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: 2021-04-21

This coursebook 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. After reading the book, students should understand the basics of quantitative research and be able to critically review simple statistical analysis. The book is intended to be self-contained but does not include mathematical proofs.

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:

1: Research Methods

Learning objectives: understand theory, model, Occam's law and types of random variable

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.

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, because people can check the findings for mistakes.

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?

Research plan

After choosing a research topic, 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
Operationalization: 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) <http://www.hku.hk/rss/HREC.htm>
Bibliography: references to prior knowledge (Note: EndNote is free in HKU!)

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

That sounds difficult and boring!

Can we really evaluate with no passion, no politics, no religion, and no ideology?

Note: the key is objectivity in evaluation, i.e. 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 repeated telephone survey for the government, asking people living in Amoy Gardens about their personal hygiene behaviour – this was interventional research, not observational research, as claimed.

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

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 physicists 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 by 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 theory states that increased education leads to increased personal income and 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 some of them 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. Note that a research question may be as specific as "Do young people protest more than older people?", which links to the hypothesis: "Young people protest more than older people" or may be a broader question, such as "Why do HK people protest?" which does not link clearly to any hypothesis.

The **research (or alternative) hypothesis** is a positive statement about what the researcher expects to find, like "Young people protest more than older people"

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, like "There is no difference in how much young and older people protest".

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 principle (or law or razor)

If we have two explanations of the world, which are equally good in predicting outcomes, 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 principle 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 (i.e. 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 in terms of the number of parameters in our statistical model or the sample size minus the number of parameters in our statistical model, which we call the degrees of freedom (we will revisit this), where simplicity means the number of parameters is small or the degrees of freedom are large.

What is proof?

Proof means that you know the truth for certain.

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 and hence cannot explain anything that does vary over the study.

What is a random variable?

A variable is 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, i.e. they are included in the research hypothesis.

What is an independent variable?

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

What is a dependent variable?

A dependent variable is an explanatory variable that is the outcome variable, presumed to be affected by the independent variable(s), if there is an independent variable.
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 that are related to independent variables

Controlled: extraneous variables that are either manipulated or included in a statistical model) so as to exclude their effect on the relationship between independent and dependent variables (for confounding variables) or to reduce the variability of the dependent variable.

Uncontrolled: extraneous 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 uncontrolled variables 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.

Mechanisms 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 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 should be similar on all uncontrolled variables and on all variables assumed irrelevant.

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

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

Measurement error: 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 eliminate or 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).

Propensity scoring: when the allocation is non-random, we can look at possible bias by examining the conditional probability that a subject will be treated given the observed explanatory variables (the propensity score). If we stratify by propensity score, we obtain estimated treatment effects that are not biased by the measured explanatory variables - see Rosenbaum and Rubin's 1983 Biometrika paper: "The central role of the propensity score in observational studies for causal effects".

Example: The classic 1950 paper by Doll and Bradford Hill in the British Medical Journal: "Smoking and Carcinoma of the Lung" examined frequency tables of lung cancer or control (other hospital patients matched by gender, age and hospital) by gender and smoking or not. The research hypothesis is that smoking is associated with lung cancer while the null hypothesis is that there is no association. Lung cancer or not is the dependent variable, smoking or not is the independent variable, while gender, age and hospital are control variables. Whether a random variable is an independent variable or control (cannot be both) may not be obvious without checking the research hypothesis (independent variable is included, control is not).

Mendelian randomization: Mendel discovered that when he crossed purebred white flower and purple flower pea plants, the result was not a blend. Rather than being a mix of the two, the offspring (known as the F_1 generation) was purple-flowered. When Mendel self-fertilized the F_1 generation pea plants, he obtained a purple flower to white flower ratio in the F_2 generation of 3 to 1, suggesting that separate genes for separate traits are passed independently of one another from parents to offspring. We can use the random assignment of an individual's genotype from parental genotypes that occurs before conception to make causal inferences (assuming that the genotype is associated with the exposure of interest and independence of the genotype from confounding factors and from the outcome given the exposure and confounding factors). The strength of this is that it does not rely on allocation.

2: Association and causation

Learning objectives: understand association & causation

What is association?

Association is observed linkage, i.e. two outcomes X and Y tend to both occur (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?

X is a cause of Y if when X occurs, Y must occur. This is called X being a sufficient condition for Y (e.g. breaking your spinal cord is sufficient to cause paralysis). If Y only occurs if X occurs, then X is a necessary condition for Y (e.g. drinking alcohol is currently necessary to know if you are an alcoholic). 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!

Note: if X occurs (or not) after Y, X cannot be a cause of Y (unless you believe in time travel!), so let us consider situations where X occurs (or not) before Y occurs (or not).

If X and Y sometimes both occur and sometimes only one of them occurs, we can 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) whether Y occurs, given X occurs (or not).

Examples: Predict which

1. 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
2. Juvenile offender will reoffend depends on his background, family support and offending history
3. Patients will respond positively to a specific treatment depends on the medical history of the patient, genetic profile.

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 was 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!

Example: having breakfast and school performance – a newspaper reported that the researcher claimed that eating breakfast made almost one year's difference in school performance, with no mention of other potential underlying causes.

What is Granger causality?

Granger defined the causality relationship as meaning that the cause happens prior to its effect and the cause has *unique* information about the future values of its effect. We say that a variable X that evolves over time *Granger-causes* another evolving variable Y if predictions of the value of Y based on its own past values *and* on the past values of X are better than predictions of Y based only on its own past values. For example, if we want to predict whether it rains tomorrow (prediction means using any form of statistical model to make an estimate forward in time), we could check whether wind direction Granger causes rain. This would mean that knowing the past wind directions as well as past rainfall helps us make better predictions of rainfall, than just knowing the past rainfall. For more discussion of prediction, please see the chapter on time series, later in the coursebook.

