, but also geological, oceanographics etc. Real importance is when data is expensive or impossible to collect again.

Observation versus participation:

People and organisations may respond differently when they are aware of data collection, but observation may not provide some of the detail. Conversely, observation may be essential to see how people behave, rather than how they think or tell others they behave!

Example: study that drops envelopes (with or without stamps) in different cities and sees what proportion of them is posted as an assessment of social behaviour in different cities.

Qualitative versus quantitative:

The dividing line is sometimes not clear, but one distinction sometimes made is the type of data collected - directly measurable (quantitative) vs. recordable (text, audio, video etc.) (qualitative). However measurement can often be applied to text etc. by counting events (e.g. word occurrence), so the same data may be used for different types of analysis. The qualitative paradigm is arguably more concerned with context than counts and provides richness not easily achieved with quantitative measures. Generalisability is much harder with qualitative analysis because it does not use probability samples (see below). Often use mixed methods (combination of qualitative and quantitative) by using qualitative first (identifying the issues) and then quantitative (measuring responses for the identified issues).

V: Basics of qualitative research

Qualitative research is an inquiry process of understanding based on a methodological tradition of inquiry that explores a problem, which enables construction of a complex, holistic picture, analyses words, reports detailed views of informants and conducts the study in a natural setting. Qualitative research usually involves many variables and few cases (versus many cases and few variables for quantitative research).

Table of Qualitative methodologies[1]

Dimension	Biography	Phenomenology	Grounded Theory	Ethnography	Case Study
Focus	Explore life of individual	Understanding essence of experiences about a phenomenon	Develop theory grounded in data from the field	Describe and interpret a cultural or social group	Indepth analysis of a single or multiple cases
Disciplinary origin	Anthropology	Psychology	Sociology	Cultural anthropology	Political science
Data collection	Interviews and documents	Long interviews with up to 10 people	Interviews with 20-30 individuals to saturate categories and detail a theory	Observations and interviews during extended fieldwork (e.g. 6m-1yr)	Multiple sources including documents, interviews, artefacts
Data analysis	Stories, epiphanies, historical context	Statements, meanings, themes, general descriptions	Open, axial, selective coding, conditional matrix	Description, analysis, interpretation	Description, themes, assertions
Narrative form	Detailed picture of individual's life	Description of essence of experience	Theory or model	Description of cultural behaviour of group or individual	In-depth study of case or cases

There is a very wide range of methodologies (approaches to collecting and analysing qualitative data), which include:

[1] Taken from "Qualitative Inquiry and Research Design" by John Creswell

Biography or narrative research

This refers to the collection of people's stories about experiences that have a significant impact on their lives.

Example: Cancer patients reveal their own experiences of being treated for cancer, which offer insights into different aspects of the care and the people who provide it.

Phenomenology

This is observing what happens in order to gain understanding of what really happens rather than what people tell you happens. It enables the researcher to gain some understanding of what it feels like for the subjects to be living a particular experience.

Example: "If you want to know why athletes are willing to take steroids – you need to understand their lived reality of winning and losing. If you want to help someone through breast cancer – you need to know how they feel about their body, their self esteem, their future. If you want to understand how you can help motivate struggling students – you need to know what it is really like for them at the bottom of the class." (O'Leary, Z. (2010). *The Essential Guide to Doing Your Research Project*. London: SAGE, p.119)

Grounded theory

This is a form of analysis constructing new theories from the data, i.e. qualitiative induction. It generally consists of four stages (Glaser, B. & Strauss, A. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Chicago: Aldine):

1. Observing the data to identify patterns that lead to the emergence of categories, then identifying the underlying properties of the categories.
2. Integrating categories and their properties: Comparing an incident to the underlying properties of the category.

3. Delimiting the theory: Fewer and fewer modifications are needed as the categories are confirmed, then the number of categories can be reduced as further refinements take place. In this way the theory begins to solidify.
4. Writing theory: Hypotheses and generalizations emerge from the analysis, as opposed to starting with and testing hypotheses.

Example: O'Reilly, Paper & Marx (O'Reilly, K., Paper, D. & Marx, S., Demystifying grounded theory for business research. *Organizational Research Methods.* 15, 247, 2012) described a study of business segments that lack effective communication and cooperation. They wanted to understand the views, perceptions, and beliefs of front-line employees (FLEs) and how these perspectives might help improve customer-company interactions. They examined differences in service levels, service outcomes, and service attitudes of the participant companies and FLEs *who work* there and, by comparing the participants' stories they identified the constraints within the organizations and their impacts.

Ethnography

This is the anthropological approach of becoming part of the culture in order to understand it.

Example: working in a factory in order to understand what it means to be a worker both from direct experience and asking/observing others.

Case study

This involves studying a small number of cases in great depth in the expectation that this gives deep insights into the process

In all cases, the focus is on understanding the full scope of the problem, rather than quantifying the problem. Arguably, qualitative research is a necessary precursor for quantitative research unless the scope of the problem is already well understood (e.g. a study aiming to understand how sports participation has changed over time might not need qualitative research unless we think the underlying drivers of participation might have changed).

Example: Silverman (Silverman, D. (2006). What is Qualitative Research? http://www.sagepub.com/upm-data/44074_Silverman_4e.pdf and http://www.sagepub.com/upm-data/11254_Silverman_02.pdf) conducted a case study of British cancer clinics to form an impression of the differences in doctor-patient relationships when the treatment was private or public. He cautioned that his data could not offer proof of the differences he identified, but that they provided strong evidence to support them.

Qualitative sampling

Probabilistic notions of sampling are not relevant for some methodologies. It is not always possible to achieve representative sampling because of the exploratory nature of the research and the sheer logistics. Depending on the nature of the research, it is sometimes necessary to select a sample that meets a particular need. Usually, with this kind of sampling, it is not appropriate to generalize the data to wider populations. The key ideas of qualitative sampling are:

Saturation

The idea is to collect data until no new perspectives are being obtained. This means that the sample size cannot always be predetermined.

Example: You are interested to find out how young children learn to write Chinese characters. You observe children engaged in writing tasks and analyse samples of children's work. Eventually you become aware that the same patterns are being repeated and there is nothing new that has not arisen before. There is no point to collect further data as no new perspectives are being obtained.

Theoretical or purposeful sampling

The idea is to select a sample with the intention of collecting a wide range of responses by sampling across all factors likely to influence outcomes.

Example: A mathematics department in a particular school has an excellent reputation for students achieving high scores on national tests. The researcher chooses this particular department with a particular purpose in mind. There would be no point investigating a random sample of mathematics department because they will not necessarily show the characteristics of interest for the research (Plowright, D., 2011. *Using Mixed Methods: Frameworks for an Integrated Methodology*. London: SAGE)

Convenience sampling

This is sampling driven by the feasibility and convenience of the selection process. Some people criticize that it does not have a place in 'credible research' (O'Leary, 2010), but it may be the only option for a small, low-budget study or a pilot.
Example: A group of recent graduates is invited to volunteer to attend an interview about the impacts of their undergraduate programme on their professional lives. Only a limited number of students is able to be contacted or willing to make themselves available, so the researchers need to utilize those who can be accessed and are willing to participate.

Snowball sampling

Snowball sampling assumes relevant respondents are connected so that we can use those connections to construct a sample from a small initial sample. In other words, it involves building a sample through referrals, as each respondent recommends others.

Example: A population of homeless people might not be easy to identify, but a sample can be built by using referrals (O'Leary, 2010).

Observation, interviewing and other means of collecting qualitative data

Qualitative data collection does not usually follow such strict predetermined rules as in quantitative methods, but is more concerned with obtaining a complete picture within the agreed domain. This necessarily requires that the observer/interviewer is well trained in engaging in the data collection process and understands the domain well enough to ensure collection deep, relevant data.

Observation

Observation is the collection of existing data. It usually takes place in a real situation, not a contrived context and captures first-hand what people actually do in the situation as opposed to telling the researcher about what they do.

Example: A school district introduces a child-centred learning approach and wants to collect authentic data about how the teachers are actually implementing this approach.

Interviewing

Interviewing can be used to provide rich qualitative data and provides flexibility to explore tangents (O'Leary, 2010). Good interview questions can elicit data about who, when, why, where, what, how and with what results (Hutchison, A., Johnston, L. & Breckon, J. Using QSR-NVivo to facilitate the development of a grounded theory project: An account of a worked example. *International Journal of Social Research Methodology*. 13, 4, 283-302) Interviews can be structured, semi-structured or unstructured.

Example: What are the perceptions of carers living with people with disability, as regards their own health needs? (Lacey, A. & Luff, D. (2009). *Qualitative Data Analysis*. The NIHR RDS for the East Midlands/Yorkshire & the Humber. http://hk.bing.com/search?q=the%20NIHR%20RDS%20for%20the%20East%20Midlands/Yorkshire%20%26%20the%20Humber%202009%20Qualitative%20Data%20Analysis&FORM=AARBLB&PC=MAAR&QS=n)

Recording and analyzing qualitative data

Audio and video recording enable the raw data to be recorded for later review, and are often used to ensure that the full context is collected, such as the tone of voice, hand gestures etc. Notetaking can range from highly structured (codes to represent common responses, concept maps) to open and interpretive (jotting down extensive notes during an interview or an observation).

Analysis does not usually start from pre-defined hypotheses, but instead tries to produce undistorted non-judgmental summaries of the issues, accepting that different people/organizations may frame issues in very different ways.

Example: In the research about the perceptions of carers living with people with learning disabilities about their own health needs, there are different ways in which the data could be analyzed depending on what the researcher is interested to explore (from Lacey & Luff, 2009):
Content analysis: Count the number of times a particular word or concept (eg loneliness) appears – categorise these quantitatively and do statistical analyses.
Thematic analysis: Find all units of data (eg sentences or paragraphs) referring to loneliness, code them and look for patterns (eg certain times and conditions where carers feel lonely).
Theoretical analysis (eg grounded theory): This goes further to develop theories from the patterns in the data. It may include data that contradict the theory. Gradually the theory is built and tested.

Computer-Assisted Qualitative Data Software

Computer-Assisted Qualitative Data Software (CAQDS) can be used to help researchers to analyze their data but they cannot analyze the data for researchers. The researcher also needs to exercise flexibility, creativity, insight and intuition (Denzin & Lincoln, 2005, Eds. *Sage Handbook of Qualitative Research.* (2nd ed.) Thousand Oaks, CA: Sage, p.578).

Examples of CAQDS are NVivo 9; QSR International Pty Ltd, 2011; QDA Miner 3.2; Provalis Research, 2009. Excel and SPSS can also be utilised for qualitative analysis, but they are not designed for that purpose.

VI: Measuring instruments

Sampling versus non-sampling errors:

These errors indicate the difference between the sample and the population. Sampling errors relate to the use of a sample rather than the whole population. For probabilistic sampling, the sampling errors can be easily quantified in a probabilistic way and related to the sample design and the sample size (sampling accuracy increases proportional to the square root of the sample size, or equivalently the sampling error decreases proportional to the inverse of the square root of the sample size). The non-sampling errors include non-contacts, refusals, misunderstandings, lies, mistakes, coding errors etc. Increasing the sample size reduces the sampling error, but increases the cost and may even increase non-sampling error as it gets harder to supervise the data collection process adequately. It is essential that the research design takes into account non-sampling errors as well as sampling errors, as the non-sampling errors are particularly damaging to research as they are hard to quantify or take into account. Note that even including all respondents does not help with non-sampling errors. We may be able to get some idea of the severity of nonresponse bias by seeing how responses differ by number of contact attempts.

Example: any telephone survey that samples numbers directly from the Chinese language telephone directory in HK misses many Chinese writing families that prefer not to be easily found and all families that do not write Chinese so cannot claim to represent the Hong Kong population.

Example: any household survey that fails to make repeated (at least 5) attempts to contact households will under represent poor and single working households (who are often out at work) and young households without children (out at work or play).

Example: a self-selected Internet survey omits most of the population who are not interested in the topic, do not see the survey or do not trust the website, so is likely to be of very limited research value.

This is not just a problem with social research, although the problems are most often acute in that context. For example, if you fail to calibrate your equipment before use in a laboratory, there is a risk of bias affecting all observations and taking more measurements does not reduce this bias. Similarly, if you use equipment that is not fit for the proposed measurement purpose, this cannot be fixed at the statistical analysis stage.

Three key criteria for a measuring instrument: reliability, validity & precision

What is reliability?

Reliability in this context (it has other meanings in statistics and in daily life) means consistency: do we get the same result if measured repeatedly - various types of reliability:

Test-retest: ask respondents again, i.e. measure a second time using the same sample and same instrument.

Split-half: compare results from different sets of items (applies to survey instruments, where we often assess using a combination of questions)

Inter-rater: compare the same sample across different interviewers/instruments

Other more mathematical definitions: look for consistency across time, instruments/interviewers, items etc.

What is validity?

Validity is about whether our measurement really measures our concept (or something else)? Is it meaningful as a measurement tool for this concept?

Various types:

Face validity: does it even look like it is measuring the right thing (e.g. Assess weight by asking how much money in your pocket!)

Criterion (predictive) validity: does it predict outcomes that we believe relate to the concept (e.g. If heavy people have more diabetes, does our measure of weight predict diabetes?)

Construct validity: does it relate to other variables in the way our theory predicts? (e.g. if we are measuring marital satisfaction and our theory says that it should be associated with marital fidelity, is that true?)

Content validity: does it cover the range of meanings contained in the concept (e.g. the concept prejudice contains prejudice on grounds of race, minority, gender etc.)

What is precision?

How precise is the measurement - e.g. is age measured to the nearest decade, year, or month. There is little point in having precision that is not needed, but if the precision is too low, cannot fix it later. Depends on situation, so age in months is too broad for new-born babies, too narrow for the elderly! Often has little impact on cost, unless it is sensitive data which people are reluctant to reveal (e.g. exact age or income). However, very precise measurement may be impractical, e.g. exact income may be hard to calculate for people who do not have a fixed monthly income.

Note that there may be some trade-off between reliability and validity. Highly reliable measures are likely to be factual and give little insight and may not have high validity (they are often quantitative). Measures that have high validity may be very personalised and have relatively low reliability (they are often qualitative). A good strategy is often to use multiple measuring tools simultaneously to measure different dimensions.

Operational choices:

Unipolar versus bipolar: e.g. neutral to positive or negative to positive (note that the midpoint of a bipolar scale can be tricky - neutral and indifferent may not be the same, in Chinese often use “neither agree nor disagree”). This can make it difficult to compare across response scales – consider the difference in asking how much you support something (from not at all to complete support) versus asking how much you agree with something (from completely disagree to completely agree)

Detail: how much detail is useful and collectable, not just units (as in precision), but detail of categories

Example: Single, married, others - do you need to break down others? Interesting study of mental health – for men, deterioration after divorce or widowhood, while for women, deterioration after widowhood, but not divorce!

Dimensions: a concept may have many dimensions, which are of interest in the research?