Magnitude and consistency of association

If the magnitude of association is high, then this implies some form of causation, but it does not address the direction of causation and there may be other causal factors. In general, though, stronger associations make causal relationships more plausible. If the consistency is high, this means that the association appears under a variety of different conditions, which also 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 decreased 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 later in this book.

Further reading: http://www.stat.columbia.edu/~gelman/research/published/causalreview4.pdf
This review by Andrew Gelman discusses 3 key books by Judea Pearl, Morgan & Winship, Steven Sloman on causal inference.

3: Research design

Learning objectives: understand representativeness, probability sampling & non-sampling error

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 usually the sampling elements, often people or households (need to be clear which), but can be rats, words, and songs – any countable objects (although sometimes we can analyse at multiple levels, e.g. students and schools).

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) or collect data on the whole population.

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. Below we discuss different mechanisms of probability sample.

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 ($10 \times 9 \div 2$) 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 of all students. 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 additional cost of sampling an extra student 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 (the effective sample size is smaller than a simple random sample).

What is a stratified sample?

Stratified sampling means sampling within subgroups of known size that are relatively homogenous, to yield more precise results than a simple random sample.
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 a systematic sample?

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 a network sample?

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. It is important as it yields representative samples of difficult to sample populations, if the assumptions are valid.

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 an initial set of drug addicts for introductions to their drug-using friends and then asking a sample of those friends, etc.

What is a distance sample?

Distance sampling involves randomly placing points or lines (transects) on a map and then measuring distances to objects. There are two main approaches – point transects and line transects. In both cases, we are trying to estimate the density of objects by counting how many we can see (from a point or line). The complication is that we need to account for the chance of detection being different at different distances. Examples would be estimating the number of people marching along Hennessy Road (usually count the number going under a bridge in random sample of time periods) or the number of people in Victoria Park (usually look at density in random sample of small areas).

Note: In the rest of this course, we will assume simple random sampling, for simplicity.

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 <u>essential</u> that the research

design take 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 non-response bias by seeing how responses differ by number of contact attempts. 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.

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.

Telephone versus face-to-face interviews:

Most interviews now use 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 high (70%+) 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). There are data archives in UK,

US, Norway etc. that mainly cover socio/economic/political data, but also geological, oceanographic etc. Critical value when data is expensive or impossible to collect again.

Choosing sample size:

When choosing the sample size for a study, we need to consider sampling error, non-sampling error and power (see later for an explanation of this).

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. Generalizability 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).

4: Basics of qualitative research

Learning objectives: understand key concepts of qualitative 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	In-depth 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

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

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).

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

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, and 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. qualitative 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 analyze 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 are 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-centered 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 whom, 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. Note taking 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 (e.g. loneliness) appears – categories these quantitatively and do statistical analyses.
Thematic analysis: Find all units of data (e.g. sentences or paragraphs) referring to loneliness, code them and look for patterns (e.g. certain times and conditions where carers feel lonely).
Theoretical analysis (e.g. 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 analyse their data but they cannot analyse 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 by QSR International Pty Ltd; QDA Miner by Provalis Research. Excel and SPSS can also be utilised for qualitative analysis, but they are not designed for that purpose.

5: Measuring instruments

Learning objectives: understand three key criteria for a measuring instrument: reliability, validity & precision, and scales of measurement

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 the same sample using different parts of the same instrument (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.

Note that mathematical formulae should be perfectly reliable, as they always give the same answer.

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?

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.)

Note that the word validity is also used in evaluating a research design – does it have internal validity (are the conclusions about the independent variables causing the dependent variables correct, see the earlier discussion on association versus causation) and does it have external validity (can the findings be generalized outside the conditions of the study).

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 newborn 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). Personalised does not necessarily mean it is not objective, it can just mean that the measuring instrument does not work the same way for all sample elements. Objective assessment will almost always be more reliable, but we should not assume that it must be more valid than subjective assessment (consider measuring quality of culture or a user interface). A good strategy is often to use multiple measuring tools simultaneously to measure different dimensions.

Making operational choices (how to measure something)

Some examples of operational choices:

Research Question: Is the lecture room at a comfortable temperature?
Concept: “comfort of the temperature in the lecture room”
Possible operationalizations:
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 – this is measuring temperature, not comfort - 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
ask everyone in the room how they feel about the temperature 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 operationalizations:
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, making the figures imprecise – 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 so it may be measuring connections and abilities of the individual instead and also the US and parts of Europe have many uncounted illegal immigrants, so migration rates are consistently underestimated.

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

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

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

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!

Operational choice considerations:

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 counts 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 reference point, so ratios and percentages have a consistent meaning.
Examples: income, height, weight, age, 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).

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, and 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

6: Probability

Learning objectives: understand three laws of 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, and 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.

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

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' Law

What is Bayes' Law?

Bayes' Law tells us that:

$$Pr(A \mid B)=Pr(A\&B)/Pr(B)=Pr(B \mid A)Pr(A)/Pr(B)$$

It is easier to remember as:

$$Pr(A \text{ and } B)=Pr(A \mid B)Pr(B)=Pr(B \mid A)Pr(A)$$

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

$$Pr(A \text{ and } B)=Pr(A)Pr(B)$$

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' 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 (see https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering for details)

Bayesian updating of evidence

Bayes' 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/not D | S) =Pr(D | S)/Pr(not D | S) and
Odds(D/not D | not S) =Pr(D | not S)/Pr(not D | not S)

Bayes' law gives us:
Pr(D | S)=Pr(S | D)Pr(D)/Pr(S)
Pr(not D | S)=Pr(S | not D)Pr(not D)/Pr(S)

so
Pr(D | S)/Pr(not D | S) =Pr(S | D)/(Pr(S | not D) x Pr(D)/Pr(not D) =0.98/0.02 x 10^{-3}
(statistical jargon is posterior odds = likelihood ratio x prior odds)

which is about 0.05 or 1 in 20
Pr(D | not S)/Pr(not D | not S) = Pr(not S | D)/Pr(not S | not D) x prior odds = 0.02/0.98 x 10^{-3}

which is about 2 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 operators 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.