Example: corruption: how much, what causes it, what should be done, what would they do personally etc.

Example: when assessing outcomes from medical treatment: life/death, survival time after treatment, quality of life, pain, blood pressure, white cell count etc.

Levels/scales of measurement:

Attributes must normally be mutually exclusive (no overlap) and exhaustive (cover all possibilities)

Example: underemployed, employed or unemployed, cannot fit into more than one category and must fit into at least one category if economically active.

What is nominal scale?

Nominal scale means that there is no ordering of attributes

Examples: gender, religion, race

What is ordinal scale?

Ordinal scale means that attributes can be ranked (although the ranking may be arguable, consider education which can be hard to compare across systems)

What is interval scale?

Interval scale means that differences between attributes have consistent meaning

Examples: temperature, net profit

What is ratio scale?

Ratio scale means that we have an interval scale, plus a meaningful zero

Examples: income, height, weight, no. of hospital visits, area of flat

These levels of measurements are ordered, where the later in the order, the wider range of statistical tools that are usable. You need to ensure that you have used a level of measurement that allows the statistical analysis that you wish to use (or vice versa).

Making operational choices (how do we measure something?)

Some examples of operational choices:

Research Question: Is the lecture room at a comfortable temperature?
Concept: “comfortable temperature in the lecture room”
Possible operationalisations:

1) use the thermometer in the room (which uses indirect physical measurement using the length of a metal strip based on how the metal expands as the temperature increase) Reliability is high but validity is questionable – one person may be wearing a T-shirt and another person is wearing a coat and may feel very differently about what is a comfortable temperature
2) ask everyone in the room how warm they feel on the scale: very cold, cold, about right, warm, very warm. The subjective scale would take into account how people feel, how they are dressed, whether they are sitting under the aircon vent, whether they are feeling unwell. Validity is high, but reliability is questionable – on another day, I may come in wearing more or less clothes

Research Question: Is HK or Singapore a better place to live?
Concept: "quality of life in a specified place"
Possible operationalisations:

1) migration rates – reliability: in HK and Singapore we know population flows quite accurately, but when someone leaves, we do not know if is temporary or permanent, while in Europe, do not know population flows accurately because of weak national border controls and in US have many uncounted illegal immigrants – validity: positive is that if it is a good place to live, we expect people to come and if it is a bad place, we expect people to leave, negative is that it can be hard to get approval to move countries legally unless you have skills, money or family reunion
2) poll people in HK, Singapore and a 3rd place – reliability: depends heavily on media reports, if they change then the poll will change – validity: is it based on personal experience as a resident or tourist, family, friends or media reports, all of which may be biased
3) suicide rates – reliability: death rates are very reliable, but not all suicides leave notes – validity: suicide can be seen as only way to escape terrible life, but it also can be seen as untreated mental illness, so arguably both elements imply very poor quality of life for some people
4) air pollution levels – reliability: quite high as based on physical/chemical processes – validity: high for people

who are asthmatic, but clearly this is only one element in quality of life

5) ask experts to rate each place on a set of living characteristics – reliability: probably high if they are experts on those characteristics – validity: how do we know if those characteristics are relevant to people considering living in that place?

Note: quality of life has a strong personal element, so it is possible that for so people Singapore gives higher quality of life (those who value air quality?) while for other people Hong Kong gives a higher quality of life (those who value freedom of speech?)

There are many other important operational issues for questionnaires including clarity of questions, ensuring that respondents are competent to answer, questions should be relevant, keep questions short, avoid negative items, avoid biased items (wording is critical, positive vs negative words). Questionnaire design is an art as much as a science and it is essential to do a pilot (pre-test) to check that the questionnaire works as designed (pre-tests usually involve qualitative analysis)

(Questionnaire design is worthy of a course itself and the Faculty of Social Science runs a course each summer)

Similarly, when choosing how to measure something in a laboratory, it is very important to choose the right equipment for the task, which may require considerable knowledge of the strengths and weaknesses of different pieces of equipment – accuracy, stability, ease of calibration, portability, sensitivity, robustness, let alone capital and recurrent costs!

Index versus scale measures

They are both ordinal measures.

Index counts the number of positive (or negative) responses, while a **scale** allows for intensities (levels of response). When creating an index, we want the underlying items to be measuring the same thing - usually weight items equally (i.e. just add them up). Often do item analysis (reliability check if each item is consistent with the rest of the items) and external validation of complete scale. In Hong Kong, often take scale from other cultures/languages, but the translated scales may not work in HK (concepts may not translate or attitudes may be different).

What is a Likert scale?

Likert scale means using ordered response levels labelled to ensure clear ordering

Example: strongly disagree, disagree, agree, strongly agree

Note: Likert scale does not guarantee that the difference between strongly disagree and disagree is the same as between agree and strongly agree, so should not treat it as interval scale without checking if this is valid

What is semantic differential scale?

Semantic differential scale means asking people to choose positions between two polar opposites

Examples: love to hate, simple to complex

What is Guttman scaling?

Guttman scaling looks for items of increasing extremity and see if people respond consistently. If the ordering is consistent, then use this to create a scale.

Example: everyone who agrees to abortion on demand should logically also agree to abortion after rape

VII: Quantitative research (statistics)

Key Elements:

1) Counting
Example: how many boys in our sample?

2) Testing Hypotheses
Example: are there more boys than girls in the population?

3) Sample size and power
Example: how big a sample do I need to have a good chance of reaching a useful conclusion from my experiment?

4) Modelling Variability
Example: how much do Age & Gender determine academic results?

5) Prediction
Example: can I predict whether a boy aged 16 from a band 1 school will pass an exam?

This course only has time to cover the first 3 elements

Computer packages for statistical analysis

There are many packages for personal computers that provide the methods needed for this course. They include:

SPSS - very popular for desktop use, but expensive and not very flexible. It has a student edition, but it is quite limited in functionality and has dataset size limits, - the HKU license for the full version is a much better deal (about HK$500 for a 3 year license to Base +Regression+Advanced Statistics). The license is purchased from IT services through departments.

SAS - very popular for large computers, very flexible but less user-friendly than SPSS

Stata - very popular for PCs and Macs –popular with epidemiologists and easy to add new procedures

JMP - very user-friendly. It has a free demo version available at http://www.jmp.com. There is a free student version that comes with some books – see http://www.jmp.com/academic/books.shtml. The student version has a few limitations, in particular an expiry date of 14 months after installation. You can also purchase an educational license for the full version at http://www.onthehub.com/jmp/ for US$49.95 for a 12 month license (US$29.95 for 6 months).

R – open source, very popular with statisticians, but not so easy to use

All of these have PC and Mac versions.

I teach practical (hands-on) statistical methods in my Applied Quantitative Research Methods course for the Graduate School using JMP and use JMP to generate the results in this book.

Please note that thanks to computers and statistical packages, we rarely need to do statistical computations by hand and hence the formulae are largely irrelevant.

What do we need to understand when using statistical methods?

1) Appropriateness: which methods to use? - depends on how data were collected and what type of data.

2) Interpretation: how to interpret the results? - need to relate back to theory.

3) Diagnostics: how to check whether the methods are appropriate - if not appropriate, want to find a better method.

VIII: Summarising data

Graphical data summary:

We can summarise data graphically or using numbers. For graphs, we will examine the stem-and-leaf display, bar chart, histogram and box plot.

Stem-and-leaf: display sorted list of numbers, where for 2 digit numbers, the tens digit is the "stem" and the ones digit is the "leaf" to illustrate the data frequencies (the row length) without losing sight of the underlying data. Easy to do by hand with paper and pencil!

Example: we collect data on the size in cm of 10 objects (14,12,16,24,20,40,42,46,49,46), the display is:

```
4|02669
3|
2|04
1|246
```

This shows that there may be two distinct populations.

Notes: stems and leaves must be sorted, must include duplicates, stems with no leaves must be included.

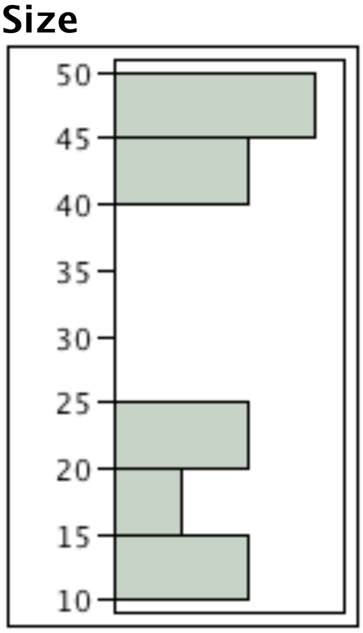
Variations include splitting the stems into 2 stems, i.e. put leaves 0-4 on one stem, 5-9 on another, so the previous plot becomes

4 | 669
4 | 02
3 |
3 |
2 |
2 | 04
1 | 6
1 | 24

Bar chart: chart with height of bars proportional to frequency (standard chart in Excel)

Histogram: bar chart with area of bar proportional to frequency (area is equivalent to height if we have equal widths)

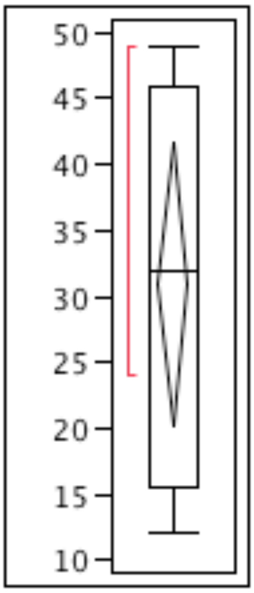
Example: We choose (equal) group widths of 5, so the bar chart and histogram are both:



Box Plot: box shows central half of the data, with a line showing the middle of the data and with whiskers and dots showing more extreme data.

Example:

Size



Note that this does not show the separation of the data.

Numerical summaries for centre of a distribution:

What is the mean?

The mean is the average value (i.e. Add all the values up and divide by the number of values). We often use the Greek symbol mu: μ for the population mean, often write the sample mean as a X with a bar above it, $\bar{X}$ and the formula for the sample mean as $\sum x_i/n$, which means adding up all the values and dividing by n, the sample size.

What is the median?

The median is the value such that half the values are less and half the values are more. It is shown as a line in the box on the box plot.

If n is the sample size and is odd, the median is the (n+1)/2 biggest (or smallest) value, if n is even number of values, the median is the average of the 2 middle values (the n/2 and n/2+1 values). The median can easily be read from the stem-and-leaf plot, as the data points are already sorted.

What is the mode?

The mode is the most common value.

Comparing the mean, median and mode

The mode is not always unique and is not robust (changing one data value slightly can cause a large change), so usually use mean or median.

Mean is easier to deal with mathematically, so most theory is based on the mean, but the median is more robust (reliable).

Example:

```
Size
Median     32
Mean       30.9
Mode       46
```

If we now change the largest value by a factor of 100, we get:

```
Median     32
Mean       75
Mode       46
```

The mean changes substantially, but in this case, the median does not change at all. Also transforming the data usually transforms the median, so the median of log(income) is the log of the median of income, but this does not work for the mean.

Numerical summaries for spread/deviation of a distribution:

What is the variance?

The variance is the average squared distance of data values from the mean. This seems complicated, but it gives a simpler theory. Population variance is denoted σ^2 (sigma squared) and sample variance is denoted as s^2. The formula for the sample variance is written as $\sum(x_i\text{-mean})^2/n$.

Note: if using sample mean instead of population mean, we adjust the divisor to be (n-1) because one data point is “used up” by the sample mean)

What is the standard deviation?

The standard deviation is the square root of the variance, i.e. measure of spread in the original units. Often use Greek symbol sigma, σ for the population standard deviation and s for the sample standard deviation.

What is the interquartile range?

The interquartile range is the distance between lower and upper quartile. Quartiles are the data values that divide up the distribution into quarters in the same way that the median divides it into halves. It is the size of the box in the boxplot.

Example:

Size:	
Standard deviation	**15.0**
Variance	**225**
Interquartile Range	**30.5**

IX: Estimation and Hypothesis testing

Estimating proportions

If the sample is representative, then we can assume that our sample will on average be similar to the population, i.e. the sample proportion should be a good estimate of the population proportion, so that the sample proportion (of say males) times the population size should be a good estimate of the population size (of say males).

Note: this idea only works if the sample is representative, i.e. all individuals have equal chance of selection. If the individuals all have known (non-zero) chance, we can adjust (weight) respondents and still estimate population proportions.

We can also provide an idea of how accurate our estimates are.

We will look at an example before introducing the concepts formally.

Example: Random sample of 1000 school children in HK.

Say we observe 520 boys

What can we say about the proportion of school children in HK who are boys?

Our best estimate is 520/1000=0.52, i.e. the sample proportion is our best estimate of the population proportion.

How accurate do we think the estimate is?

We will simplify things slightly for now, by assuming that the total no. of children in our population is much greater than 1000 (e.g., if only 1000 in the population, then our estimate would be exact). It is rare that our sample size is more than 1% of the population size.

Note: If the sample size is more than 1% of the population, we should make an adjustment to the following calculation, as the estimate will be more accurate than we estimate here.

The jargon we use for describing the accuracy of our estimate is that we say that the approximate 95% Confidence Interval (C.I.) for the population proportion is:

$$0.52 \ {+/-}\ 2 \sqrt{(0.52 \times (1-0.52) / 1000)}$$

The formula for this approximate answer is:

$$p \ {+/-}\ 2 \sqrt{(p \times (1-p) / n)}$$

where p is the sample proportion and n is the sample size. This approximation is good as long as the sample size is at least 30 and the sample proportion is not too close to either 0 or 1. Otherwise, we need to use statistical tables.

This gives us 0.52 +/- 0.032

In percentage terms, our interval contains values between 48.8% and 55.2%

Calling this interval the 95% C.I. means that, on average, we would expect that in 95 out of 100 samples, our confidence interval would contain the population proportion. This is already a hard concept as it is not a probability statement about our specific interval, but instead what would happen if we repeat our sampling.

We will come back to these ideas in a more careful way later and explain where the formula comes from.

Testing hypotheses about proportions

If our concern was whether there was a sex imbalance in the schools, our hypotheses would be:

Null: Males and females equal

Alternative 1: males and females unequal

or Alternative 2: males more than females

or Alternative 3: females more than males

We assume Alternative 1 for now

Question: is there evidence that our null hypothesis is false (relative to our alternative)?

Idea: find a good summary of the data (called our test statistic) that best summarizes the evidence for choosing between the hypotheses. In this case, the sample proportion is the best test statistic. If the test statistic has a value that is unlikely to occur if the null hypothesis is true, then we reject the null hypothesis (strictly speaking we mean that this value or more extreme values are unlikely, rather than that one value is unlikely).

Implementation: find the probability of observing the test statistic (or a more extreme value) if the null hypothesis is true. If this probability is below a cut-off value (the cut-off is called the significance level), we reject the null hypothesis, on the grounds that what we observed is not consistent with the null hypothesis.

In this case, the probability of observing a sample proportion of 0.52 or a more extreme value (i.e. More than 0.52 or less than 0.48) is about 0.2 if the population proportion is truly 0.5 (this probability will usually be given in the output from our statistical package).

The usual significance levels used are 0.05 (5%) or 0.01 (1%). This means that we choose to reject the null hypothesis when it is really true 5% (1 in 20) or 1% (1 in 100) of the time. Which significance level we use as cut-off values depends on what risk we accept for making the mistake of rejecting a true null hypothesis (we will return to this topic more carefully later as well).

Population to sample

Given some knowledge of a population, what can we say about samples from that population?

The average of the sample means tends towards the population mean as the sample size increases as long as we have a probability sample.

The sample variance is a good estimator of the population variance for sample sizes greater than about 30.

The standard error (the standard deviation of the distribution of the sample mean for repeated samples) is the population standard deviation divided by the square root of the sample size.

The distribution of the sample mean is approximately Normal (a known distribution) for sample sizes greater than about 30.

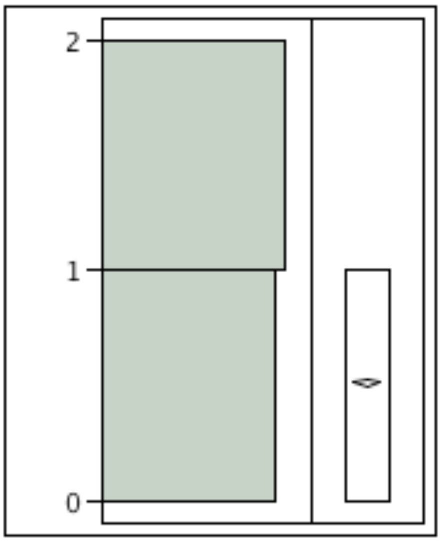
None of these statements require us to know the full distribution of values in the population – however, the population distribution is important if we deal with small sample sizes (less than 30), when we cannot use some of our approximations without careful checking.

Example:

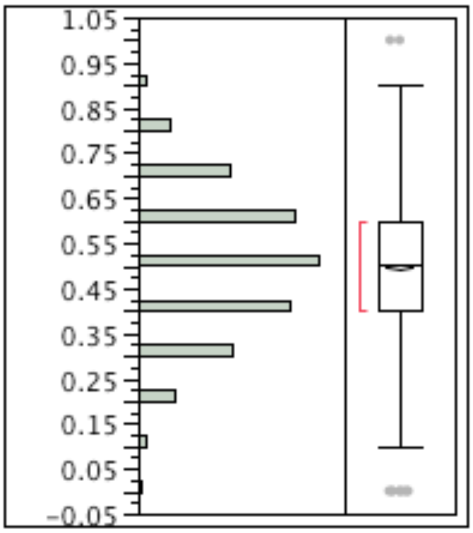
We now do a simulation of tossing a coin. We will do 5000 simulations of different experiments.

First, we consider a fair coin (i.e. Pr(H)=0.5). Let us compare the histogram and box plot we get for the proportion of Heads in our 5000 simulations if we do 1 toss, 10 tosses, 20 tosses or 80 tosses in each experiment

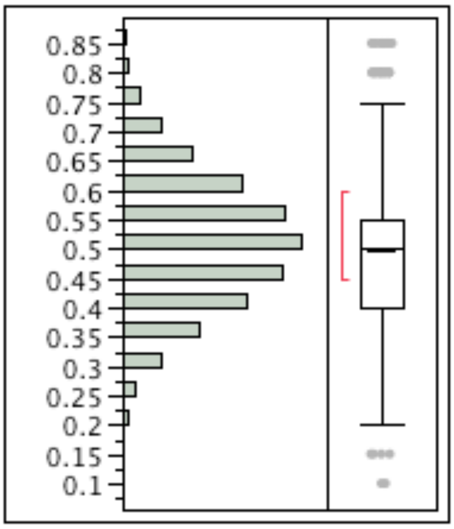
Toss 1



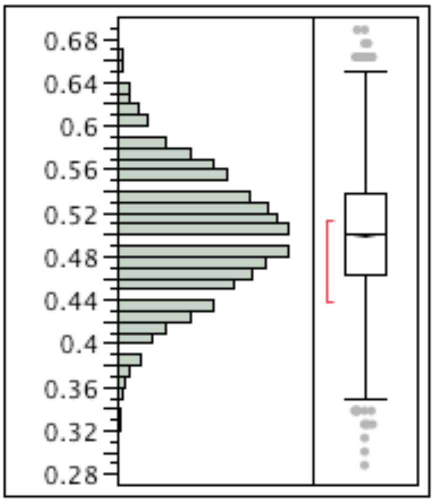
Toss 10



Toss 20



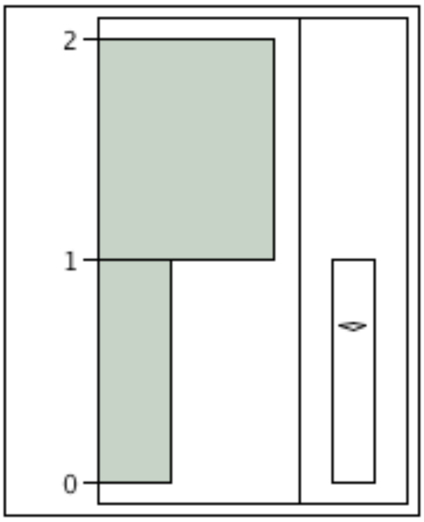
Toss 80



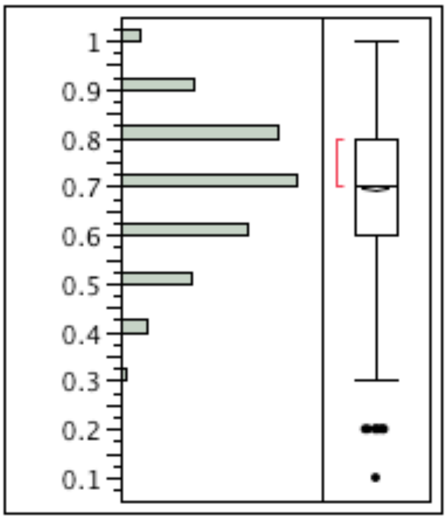
You can see that even with 20 tosses, the histogram looks bell-shaped and that as we increase the sample size, the spread (width) decreases, while the mean and median of the distribution is close to 0.5 in all cases.

We now repeat our simulations with a very unfair coin (Pr(H)=0.7)

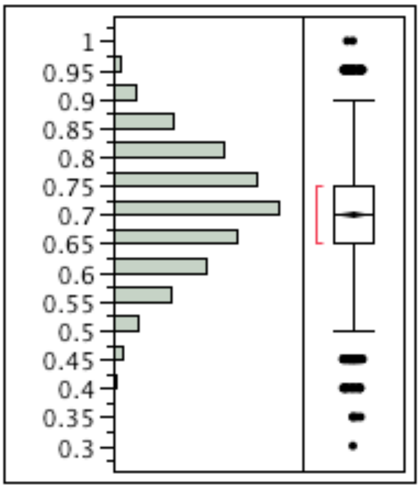
Toss 1



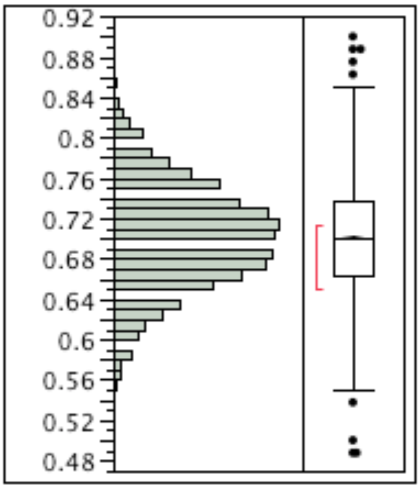
Toss 10



Toss 20



Toss 80



Notice that once the sample size increases, we again get a symmetrical histogram and decreasing spread, this time with the mean and median close to 0.7.

Sample to population

As the sample mean gets very close to the population mean for large samples, we say that the sample mean is a good estimator of the population mean.

The sample mean can be used as:

1) a point estimator for the population mean
2) the centre of an interval estimator (confidence interval) for the population mean
3) the basis for a hypothesis test of whether the population mean has a particular value.

In all cases, we are using our knowledge about how the population relates to the sample to make reverse statements about the population from a selected sample.

Sample theory

Let us now try to be more systematic about how we handle confidence intervals and hypothesis testing. We need to understand the ideas behind some statistical theory called sample theory (without the maths!).

If we look at the mean of a random sample (called the **sample mean)**, it turns out that:

The mean of all possible sample means is the population mean (we call this property unbiasedness, meaning that we get the correct answer on average)

The variance of the sample mean is the population variance divided by the sample size, or equivalently, the standard deviation of the sample mean (called the standard error) is the population standard deviation divided by the square root of the sample size.

Note: this assumes that the population size is much bigger than the sample size, otherwise we need to multiply by a correction factor = (1- sample size/population size). This is intuitive in that a random sample (without replacement) the same size as the population must have the same mean as the population! However, as long as the population size is much larger than the sample size, it does not have much effect on the accuracy of a random sample.

One very important further issue:

The distribution of the sample mean is approximately Normal (Gaussian) (i.e. bell-shaped with known centre, spread and shape). The approximation is very good for sample size of 30, (almost) regardless of the original distribution of values in the population.

This means that we can use theory developed for the Normal distribution in many situations with samples, as long as our interest is in the population mean.

For example, we can construct interval estimates:

The 95% Confidence Interval (C.I.) for the population mean using the Normal distribution for the sample mean is:

sample mean +/- 1.96 x standard error (often round 1.96 to 2)

The 99% C.I. is:

sample mean +/- 2.65 x standard error

We will explain where the 1.96 and 2.65 come from later

For another example, hypothesis testing:

To test hypotheses about the population mean, we use the sample mean as an estimate of the population mean, where we know how reliable an estimate we have by looking at the standard error (small standard error means little spread in our estimates, i.e. good estimate).

If our sample is big enough that we can assume the Normal approximation (i.e. sample size of at least 30), then we can find the chance of observing this data or more extreme data, if the null hypothesis is true. For example, to test whether the population mean has a particular value against the two-sided alternative (i.e. the population mean does not have that value), our test statistic, assuming a sample size of at least 30 is:

(Sample mean-hypothesized population mean)/standard error

which we call the z-statistic and compare against the standard Normal distribution (mean=0, standard deviation=1), i.e. we reject at significance level 5% if the value is less than -1.96 or greater than +1.96, we reject at level 1% if the value is less than -2.65 or greater than +2.65. In other words, for a two-tailed test, we reject if the confidence interval does NOT contain the hypothesized population mean.

The complication is that we need to know the standard error, which is equal to the population standard deviation divided by the square root of the sample size. This means we need to know the population standard deviation. In practice, we often do not know this, so we have to estimate this from the sample, using the sample standard deviation. In order for this estimate of the population standard deviation to be good, the sample size needs to be at least 30, otherwise we need to account for this using the Student's T-distribution instead of the Normal distribution (we'll look at how this is different later)

For example, if we toss a coin and the thing we are measuring is the proportion of heads, then:

The sample mean is the sample proportion of heads

The standard error is the population standard deviation divided by the square root of the sample size.

In the special case of a proportion, the population standard deviation has a simple formula:

Population SD$=\sqrt{(p \times (1-p))}$

So,

Standard Error$=\sqrt{(p \times (1-p)/n)}$

As $p \times (1-p)$ is always less or equal to 1/4, then the standard error will always be less than $1/\sqrt{(4 \times \text{sample size})}$.

In fact, as we do not know p, we often use this simple limit for the standard error when calculating CIs for sample proportions, except when the sample size is large, when we can substitute the sample proportion in the formula. For hypothesis testing, we usually will know p, so we can use the exact formula. For small samples or p very close to 0 or 1, we should use Binomial tables, otherwise, the Normal approximation is easier to use and good enough as an approximation.

The Normal approximation to the binomial gives an approximate 95% C.I. of

sample proportion +- $1 / \sqrt{(\text{sample size})}$

and a 99% C.I. of

sample proportion of +- $1.32/\sqrt{(\text{sample size})}$

Making mistakes/errors

When doing a hypothesis test, we can make 2 types of mistake (error). We can reject the null, when it is true (called Type I error). We know that the chance of this type of error depends on the significance level we use to reject the null hypothesis. If we reject at $p=0.01$, this means that we have a 1% chance of rejecting a null hypothesis that is really true.

The other type of error is to fail to reject the null hypothesis when it is false (called Type II error). This is harder to work out, because it depends on the alternative, which often is not specific.

Example: if we have an unfair coin, this means that the chance of heads, p is not 1/2, but does not tell us the value of p

How easy it is to detect that it is not fair depends on 2 things:

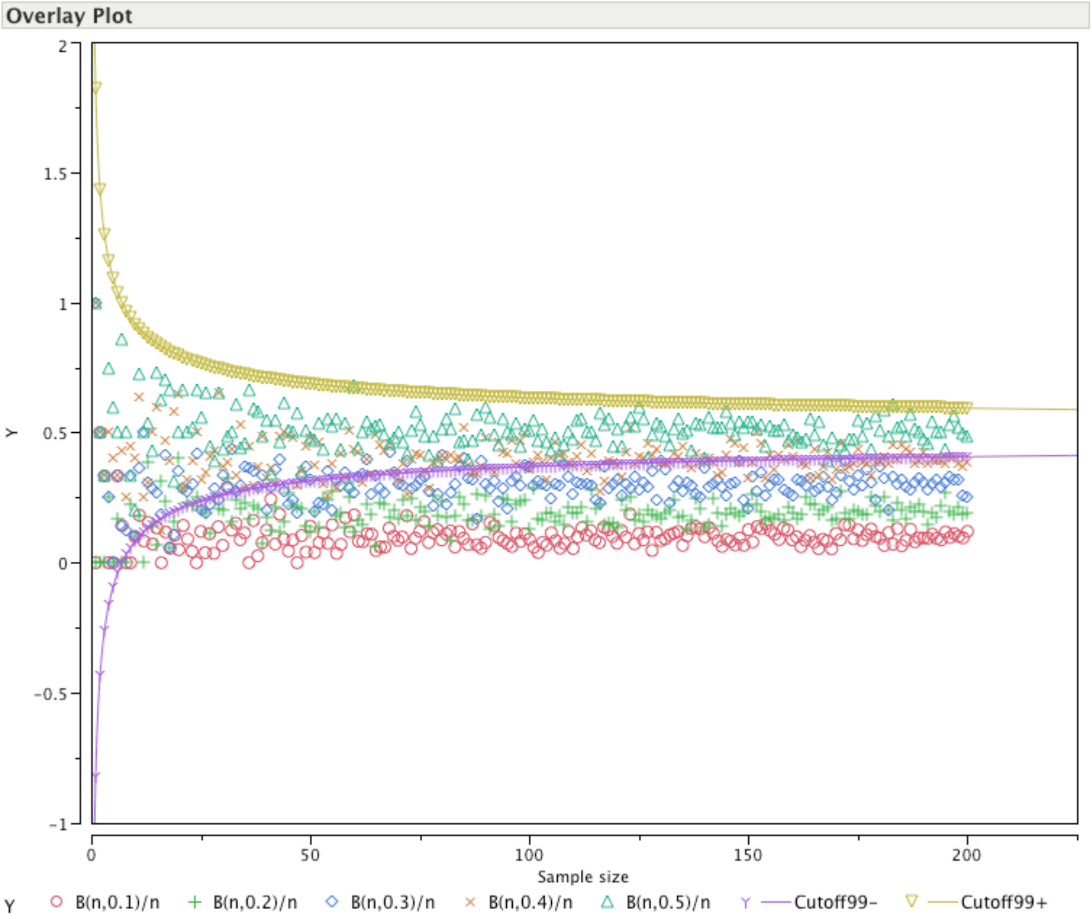
1) how unfair the coin is
2) how big a sample we take (how many tosses)

Consider the most extreme case of a 2-headed coin:

Sample proportion is 1, so we need only 8 tosses to reject fair coin hypothesis at p= 0.01

However, if the population proportion is closer to 1/2, then we need more tosses (bigger sample size).