7: Statistical Computing

Learning objectives: understand use of computers for statistical analysis

Computer packages for statistical analysis

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

IBM SPSS - very popular for desktop use, but expensive and not very flexible (in terms of adding new procedures). Free to use for current HKU staff and students: https://www.its.hku.hk/services/procurement/software/license#spss

SAS - popular in finance sector, very flexible but less user-friendly than SPSS

Stata - popular with epidemiologists and easy to add new procedures

JMP - user-friendly, with a free demo version at http://www.jmp.com. You can 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.

Many other software tools can also be used to do statistical calculations, although statistics is not their primary focus (e.g. Excel, Matlab etc.)

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.

I run a Graduate School workshop on how to use R.

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 computing?

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.

8: Summarising data

Learning objectives: understand how best to summarise 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. Stem-and leaf can be used for any number of digits, but it is simplest with 2 digits. We need to divide up the digits into those used for the stem and those used for the leaves. We can scale the numbers and note the scale below (for example with 1 decimal place, multiply by 10 so that the numbers are integers). 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 if the number of stems is small, i.e. put leaves 0-4 on one stem, 5-9 on another, so the previous plot becomes

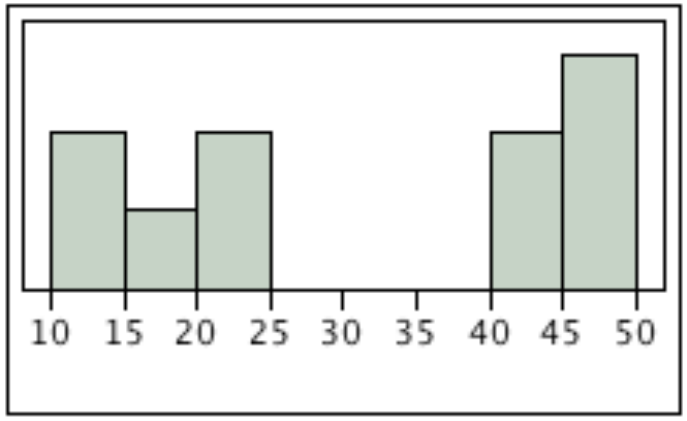
4 | 669
4 | 02
3 |
3 |
2 |
2 | 04
1 | 6
1 | 24

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)

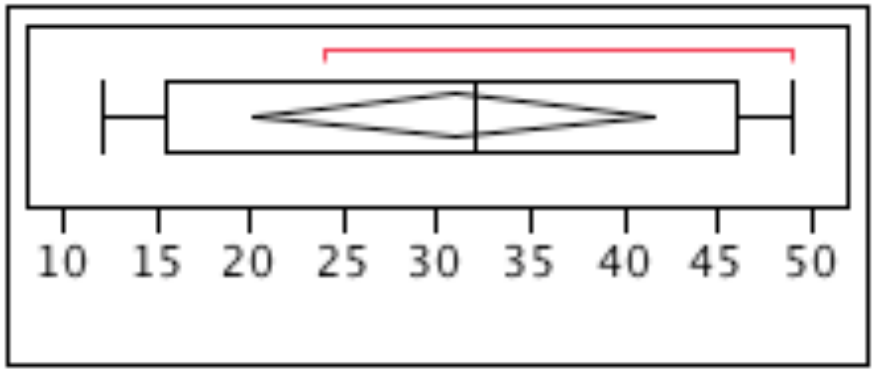
Example: We choose (equal) group widths of 5, so the bar chart and histogram are same:

Size



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.

Size



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

Numerical summaries for center 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, so sample mean=309/10=30.9 for our example.

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. Hint: you can check your answer by counting up and counting down to check that you get the same answer, which is the average of the 5^{th} and 6^{th} values, i.e. (24+40)/2=32 for our example.

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 the 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 (because log function does not change the order of the data values, the log of the middle value is the middle value of the log values), but this isn't true 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. (To get the quartiles: use the median to divide the ordered data set into

two halves - if there are an odd number of data points in the original data set, include the median in both halves, or if there are an even number of data points in the original data set, split this data set exactly in half. The lower quartile value is the median of the lower half of the data, while the upper quartile value is the median of the upper half of the data). The interquartile range is the size of the box in the boxplot.

Example:

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

9: Estimation and Hypothesis testing

Learning objectives: understand how to test hypotheses and estimate population characteristics using statistics

Estimating means or 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%

We call this interval the 95% C.I., which 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 means or 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. In fact, 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 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. 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.