Example: This is a simulation that illustrates the effect of changing how fair the coin is and how big the sample is. We will look at the results for sample size from 1 to 200 and with Pr(H)=0.1,0.2..0.5 and show the sample proportion for each sample and the 99% C.I. if the coin is fair, so we can see if we would reject the null hypothesis (fair) for each sample:



We can see that for Pr(H)=0.1 (red circles), a sample of 15 is plenty, while for Pr(H)=0.2 (green +), we need a sample of about 30, for Pr(H)=0.3 (blue diamond), we need a sample of about 100, while for Pr(H)=0.4 (brown x), even 200 is barely sufficient. Note that for the fair coin (green triangle), there are 2 or 3 cases where we would reject the null hypothesis, which is reasonable as we have a chance of 0.01 of making this mistake and we did 200 experiments, so we expect about 2 mistakes on average.

Power

The related question is how likely are we to reject the null when it is false?

This is called the power of our test and is 1 - the chance of a Type II error.

This depends on what the true population mean is, but doing this calculation for an assumed true population mean allows us to check whether our sample is likely to be useful or not. If our sample size will not allow us to reject the null, even when the true value is quite different, then our sample is of little use.

If you are seeking funding for a piece of research where the cost of doing the research is high, it is likely that the funding agency would expect you to show that your sample size is such that the chance of being able to reject the null is reasonable (usually require at least 80%) given the likely value of the population mean and variance (based on a pilot study or a literature review).

Remember our coin example – we can choose a sample size that is likely to reject if the coin has at least a given bias.

For power calculations, use tables or specialized software (webpages for simple cases can be found on the Internet): free software called GPower for example.

More precise statistical formulae

In general, we may not know the population variance, so we may not be able to do confidence intervals or hypothesis tests directly. However, for large sample sizes, we can just use the sample variance as a replacement, while for smaller sample sizes, we need to also replace the normal distribution (bell-shape) coefficients with ones using the Student's T distribution. That allows for the fact that we are using the sample variance instead of the population variance, so we cannot make such precise statements about the mean (i.e. the confidence interval will be wider).

For large samples (n>30), the 95% Confidence Interval for the population mean is:

sample mean +/- 1.96 x standard error of sample mean

where

standard error = population standard deviation/√(sample size)

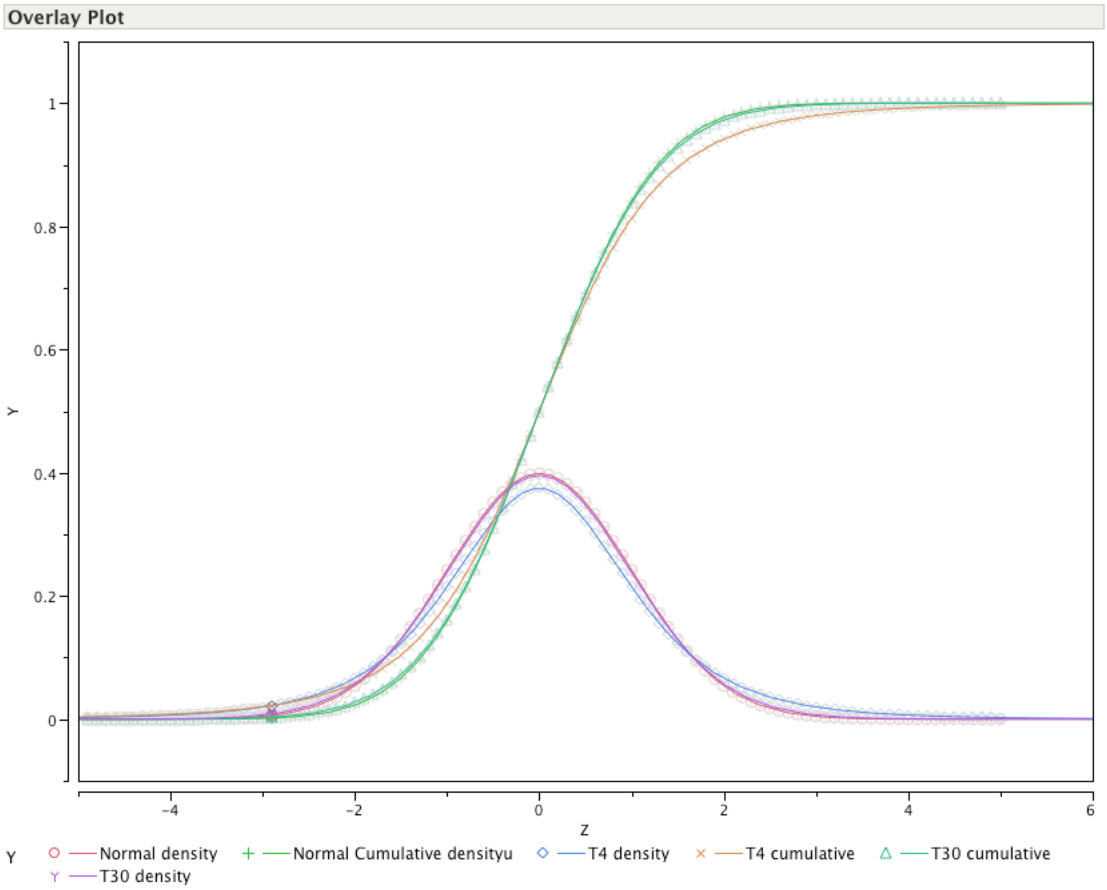
and we substitute the sample standard deviation for the population standard deviation, if necessary.

1.96 is called the Z-value. To be more precise it is $Z_{0.025}$ because it is the value such that there is a 2.5%=0.025 chance of a bigger value (the other 2.5% is the chance of a value below -Z). For a 99% C.I., we replace it by 2.65 which is $Z_{0.005}$.

Z is what we call a quantile for the standardized Normal distribution, where standardized means a Normal distribution with zero mean and unit variance.

When the population variance is not known and we need to estimate the population variance from the sample variance, we replace the Z-value by the t-value, which is slightly larger to take into account the fact that we have less information (by not knowing the population variance). The t-value also depends on the sample size (as it affects the accuracy of the variance estimate) through what is called the degrees of freedom (sample size -1). As long as the sample size is at least 30, the t-value is very similar to the Z-value.

We now show the distribution for a standard Normal distribution and T distributions with sample sizes of 5 and 31. The cumulative distribution means the probability of observing the variable to be less than or equal to that value.



The way the formula above works is that:

The sample mean approximately follows a Normal distribution with mean equal to the population mean and variance equal to the population variance divided by the sample size.

Equivalently,

(sample mean-population mean)/standard error

follows a standard Normal distribution with mean 0 and variance 1.

This is why 95% of the time, it will be between +/- 1.96

Or the sample mean will be between:
population mean +/-1.96 x standard error

Hence we get our confidence interval formula – the coverage of the confidence interval in this case is 95%.

For hypothesis testing, we are essentially doing the opposite. Instead of asking what are the likely values for the population mean, we ask whether the hypothesised value for the population mean is reasonable, or in other words, does it lie inside a confidence interval. If it does not lie in the 95% C.I., we say that we reject the null hypothesis at significance level of 5% (i.e. 100%-95%).

An equivalent way of expressing the hypothesis test is that we reject the null at 5% if:

(sample mean-hypothesised population mean)/standard error
$> +1.96$ or < -1.96.

In this version, we assume that values on either side of the hypothesised value are equally strong evidence against the null hypothesis. This is called a two-sided test.

One-tailed or two-tailed tests?

What we have described above is what is called the 2-tailed hypothesis test, i.e. our alternative allows that the coin may be biased towards either heads or tails.

If we were sure that only values on one side of the hypothesised value were evidence against the null hypothesis, our confidence interval and hypothesis test would be one-sided.

For example, if we believed that only population means greater than our hypothesised value were plausible, our 95% C.I. would cover all values up to the sample mean + 1.645 x standard error, where $1.645=Z_{0.05}$

Again, if this interval does not include the hypothesised value for the population mean, we would reject at 5% using a one-sided test

Note that it is easier to reject the null hypothesis with a one-sided test, in that the upper limit of the interval is lower. However, the decision to choose one tail should be a priori (beforehand), not after looking at the results!

When using statistical packages, they normally calculate what is called the observed significance level, which means the probability of rejecting the null incorrectly for this particular value of the test statistic.

This means we do not need to use statistical tables at all, but simply decide whether the observed significance level is sufficiently small that we reject the null hypothesis. This way, we know exactly how strong is the evidence against the null hypothesis for this data set, which is much more useful than just knowing whether we reject at 5% or 1%.

What if the population variance is unknown?

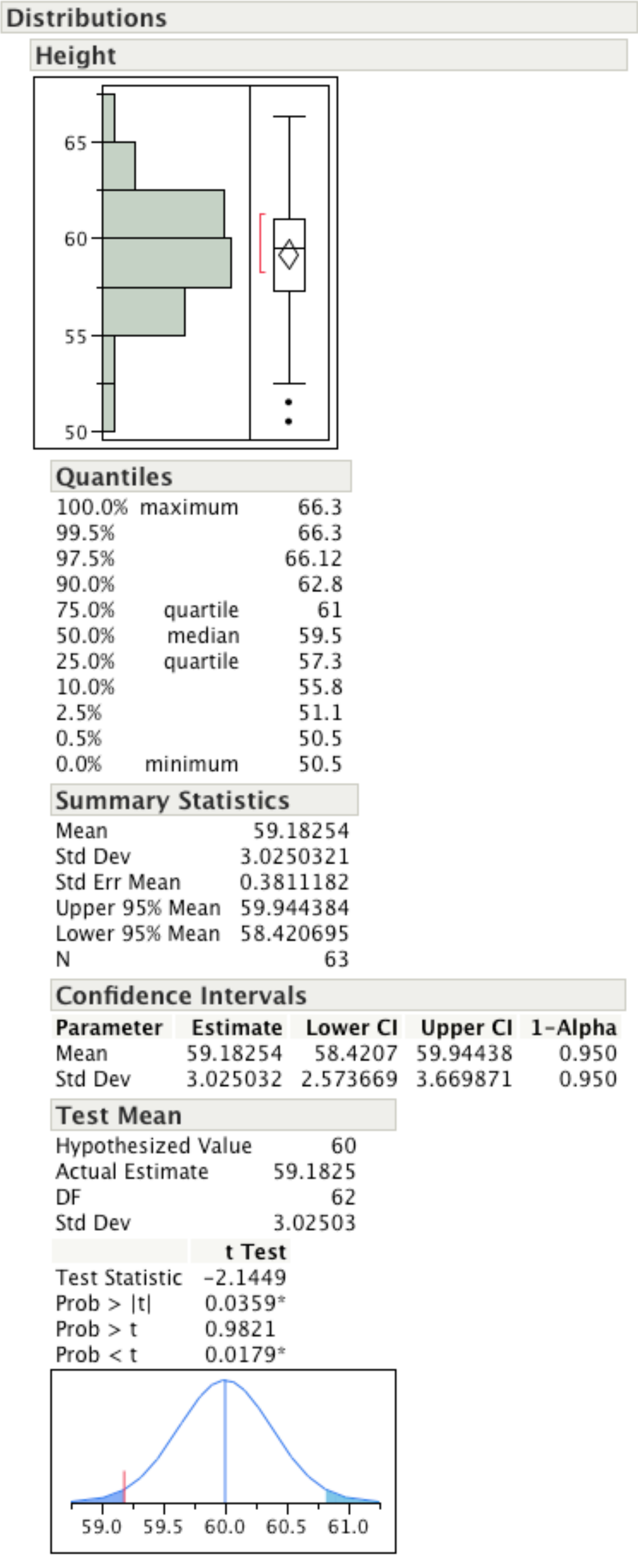
Note that if the standard error is estimated because the population variance is not known, then

(sample mean-population mean)/estimated standard error

follows a standard t distribution with (n-1) degrees of freedom where n is the sample size (instead of a standardised Normal distribution)

So, instead of +/-1.96 as interval, we will get a slightly larger number and hence a wider interval than when using the Z-value, unless the sample size is large.

Example: We have a sample of the heights of 63 12-year old children and show the 95% C.I. and the test for all 3 alternatives against a null hypothesis that the mean population height is 60



```
Quantiles
100.0%  maximum      66.3
99.5%                66.3
97.5%               66.12
90.0%                62.8
75.0%     quartile     61
50.0%     median     59.5
25.0%     quartile   57.3
10.0%                55.8
2.5%                 51.1
0.5%                 50.5
0.0%      minimum    50.5

Summary Statistics
Mean              59.18254
Std Dev          3.0250321
Std Err Mean     0.3811182
Upper 95% Mean   59.944384
Lower 95% Mean   58.420695
N                       63

Confidence Intervals
Parameter  Estimate  Lower CI  Upper CI  1-Alpha
Mean       59.18254   58.4207  59.94438    0.950
Std Dev    3.025032  2.573669  3.669871    0.950

Test Mean
Hypothesized Value        60
Actual Estimate      59.1825
DF                        62
Std Dev              3.02503
                  t Test
Test Statistic   -2.1449
Prob > |t|       0.0359*
Prob > t         0.9821
Prob < t         0.0179*
```

Note that we can apply all this theory to proportions in just the same way, using the sample proportion as our estimate for the population proportion and using either the exact formula for the population variance (=p x (1-p)) or the conservative limit (=1/4). For small samples, we can use the exact tables for the Binomial distribution to find the confidence interval and do the hypothesis test, instead of using the Normal approximation.

The danger of multiple tests

Note that if you are doing many significance tests, which are independent, the chance of making at least one false rejection of a null increases proportional to the number of tests you do.