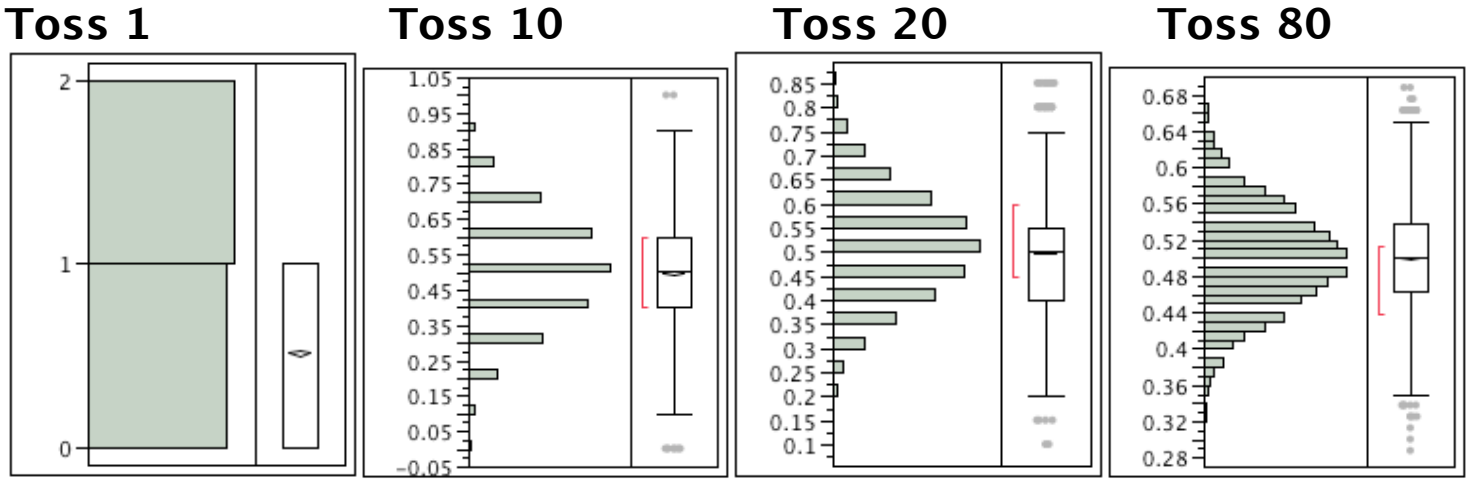
The distribution of the sample mean is approximately Normal (a known distribution) for sample sizes greater than about 30, (almost) regardless of the original population distribution[2].

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.

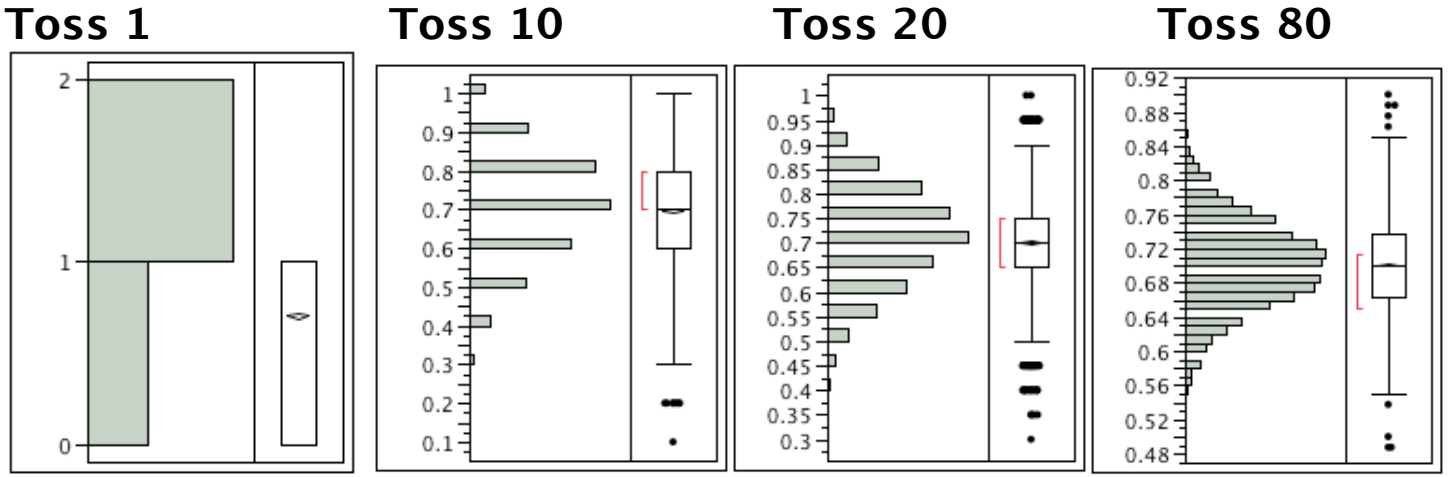
Example: Simulation of tossing a coin. We will do 5000 simulations of different experiments. First, we consider a fair coin (i.e. Pr(H)=0.5). Let us compare the histogram and box plot for the proportion of Heads in 5000 simulations if we do 1, 10, 20 or 80 tosses in each experiment

---

[2] Independent, identically distributed data with finite variance is a sufficient condition.



Even with 20 tosses, histogram is bell-shaped; as sample size increases, spread (width) decreases, centre of the distribution close to 0.5. Now unfair coin (Pr(H)=0.7)



As sample size increases, get symmetrical histogram, decreasing spread, centre 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

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

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

As p x (1-p) is always less or equal to 1/4, then the standard error will always be less than $1/\sqrt{(4}$ x 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
approximate 95% C.I. of: Sample proportion +- 1 / √(sample size)

and a 99% C.I. of: Sample proportion of +- 1.32/√(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

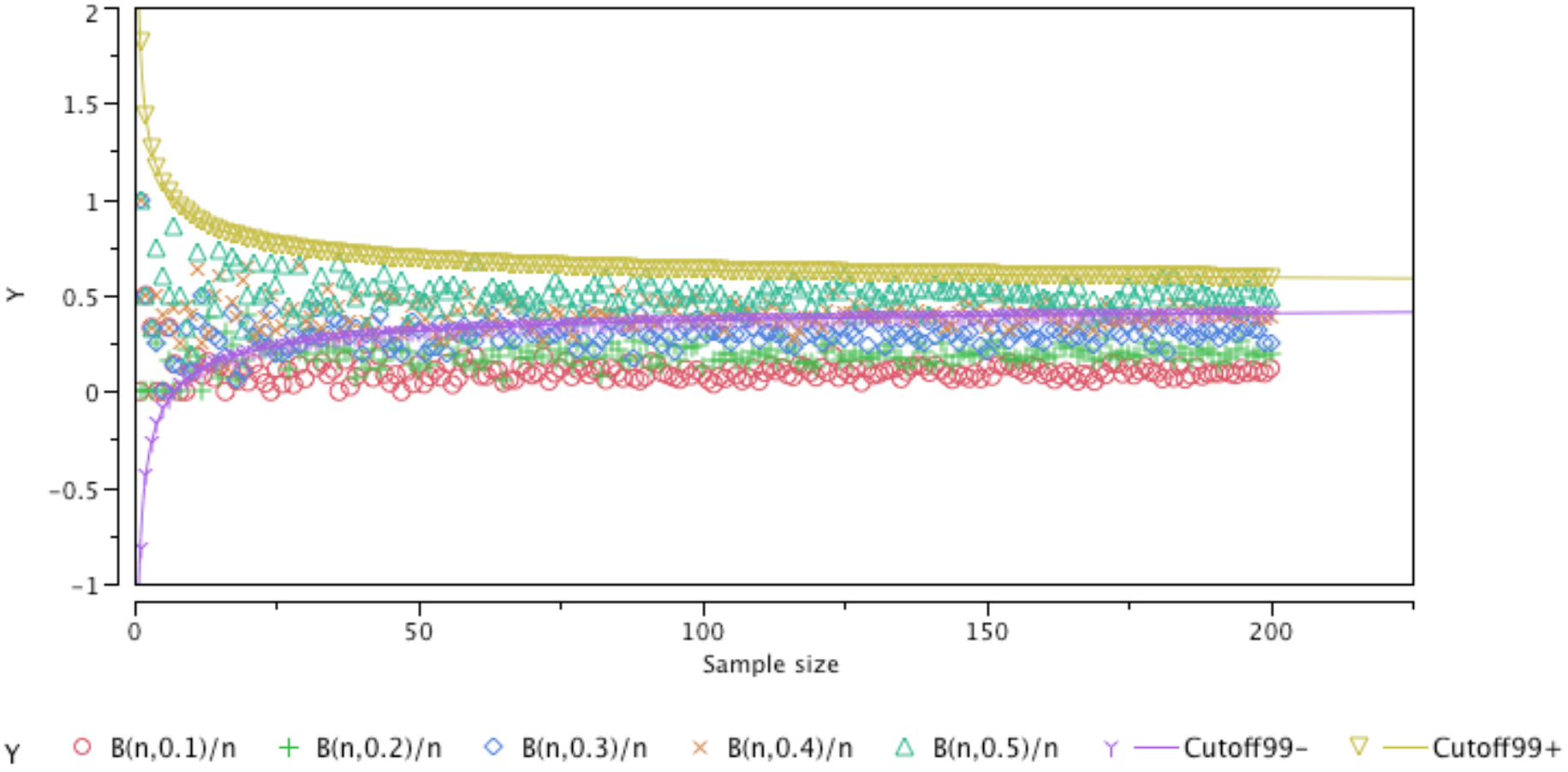
How easy it is to detect that it is not fair depends on 2 things:

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: the sample proportion is 1, so we need only 8 tosses to reject fair coin hypothesis at p= 0.01 using a 2-tailed test (2/256=0.0078<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. This means that if the marker is outside the confidence interval, we would reject the null hypothesis of a fair coin. The graph shows that it is easier to detect an unfair coin if the sample size is larger or the coin is more unfair.



For Pr(H)=0.1 (red circles), a sample of 15 is plenty, for Pr(H)=0.2 (green +), need a sample of about 30, for Pr(H)=0.3 (blue diamond), 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 200 experiments, so expect about 2 mistakes on average.

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 $Z_{0.005}$ (we can get these values from statistical tables or calculator). Z is what we call a quantile for the standardized Normal distribution, where standardized means a Normal distribution with zero mean and unit variance.

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!

Observed significance level

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?

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, because there is little extra uncertainty from estimating the variance.

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

(sample mean-population mean)/estimated standard error

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

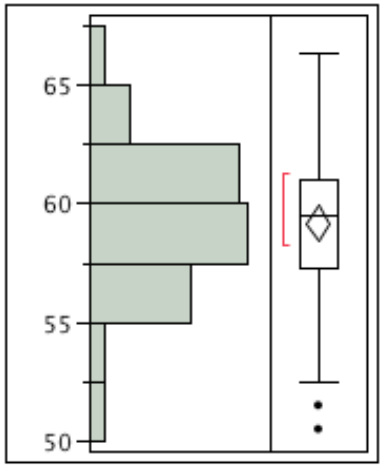
So, instead of +/-1.96 as interval, we will get a slightly larger number and hence a wider interval than when using the Z-value, unless the sample size is large.

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.

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

Distributions

Height



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

Distributions

Height

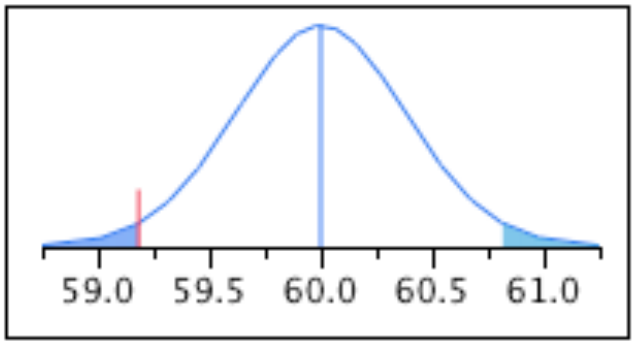
Confidence Intervals

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*



The problem 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 (known as p-hacking if you fail to adjust). This can be adjusted for using the Bonferroni correction - the effective significance level is the stated significance level multiplied by the number of hypotheses tested. However, it is then very hard to detect effects when the number of tests (m) is large (the power is low, meaning we are unlikely to detect that the null hypothesis is false). However, we can instead control the false discovery rate (proportion of discoveries which are false, where discovery means the null hypothesis is rejected) – this assumes that we care about the proportion of false rejections, rather than the number of false rejections.