Extension of hypothesis testing and confidence intervals to other situations

These ideas can be extended to other situations. The most common is testing for change in population mean. In other words, is the mean of population 1 the same as the mean of population 2 (i.e. the null hypothesis is that the difference between the means is zero)?

It is important to distinguish between two situations: namely whether the two samples contain the same individuals or not. If they do contain the same individuals, then we can analyse the change in measurements on the same individuals and get much more sensitive results. For example, to assess the effect of training, testing the same individuals before and after is a much more sensitive way of finding a change than testing a different set of individuals. This is because we are able to exclude the variability across individuals and focus on variability within individuals.

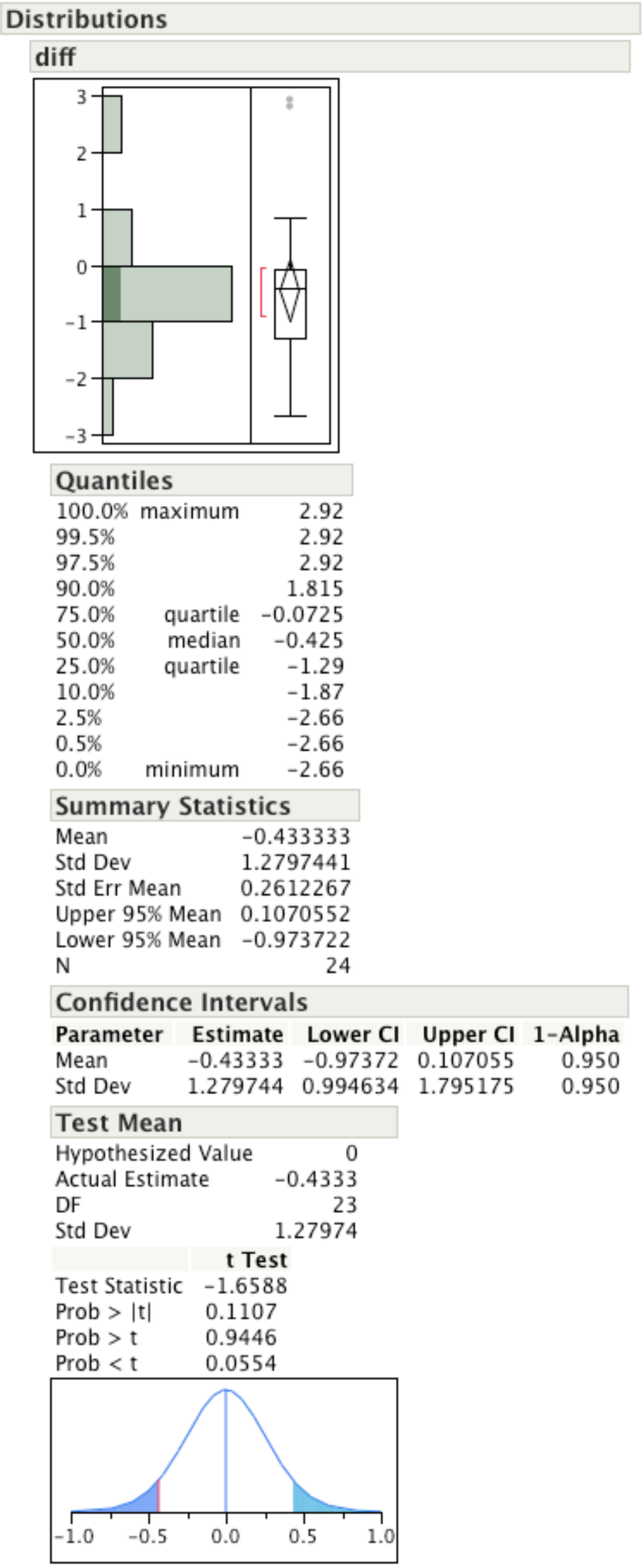
Paired T-test

Consider 64 individuals with average score of 50 and standard deviation of 8.

If we improve the score of half the individuals by 1 and half by 2, then we now have a new average score of 51.5. If we match individuals and look at the change, we have 64 differences with mean 1.5 and standard deviation 0.5, so the 99% confidence interval for the mean change is (1.33,1.67), so we are quite certain that there is a positive change (even if it is not big enough to be useful!).

In other words, by looking at the differences for each individual, we are just treating the change as our variable, and then applying the same statistical procedures to the changes to get confidence intervals and hypothesis tests for the difference in the population means. We call this a paired t-test (t-test, because we usually do not know the population standard deviation and estimate it from the sample).

Example: we have 24 measurements of the temperature inside and outside a box. We create a new variable, diff, which is the difference between inside and outside temperature. We show the 95% C.I. and the test of the (obvious) null hypothesis of no difference, i.e. the mean difference is zero.



```
Distributions
  diff

  Quantiles
  100.0%  maximum        2.92
  99.5%                  2.92
  97.5%                  2.92
  90.0%                 1.815
  75.0%     quartile  -0.0725
  50.0%     median     -0.425
  25.0%     quartile    -1.29
  10.0%                 -1.87
  2.5%                  -2.66
  0.5%                  -2.66
  0.0%      minimum     -2.66

  Summary Statistics
  Mean             -0.433333
  Std Dev          1.2797441
  Std Err Mean     0.2612267
  Upper 95% Mean   0.1070552
  Lower 95% Mean   -0.973722
  N                       24

  Confidence Intervals
  Parameter   Estimate   Lower CI   Upper CI  1-Alpha
  Mean        -0.43333   -0.97372   0.107055    0.950
  Std Dev     1.279744   0.994634   1.795175    0.950

  Test Mean
  Hypothesized Value        0
  Actual Estimate     -0.4333
  DF                       23
  Std Dev             1.27974
                   t Test
  Test Statistic  -1.6588
  Prob > |t|       0.1107
  Prob > t         0.9446
  Prob < t         0.0554
```

Two-sample T-test

If the populations are different, or we do not use the same sample, what then?

Not surprisingly, the difference in sample means is still the best estimator of the difference in population means, while the standard error of the difference is √(population variance 1/sample size 1+population variance 2/sample size2), so we can use similar ideas for producing confidence intervals and hypothesis tests for the difference in the population means. This is called the two-sample t-test (because the population standard deviations are not usually known, although we often assume that population variance 1 = population variance 2, so we can use a pooled (combined) estimate of the variability, based on both samples, which simplifies the statistical calculations).

In other words, for a large sample, the 95% confidence interval for the difference in means is:

(Sample mean1-sample mean2)+/-1.96 x standard error

Where the standard error is:

√(pop variance 1/sample size 1+pop variance 2/sample size2)

Note: the means subtract, but the variances add.

Hence, for our example above, if we do not match individuals (treat them as a new sample), then the mean difference is 1, but the standard error is

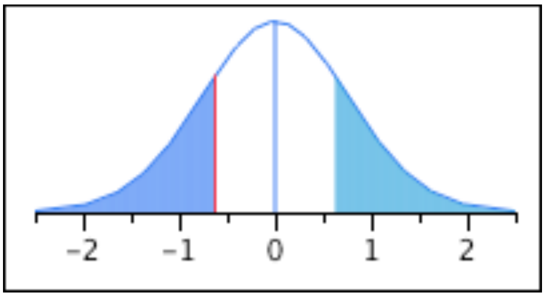
√(pop var1/sample size 1+pop var 2/sample size 2)
=√(64/64+64/64) =√2

so our 95% C.I. is 1.5 +/- 1.96 x √2 and 99% C.I. is now 1.5 +/- 2.65 x √2, which includes 0, so we fail to reject the null (i.e. no change).

Example: we look again at the heights of the 63 children, this time we examine whether the boys and girls have different heights. The results are shown twice for the difference in heights, with and without the assumption that the variance is the same in the two populations.

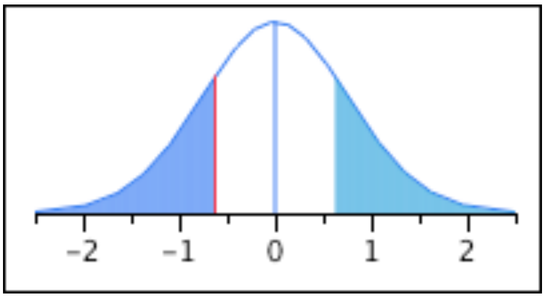
Height By Gender (Males-Females)
Assuming equal variances

```
Difference        -0.6252 t Ratio       -0.81702
Std Err Dif        0.7652 DF                  61
Upper CL Dif       0.9049 Prob > |t|      0.4171
Lower CL Dif      -2.1552 Prob > t        0.7915
Confidence           0.95 Prob < t        0.2085
```



Assuming unequal variances

```
Difference        -0.6252 t Ratio       -0.81815
Std Err Dif        0.7641 DF            60.71441
Upper CL Dif       0.9029 Prob > |t|      0.4165
Lower CL Dif      -2.1532 Prob > t        0.7918
Confidence           0.95 Prob < t        0.2082
```



Effect size

The (standardized) effect size is a standardized measure of the strength of a relationship. In this case, it is just the standardized mean difference between the two groups. This is important if we are trying to calculate the power for a two-sample case with different sample sizes for an assumed effect size.

The effect size for a two-sample t-test is:

(pop mean 1- pop mean 2)/standard deviation.

Note that this does not depend on the sample size and has no units.

Categorical data with more than 2 categories

We have already looked at the case of proportions when there are only 2 categories. In that case, the count follows a Binomial distribution, and for moderately large samples ($n>10$), the count and proportion approximately follow a Normal distribution. The obvious extension is to the situation with more than 2 categories. In this case, we usually apply a different concept known as Goodness of Fit.

The Goodness of Fit test can be applied to all situations where we have count data and the alternative hypothesis is the 2 tailed form (e.g. for 3 categories, the null hypothesis is that the population proportions are specified (for example, that they are equal) and the alternative is that at least one of the proportions is something different).
Example: if we have data for how many people in our sample live in HK Island, Kowloon or NT, we can test the hypothesis that the proportions are consistent with the proportions in the 2011 census data.

The usual test statistic is called the (Pearson's) Chi-squared (or X^2) Goodness of Fit statistic.

What is the Pearson's Chi-squared Goodness of Fit statistic?

$X^2 = \Sigma$ (Observed count for each category-Expected count for that category)^2/Expected count for that category

Where Observed means the actual count and Expected means the count we would expect if the null hypothesis is true.

You can see that this measure is a way of summarizing how close the observed data is to the data that is most likely if the null hypothesis is true, hence the name goodness of fit.

If k is the number of categories, then we compare X^2 against tables for the Chi-squared distribution with k-1 degrees of freedom (This is the distribution for the sum of k-1 independent squared standard Normal variables)

The calculation of X^2 and of the probability of observing this value of X^2 or a value more extreme (observed significance level) can be easily found using a computer package.

Strictly speaking, this assumes the Normal distribution, so the probability distribution for X^2 is only approximate, but this is a good approximation as long as the expected count is at least 5 for each category.

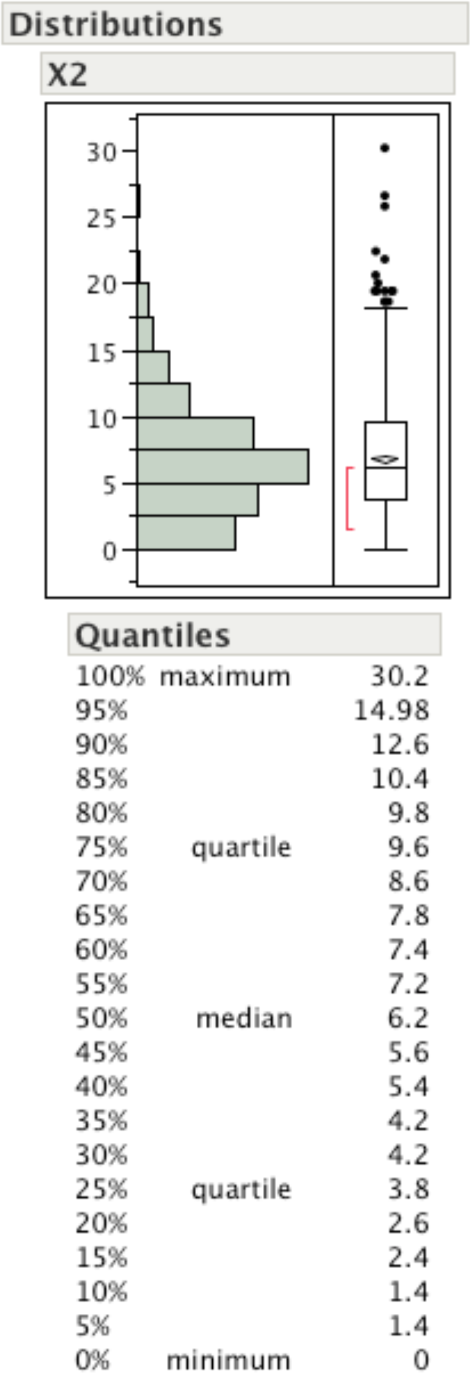
There is another statistic used in this situation, called the G^2 Likelihood ratio statistic, which is essentially just another approximation to a chi-squared statistic.

G^2=2 x Sum (Observed x log(expected/observed)).

Note: we are assuming that the data is nominal scale, not ordinal. If the data is ordinal scale, then there are more sensitive statistical methods that should be used.

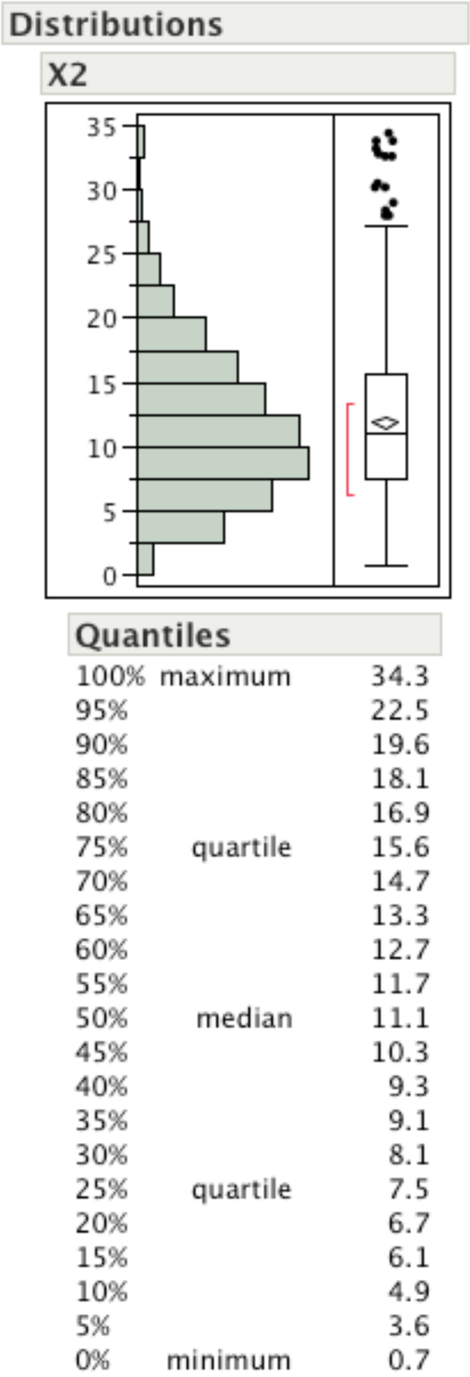
Example: We have simulated data for the mother tongue of students in an international school: English, Cantonese or Putonghua with Pr(E)=1/3, Pr(C)=1/2,Pr(P)=1/6. The first simulation of 1000 experiments uses a sample size of 30, the second uses a sample size of 60. We test a null hypothesis that the proportions are equal. The 95% cutoff for X^2 is 6.0, power for a sample of 30 is about 50%, sample of 60 is about 85%.