Benjamini–Hochberg procedure

The Benjamini–Hochberg procedure (BH step-up procedure) controls the False Discovery Rate at level alpha. It works as follows:

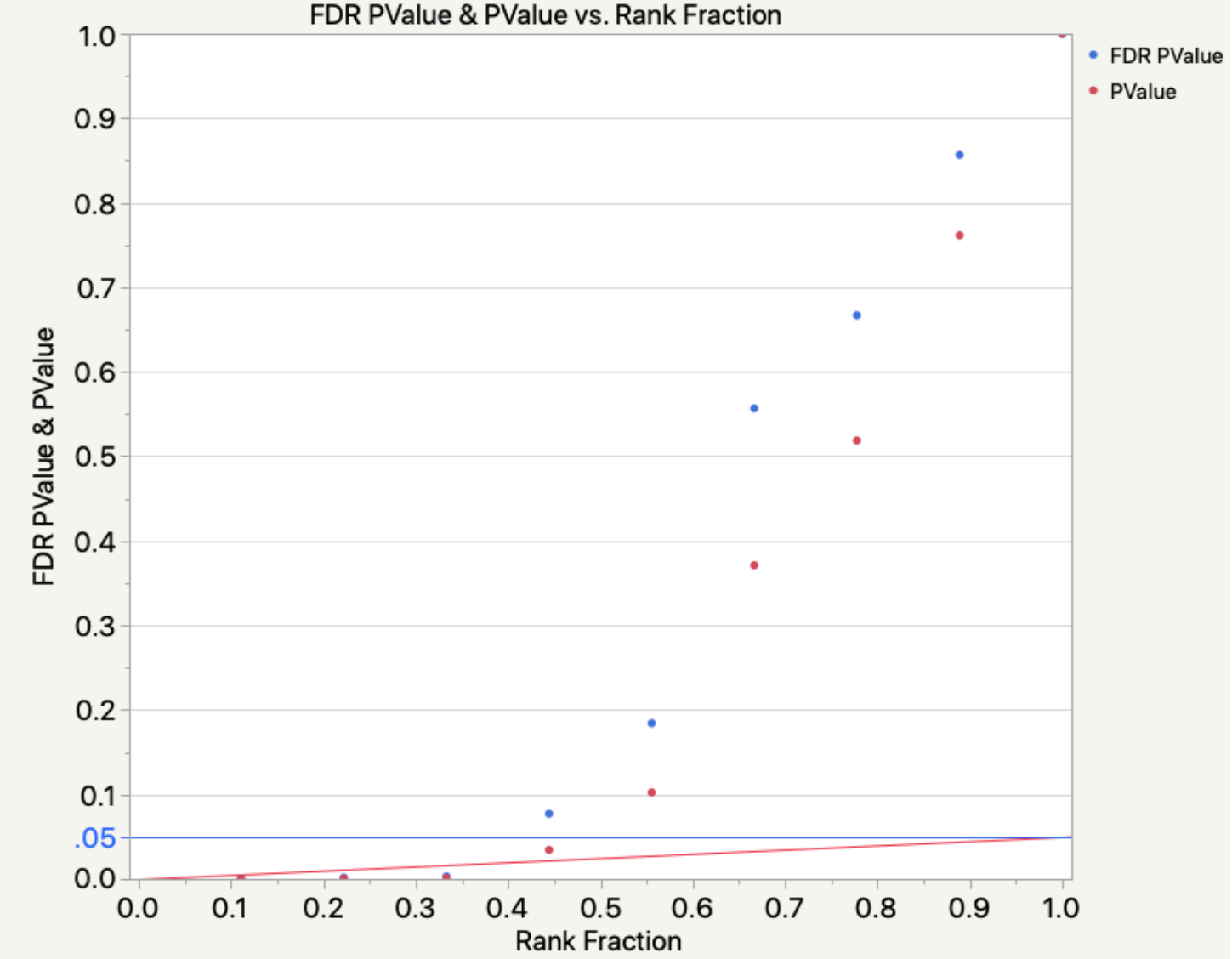
1. We have m null hypotheses tested and we list the m p-values in ascending order and denote them by $P_{(1)}$-$P_{(m)}$.
2. For a given alpha, find the largest k such that $P_{(k)}$ is less than k x alpha/m
3. Reject the null hypothesis (i.e., declare discoveries) for i=1 to k

Geometrically, this corresponds to plotting $P_{(k)}$ vs. k/m (on the y and x axes respectively), drawing the line through the origin with slope alpha, and declaring discoveries for all points on the left up to and including the last point that is below the line. This procedure is valid when the m tests are independent, and also in various scenarios of dependence, but is not universally valid.

We examine a simple example below with 9 tests and alpha =0.05. Bonferroni would suggest the unadjusted p-value should be 0.0055, while BH looks at the line with slope 0.05 and rejecting all tests until there is one above the line (i.e. reject the first 3).

Response Screening

FDR PValue Plot



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.