Sample of 30



```
Distributions
  X2

Quantiles
100% maximum    30.2
95%            14.98
90%             12.6
85%             10.4
80%              9.8
75%   quartile   9.6
70%              8.6
65%              7.8
60%              7.4
55%              7.2
50%   median     6.2
45%              5.6
40%              5.4
35%              4.2
30%              4.2
25%   quartile   3.8
20%              2.6
15%              2.4
10%              1.4
5%               1.4
0%    minimum      0
```

Sample of 60



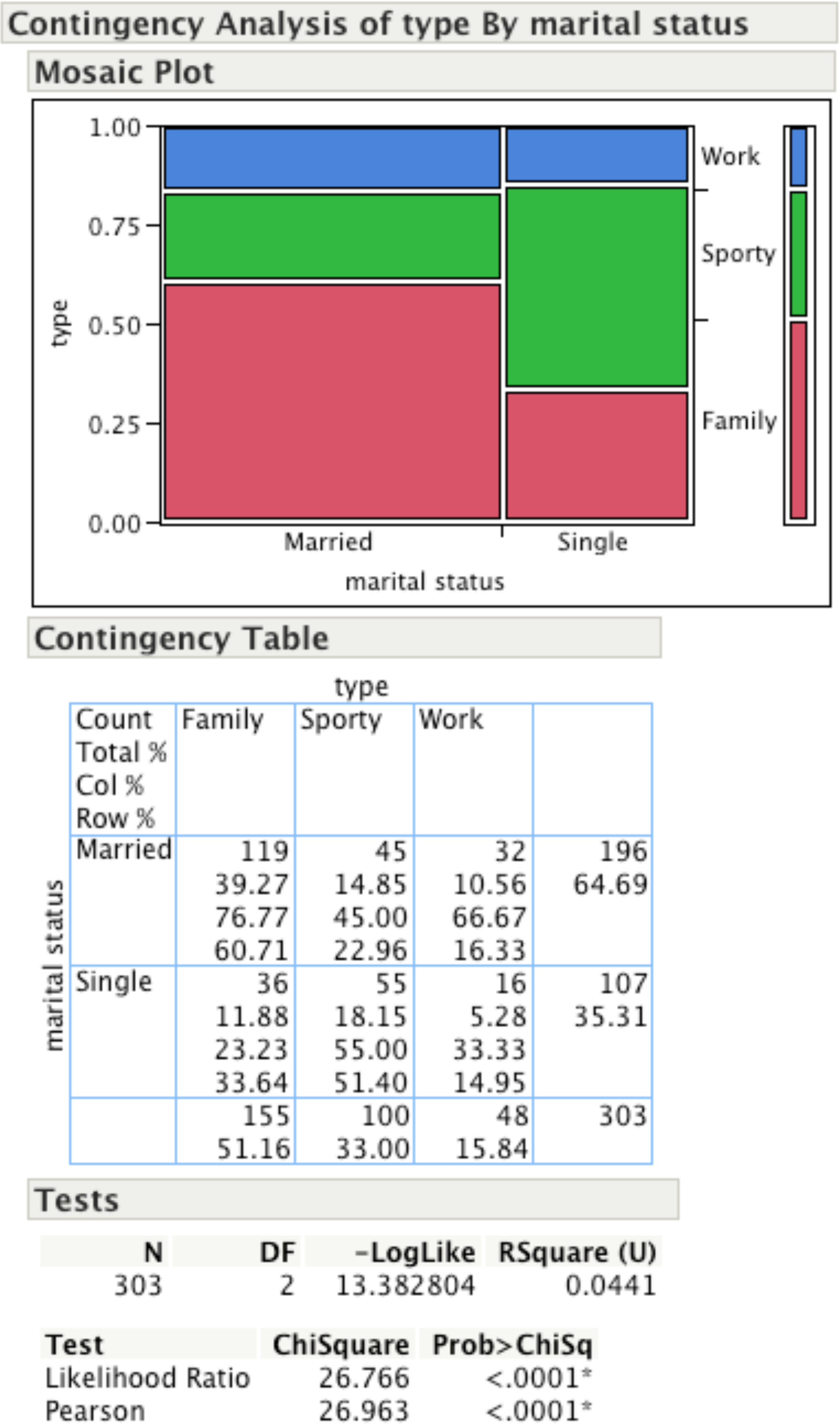
```
Distributions
  X2

Quantiles
100% maximum    34.3
95%             22.5
90%             19.6
85%             18.1
80%             16.9
75%   quartile  15.6
70%             14.7
65%             13.3
60%             12.7
55%             11.7
50%   median    11.1
45%             10.3
40%              9.3
35%              9.1
30%              8.1
25%   quartile   7.5
20%              6.7
15%              6.1
10%              4.9
5%               3.6
0%    minimum    0.7
```

X: Relationships between pairs of variables

If we do not know what type of relationship there is between a pair of variables, the best starting point is usually a graphical display or a table. For variables that are categorical, we usually use a table of counts (can use mosaic plot), otherwise we use a graphical display called a scatterplot.

Let's look at an example with two nominal scale variables and then another example with two interval scale variables.

Car poll: This shows that there is association between preferred car type and marital status and the mosaic plot shows the key difference is Singles prefer a Sporty car to a Family car.



Contingency Analysis of type By marital status

Mosaic Plot

Contingency Table

marital status	type Count Total % Col % Row %	Family	Sporty	Work	
	Married	119 39.27 76.77 60.71	45 14.85 45.00 22.96	32 10.56 66.67 16.33	196 64.69
	Single	36 11.88 23.23 33.64	55 18.15 55.00 51.40	16 5.28 33.33 14.95	107 35.31
		155 51.16	100 33.00	48 15.84	303

Tests

N	DF	-LogLike	RSquare (U)
303	2	13.382804	0.0441

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	26.766	<.0001*
Pearson	26.963	<.0001*

Testing for independence of categorical variables

A common situation with categorical data is that we have 2 nominal scale variables with a research hypothesis of association between the variables while the null hypothesis is independence (which means that probabilities multiply).

For example, we might classify people both by gender and whether they are under 30 or not. One question would be whether we can reject the hypothesis that the gender and age are independent in our population. The alternative hypothesis is that there is some (unstated) association between them.

It turns out that we can use the same concept (GoF) as we used for 1 variable with multiple categories above.

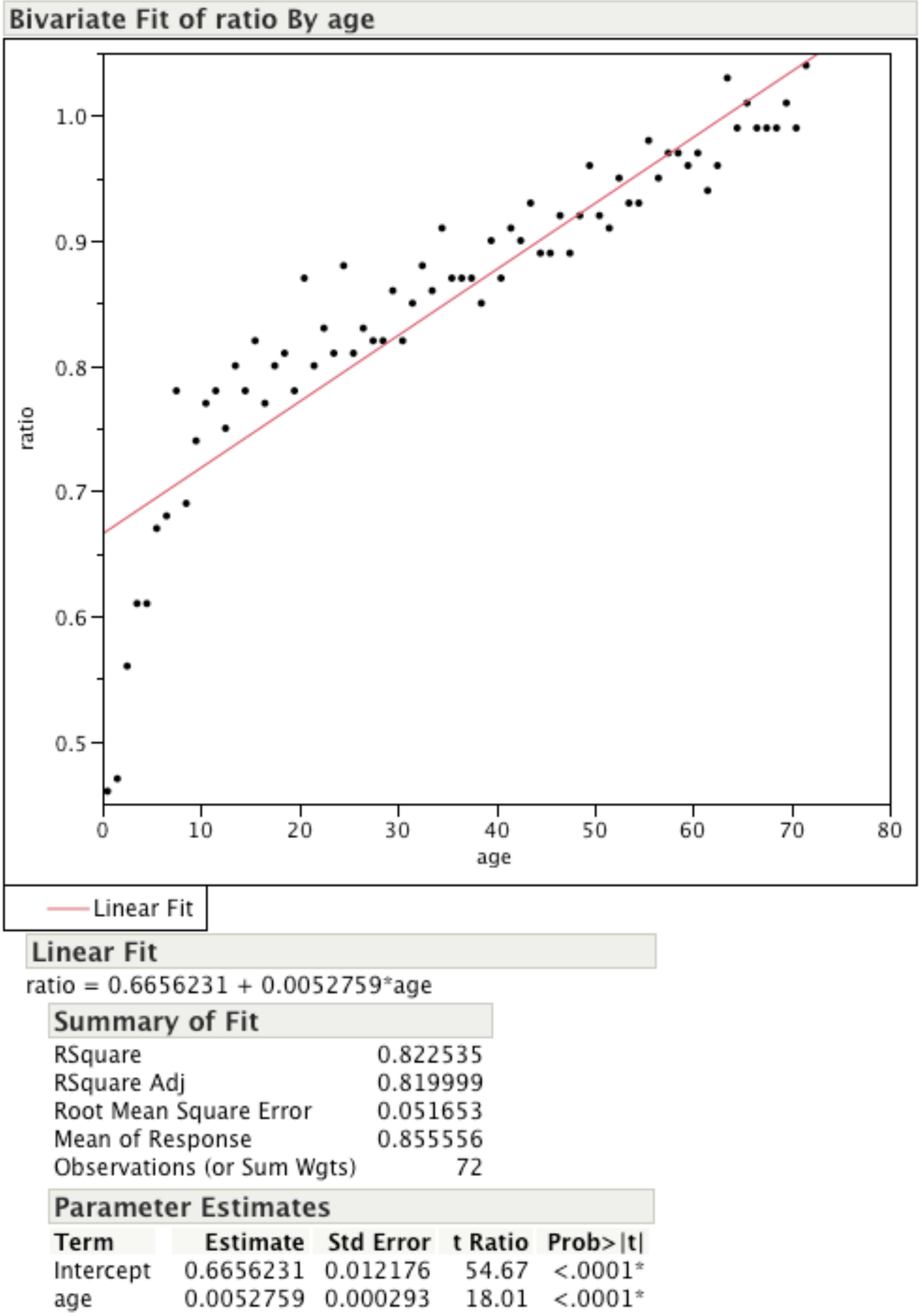
In the case of our example, the null hypothesis implies that the proportion of males who are under 30 should be the same as the proportion of females who are under 30 or equivalently, the proportion of under 30s who are male should be the same as the proportion of over 30s who are male. In other words the probability of being a male is independent of the probability of being under 30.

This means that the expected number of males under 30 is equal to the total sample size times the sample proportion of males overall times the sample proportion of under 30s overall. Similarly for the other three combinations of age and gender.

We can then use these expected and observed counts in the X^2 statistic and the degrees of freedom will be (k-1) x (m-1) where k is the number of categories for one variable and m is the number of categories for the other variable, so in the 2 x 2 case, the degrees of freedom are 1.

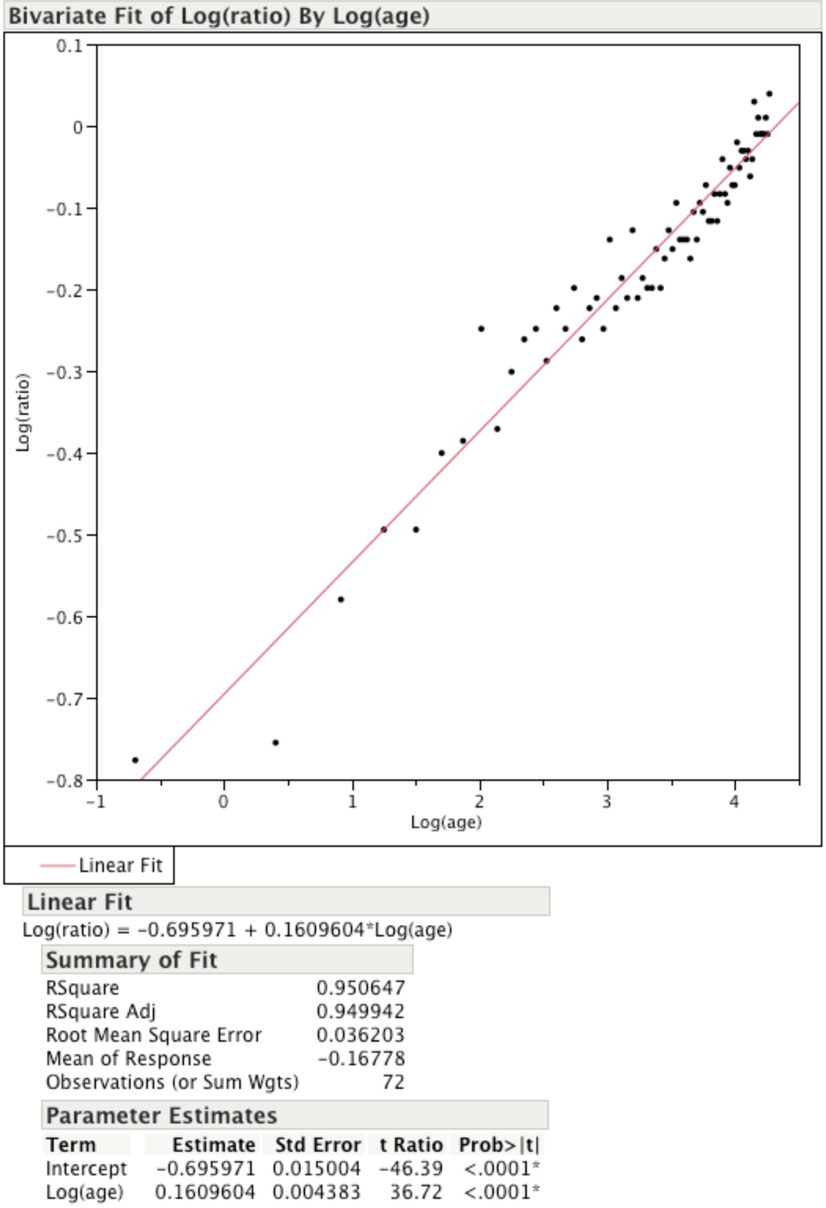
Need expected count of at least 5 in each cell.