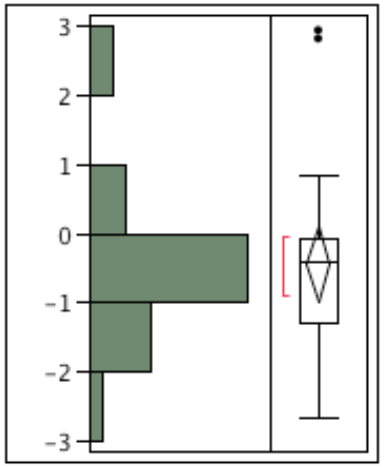
If we improve the score of half the individuals by 1 and half by 2, then we now have a new average score of 51.5. If we match individuals and look at the change, we have 64 differences with mean 1.5 and standard deviation 0.5, so the 99% confidence interval for the mean change is (1.33,1.67), so we are 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 almost always do not know the population standard deviation for change and need to 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

Distributions

diff

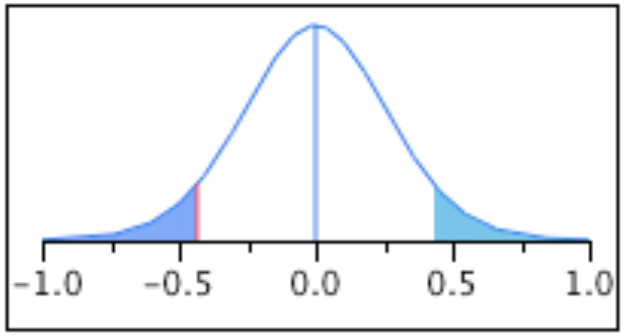
Confidence Intervals

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 and need to be estimated from the samples

Note: If we assume that population variance 1 = population variance 2, we can use a pooled (combined) estimate of the variability, based on both samples, which simplifies the statistical calculations, but that is more complex in that we need first to check whether the variances are equal using Levene's test, so this is approach is not recommended. Instead use Welch's two-sample t-test, which does not assume equal variances, for which the degrees of freedom calculation is more complex, but this is handled by the statistical package automatically.

Reference: Welch, B. L. (1947). "The generalization of "Student's" problem when several different population variances are involved". *Biometrika*. **34** (1–2): 28–35. doi:10.1093/biomet/34.1-2.28.

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.

Oneway Analysis of Height By Gender

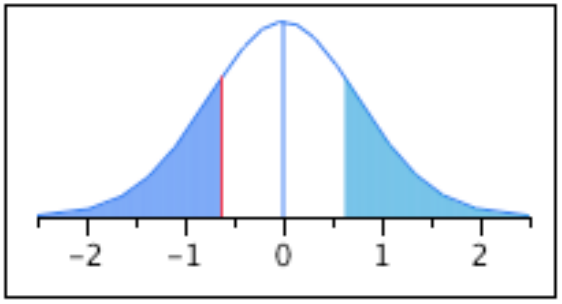
Oneway Anova

t Test

m-f

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



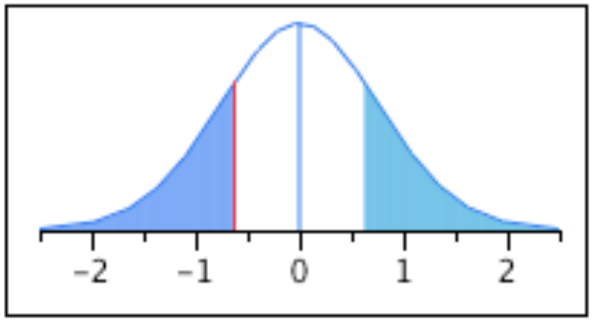
Oneway Analysis of Height By Gender

t Test

m-f

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 for that category and Expected means the count we would expect for that category if the null hypothesis is true. X^2 sums over all categories.

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)).

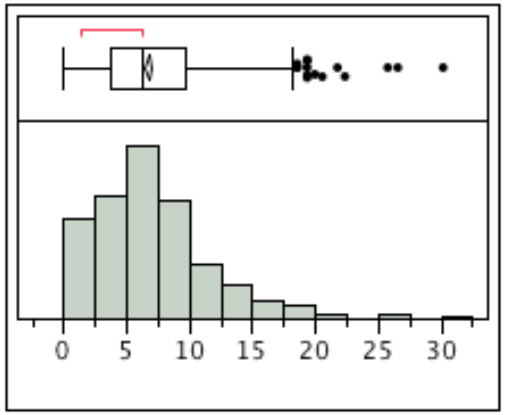
Note: we are assuming that the data is nominal scale, not ordinal. If the data is ordinal scale, then there are more sensitive statistical methods that should be used.

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. If the counts are E=18,C=31,P=11 for a sample size of 60, then under the null, the expected values are all 20, so $X^2 = ((18-20)^2+(31-20)^2+(11-20)^2)/20 = (4+121+81)/20=10.3$.

Sample of 30

Distributions

X2



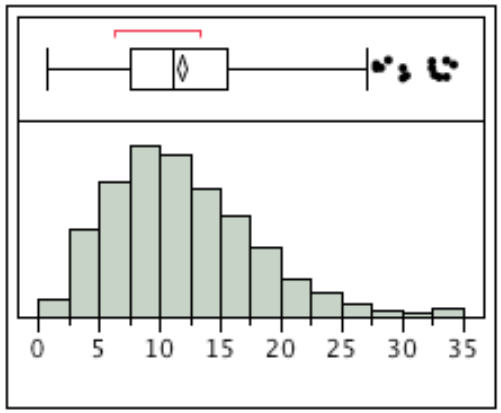
Quantiles

100.0%	maximum	30.2
99.5%		21.794
97.5%		18.165
90.0%		12.6
75.0%	quartile	9.6
50.0%	median	6.2
25.0%	quartile	3.8
10.0%		1.4
2.5%		0.605
0.5%		0.2
0.0%	minimum	0

Sample of 60

Distributions

X2



Quantiles

100.0%	maximum	34.3
99.5%		32.699
97.5%		25.59
90.0%		19.6
75.0%	quartile	15.6
50.0%	median	11.1
25.0%	quartile	7.5
10.0%		4.9
2.5%		2.8
0.5%		1.3015
0.0%	minimum	0.7

If we use statistical tables for Chi-squared with 2 degrees of freedom, $Pr(X^2>5.99) = 0.05$, so the power for a sample of 30 is about 50%, sample of 60 is about 85% (based on the proportion of values greater than 5.99).