Growth of babies: This plot shows a strong non-linear pattern to the relationship between growth ratio and age.



```
Bivariate Fit of ratio By age

Linear Fit
ratio = 0.6656231 + 0.0052759*age

Summary of Fit
RSquare                      0.822535
RSquare Adj                  0.819999
Root Mean Square Error       0.051653
Mean of Response             0.855556
Observations (or Sum Wgts)         72

Parameter Estimates
Term        Estimate   Std Error  t Ratio  Prob>|t|
Intercept  0.6656231   0.012176    54.67   <.0001*
age        0.0052759   0.000293    18.01   <.0001*
```

Once we log transform both growth ratio and age, we see an approximate linear relationship



Linear Fit

Log(ratio) = -0.695971 + 0.1609604*Log(age)

Summary of Fit

RSquare	0.950647
RSquare Adj	0.949942
Root Mean Square Error	0.036203
Mean of Response	-0.16778
Observations (or Sum Wgts)	72

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.695971	0.015004	-46.39	<.0001*
Log(age)	0.1609604	0.004383	36.72	<.0001*

Those examples show that the relationships can have many forms, but they often are approximately linear (possibly after simple transformation, e.g. taking logs), which allows us to build simple models.

Use of correlation

We can summarise the linear relationship using correlation (r). Correlation measures association in terms of how strong the linear relationship is, where a correlation of:

+1 means a perfect positive linear relationship
-1 means a perfect negative linear relationship
0 means no linear relationship, although there may still be a non-linear relationship.
Between 0 and +1 means a positive linear relationship that does not fit a line perfectly
Between –1 and 0 means a negative linear relationship that does not fit a line perfectly

Note that if r is not equal to +/-1, this can be because

a) relationship is not linear (curve?)
b) (random) errors in either x or y

r by definition is:

$$\text{Covariance}(X,Y)/\surd(\text{Var}(X)\text{x}\text{Var}(Y))$$

Where:

$$\text{Covariance}(X,Y)=E((X-\text{Mean}(X))(Y-\text{Mean}(Y))$$

$$\text{Var}(X)=E((X-\text{Mean}(X))^{2})$$

i.e. correlation is the joint variability of X and Y scaled by the variability of each of X and Y.

Note: no need to memorize these formulae

Estimation of correlation, r, based on a sample is:

$r=\sum((Y\text{-Mean}(Y))(X\text{-Mean}(X))/\sqrt{}(\sum (X\text{-Mean}(X))^2x\sum(Y\text{-Mean}(Y))^2)$

It is straightforward then to show that if Y=X, r=1 and if Y=-X, then r=-1

If Y=A+BX, then r=+/-1, depending on the sign of B

Obvious hypothesis test is if the population correlation is zero, but this test requires us to assume that both variables (X and Y) follow a Normal distribution.

The test relies on the fact that if X and Y follow a Normal distribution with zero correlation, then:

$r/\sqrt{}((1-r^2)/(n-2))$ follows a Student's t distribution with (n-2) degrees of freedom. If the sample size is at least 30, the t distribution is very close to a Standard Normal distribution (otherwise the tails of the distribution are "thicker" to account for our estimating the population variance using the sample variance)

If X and Y both follow a Normal distribution, then it turns out that independence is identical to zero correlation, but this is not necessarily true in other situations (zero correlation just means no linear relationship, remember, while independence means no relationship at all, linear or non-linear).

We sometimes use r^2 as a way of summarizing the proportion of variability of one variable "explained" by the other variable.

However, note that r^2 does not tell us about the direction of causation. It is possible that

X => Y

Y => X

Or that W causes both X and Y (W is a common cause).

There is a whole field of statistical methodology that looks at correlations for large sets of variables (factor analysis or principal component analysis), but the mainstream approach is to build models for linear relationships.

Simple (bivariate) linear model

We write the simple linear model as:

$Y=A + B X + \varepsilon$

where ε is the random error, A is the intercept, B is the slope, X is the independent variable, Y is the dependent variable and we assume that random errors from different observations are independent and have constant variance.

This is like elementary algebra for a line, except for the random error term.

Key question:

How do we estimate A and B? Clearly we want the "best fitting" model.

There is a very general mechanism for fitting statistical models with constant error variability. This method is called "least squares". It minimizes the squared error, or in other words, it finds estimates for A and B (we will call them a and b) that minimize the squared error for Y:

$\sum(\text{observed y-fitted y})^2$

This can also be written as:

$\sum r_i^2$

where r_i is called the residual for the ith data point, being the difference between the observed and fitted y for that point.

Hence the name least squares. It is possible to prove that this is the best way to fit a linear model if the error variance is constant.

Also, the error variance can be estimated from the sample as:

$s^2 = \sum r_i^2/(n-2)$

where n is the sample size.

(we divide by n-2 because we "use up" 2 observations to account for the 2 parameters we estimate – remember that we can always fit a line perfectly to any 2 points)

The formulae for a and b are:

b=∑(Y-Mean(Y))(X-Mean(X))/ ∑(X-Mean(X))(X-Mean(X))
=Cov(X,Y)/Var(X)

a=Mean(Y)-b x Mean(X)

(because Mean (Y)= a + b x Mean(X))

In practice, we do not do the calculations by hand – even simple calculators can do the calculations for us, or a spreadsheet.

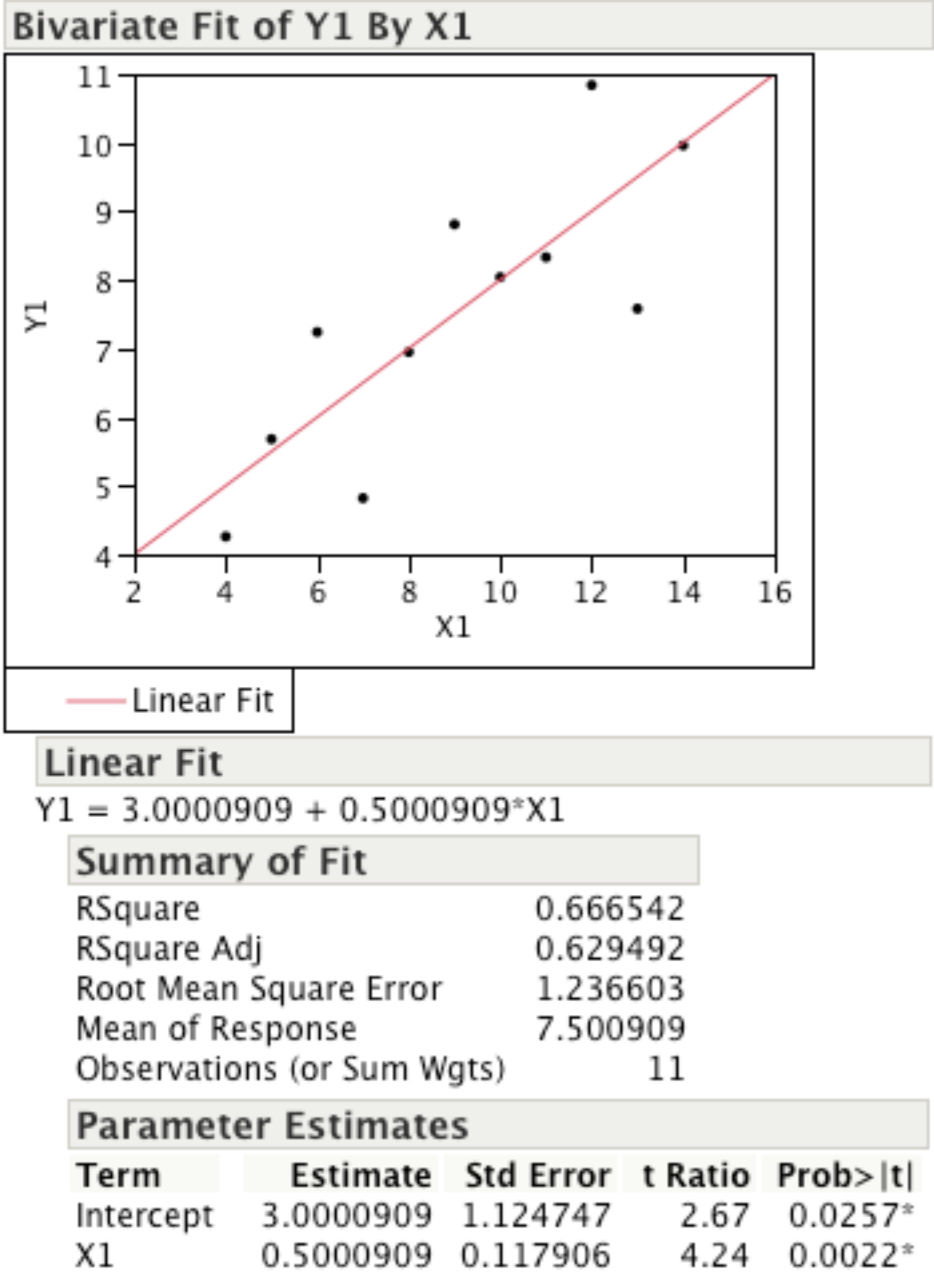
For those of you with good mathematical skills, it is easy to show that these estimates are optimal by using calculus to find the values of a and b that minimise s^2

However, an important word of warning:

Numbers (such as a,b and s^2) alone (without graphics) may not be a good summary of what is going on.

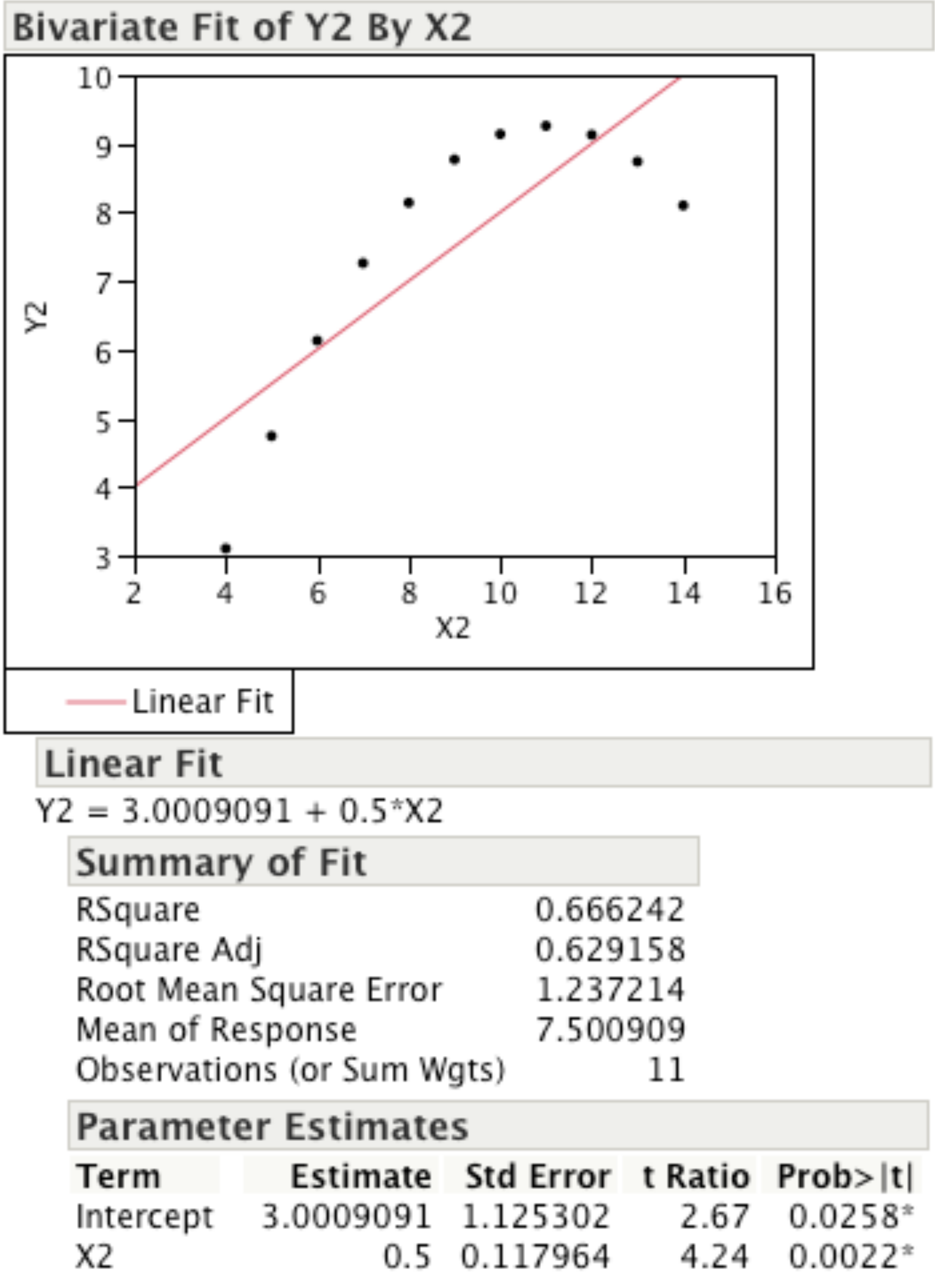
Examples: three samples, each with a sample size of 11.

The first one shows a moderate linear relationship and no unusual patterns.



```
Bivariate Fit of Y1 By X1

Linear Fit
Y1 = 3.0000909 + 0.5000909*X1

Summary of Fit
RSquare                        0.666542
RSquare Adj                    0.629492
Root Mean Square Error         1.236603
Mean of Response               7.500909
Observations (or Sum Wgts)           11

Parameter Estimates
Term        Estimate   Std Error  t Ratio  Prob>|t|
Intercept   3.0000909  1.124747      2.67   0.0257*
X1          0.5000909  0.117906      4.24   0.0022*
```

The second example shows a strong pattern:



Bivariate Fit of Y2 By X2

Linear Fit

Y2 = 3.0009091 + 0.5*X2

Summary of Fit

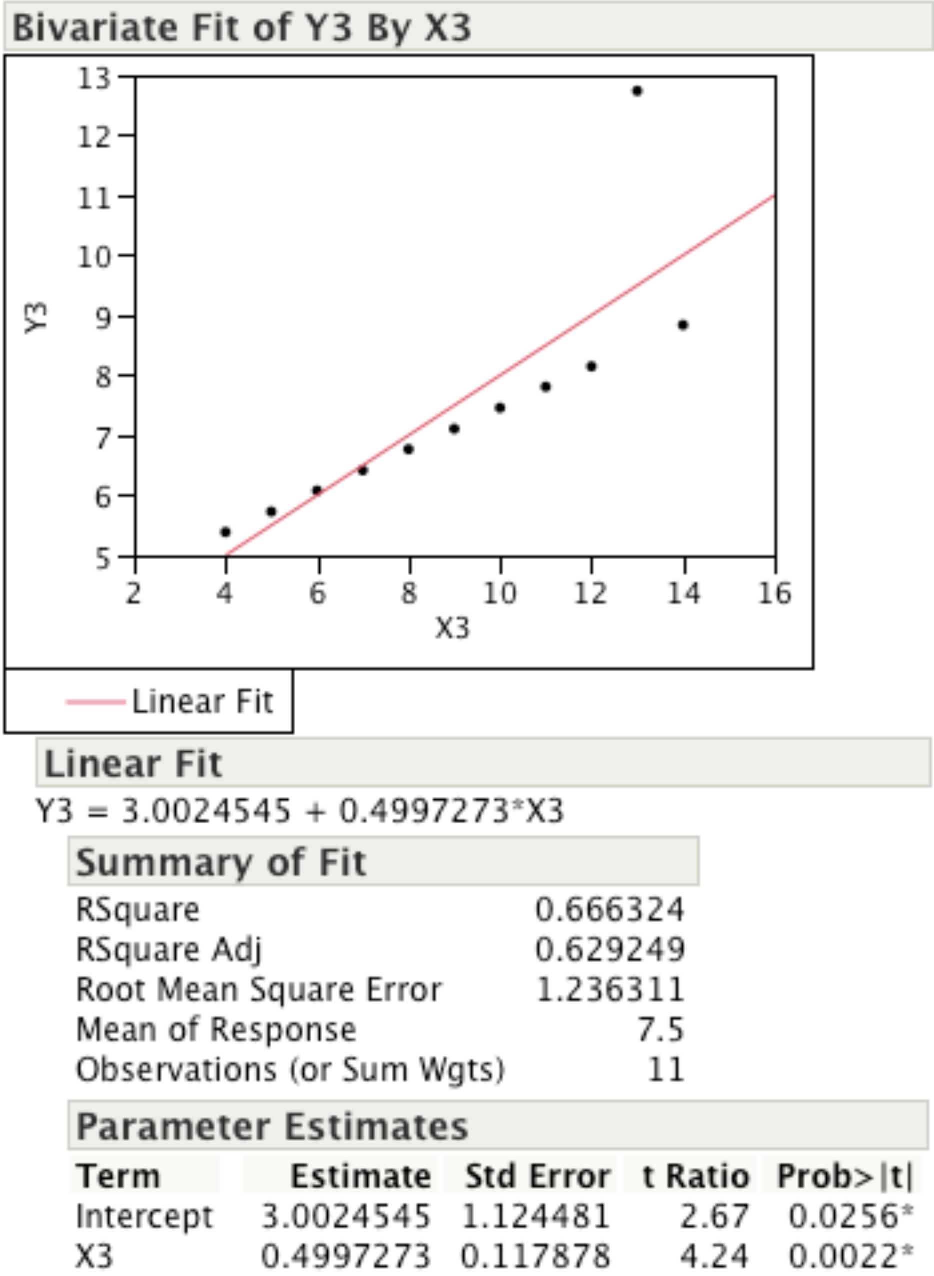
RSquare	0.666242
RSquare Adj	0.629158
Root Mean Square Error	1.237214
Mean of Response	7.500909
Observations (or Sum Wgts)	11

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.0009091	1.125302	2.67	0.0258*
X2	0.5	0.117964	4.24	0.0022*

Clearly, there is a quadratic relationship here

The third example shows a different pattern:



```
Bivariate Fit of Y3 By X3

Linear Fit
Y3 = 3.0024545 + 0.4997273*X3

Summary of Fit
RSquare                          0.666324
RSquare Adj                      0.629249
Root Mean Square Error           1.236311
Mean of Response                      7.5
Observations (or Sum Wgts)             11

Parameter Estimates
Term         Estimate   Std Error   t Ratio   Prob>|t|
Intercept   3.0024545    1.124481      2.67    0.0256*
X3          0.4997273    0.117878      4.24    0.0022*
```

Clearly, there is one data point inconsistent with our model.

Also, note that (unlike correlation) the linear model is not symmetrical – if you reverse the roles of X and Y, you get a different model. This is because this model assumes that all the (measurement) error is in Y, not X.

If we can also assume that the errors follow a Normal (bell-shaped distribution) (in addition to our assumption of independent errors with constant (unknown) variability), then we can develop statistical inference for the slope. The key hypothesis to test is whether B=0 as this simplifies our model to mean that X and Y are independent. This turns out to be the same test as testing r=0, even though here we are only assuming that the errors for Y given X are Normal, not that X and Y are both Normal.

Essentially, we are asking what values of b we might expect if there is really no linear relationship between X and Y, and the errors follow a Normal distribution with constant variance, and rejecting B=0 if the value we observe for b is too far away from 0 to be likely under the null.

The test for B=0 is based on the fact that,

$(b-B)/s_b$

follows a Student's t distribution with n-2 degrees of freedom, where

$s_b=s/\sqrt{(\sum(X-Mean(X))^2)}$ is the standard error of b.

Note that s_b decreases if the variance of X increases.

This, of course, also provides the basis for finding a 100(1-α)% confidence interval for B of

$b+/-s_b$ x $t_{\alpha/2}(n-2)$, where $t_{\alpha/2}(n-2)$ is the cutoff value of a t distribution with (n-2) degrees of freedom

If the sample size is above 30, then the t distribution is very close to a Standard Normal distribution.

(We could also test separately for A=0 in a similar way, which would mean we are testing if Y is proportional to X, which is rarely meaningful).

Residuals

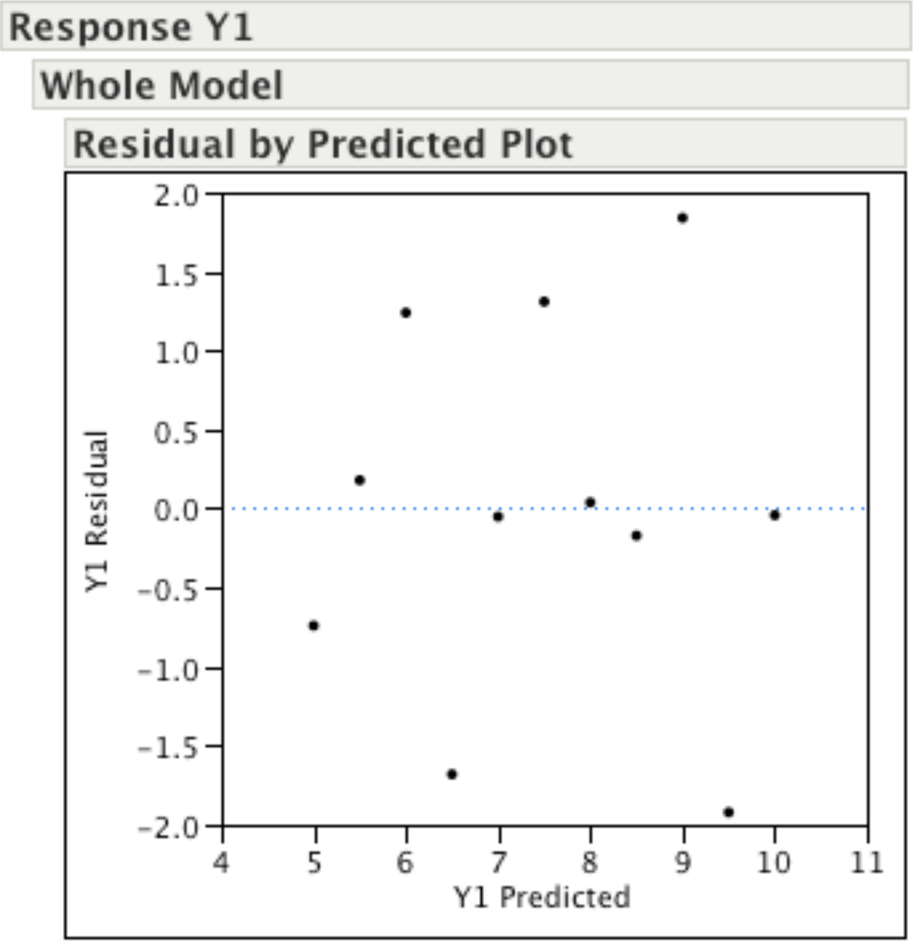
As with any tool, diagnostics are important. The simplest diagnostics for regression are the residuals.

2 key ideas:

Look for particular data points that look to fit badly (possible error?)

Look for patterns that suggest model errors.

Examples: same as before, this one shows no pattern

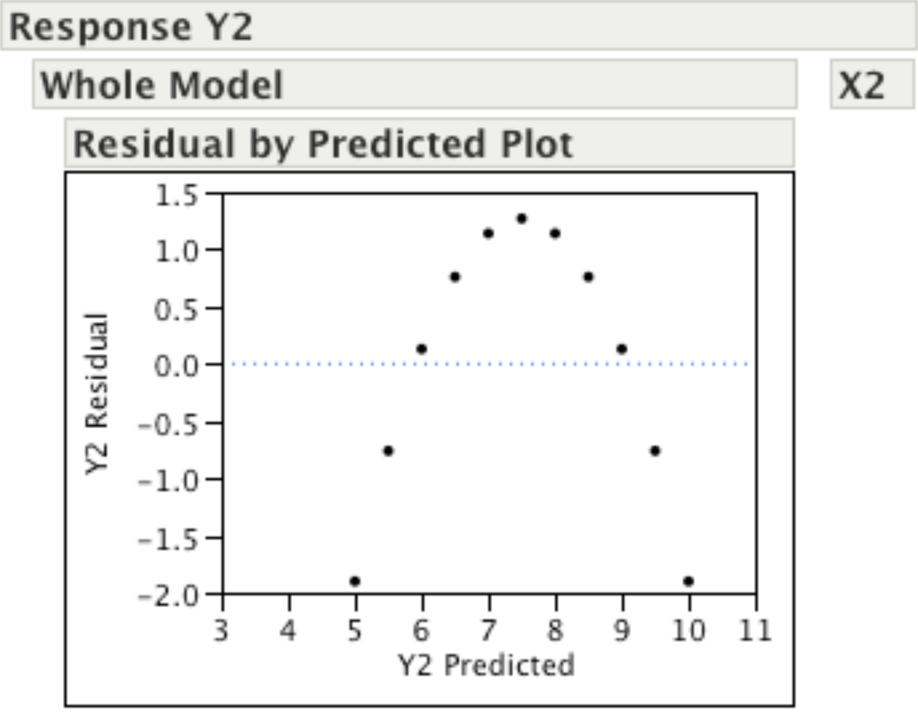


Response Y1

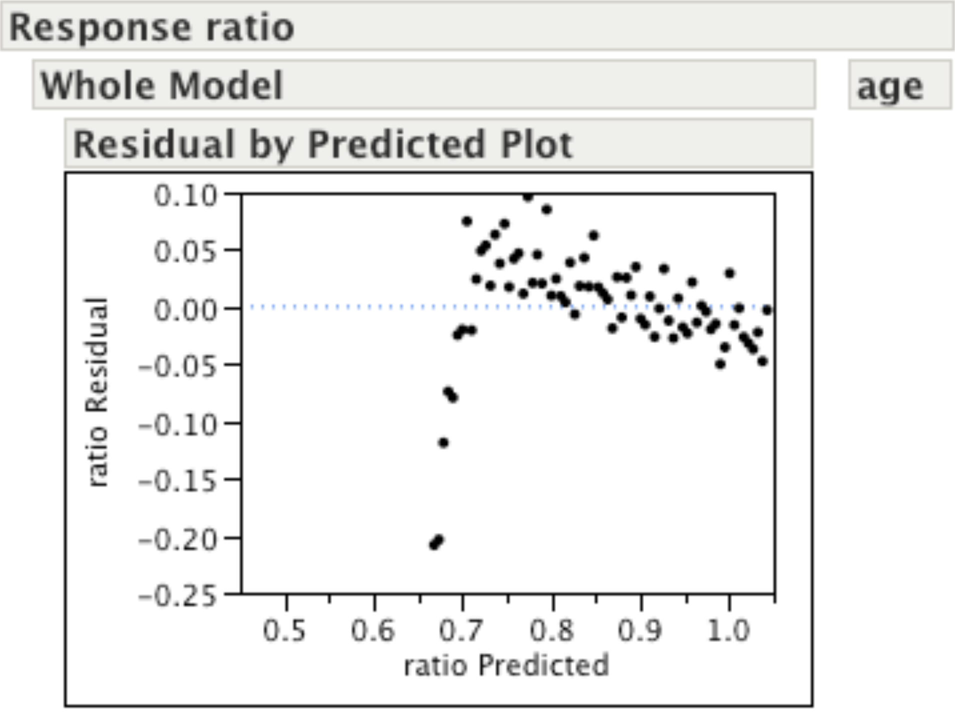
Whole Model

Residual by Predicted Plot

Here we can see the quadratic pattern so clearly!



Here the growth example shows the strong non-linear relationship



Meaning for r^2

r^2 summarises how good the linear fit is on a scale from 0 to 1

r^2 can be written as:

$1-\sum r_i^2/\sum(y\text{-Mean}(y))^2$

This shows clearly that a perfect fit will give $r^2=1$ and if the fitted line is no better than a straight horizontal line through the Mean of y (i.e. yfit=Mean(y)), then $r^2=0$

r^2 is often interpreted loosely as the proportion of variability “explained” by x. This is loose in that this assumes that x causes y, which may not be true.

Prediction

How can we predict y for a value of x?

We can just substitute into our equation, using our estimates, a and b:

Y predicted=a + b × X predictor

We can also estimate the prediction variability, given that the distribution of Y predicted, given an X predictor is:

(Y predicted-true Y)/s_{pred} follows a t distribution with n-2 degrees of freedom, where

$s_{pred}=s\ \sqrt{}(1/n+(\text{Xpred-Mean}(X))^2)/\ \sum(X\text{-Mean}(X))^2)$

This can be used for hypothesis testing or constructing confidence intervals for the mean of Y given X.

Note that s_{pred} is smallest for Xpred=Mean(X) and increases as Xpred moves away from Mean(X)

Similarly for predicting y for a new individual

$s_{\text{predi}} = s\ \sqrt{(1+1/n+(\text{Xpred-Mean(X)})^2)/\sum(\text{X-Mean(X)})^2)}$

(look at example of confidence intervals in JMP – show mean confidence interval versus individual confidence interval which is the prediction for a new individual measured)

However, note that this assumes that our linear model is correct for that value of x. This may be safe if x is within the range of data we observed, but what if x is much greater (or smaller) than our dataset? In that case, we have no real support from the data for believing that we can extrapolate like this.

It should be obvious that a wider range of x values will produce more precise results both for estimating B and for prediction.

Statistical Advice Centre for Students (STACS)

All HKU research students are entitled to a maximum of 4 hours free statistical consulting (over their time as a graduate student), provided by me.

Details can be found on the Graduate School website,

http://www.hku.hk/gradsch/web/student/support/stat.htm

where you can find the link to download the form for your supervisor to sign and return to the Graduate School for checking. The form is called a support services form.

Graduate School
Room P403, Graduate House,
The University of Hong Kong,
Pokfulam Road, Hong Kong
Tel: (852) 2857 3470
Fax: (852) 2857 3543
Email: gradsch@hku.hk
Homepage: www.hku.hk/gradsch

ISBN-10: 988-12813-0-X
9 789881 281302