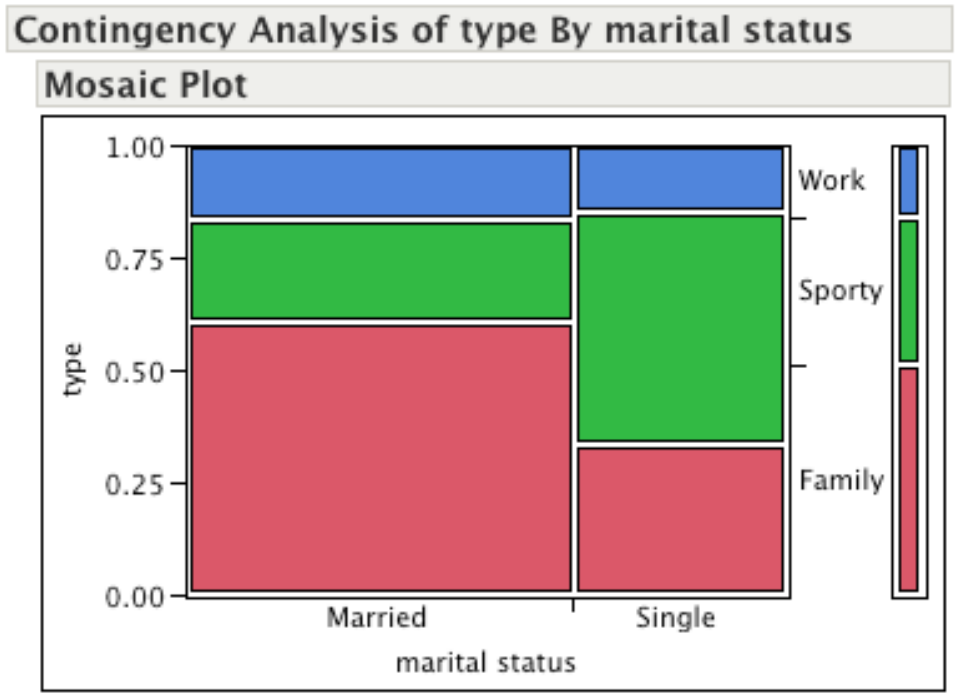
10: Relationships between pairs of variables

Learning objectives: understand statistical tools for relationships

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

Contingency Table

marital status	Count / Total % / Col % / Row %	type: Family	type: Sporty	type: 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*

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.

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).

In the example above, the question would be whether we can reject the hypothesis that the preferred car type and marital status are independent in our population. The alternative hypothesis is that there is some (unstated) association between them.

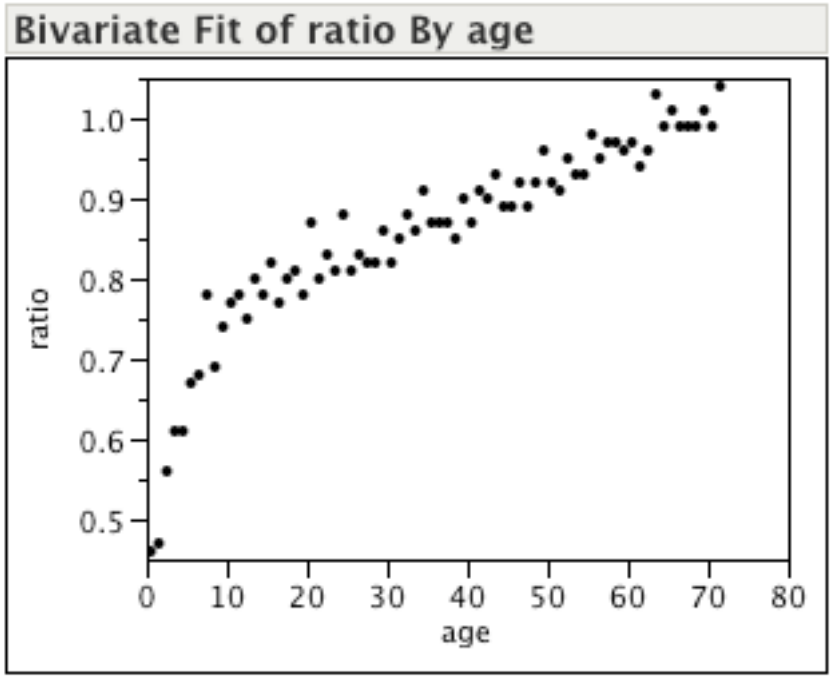
It turns out that we can use the same concept (GoF) and formula as we used for 1 variable with multiple categories above.

In the case of our example, the null hypothesis implies that the proportion of married people who prefer family cars should be the same as the proportion of singles who prefer family cars or equivalently, the proportion of those preferring family cars who are single should be the same as the proportion of those preferring sporty cars who are single. In other words the probability of preferring a type of car is independent of the probability of being married. Given that independent probabilities multiply, this means that the expected number of married people preferring a family car is the total sample size times the sample proportion of married people overall times the sample proportion of those preferring family cars overall; similarly for the other five combinations of marital status and car preference.

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 this 3 x 2 case, the degrees of freedom are 2. The expected count for the first cell (i.e. married man prefers family car) is: 303 x (196/303) x (155/303) = 100.3 so the X^2 contribution for the first cell is $(119-100.3)^2/100.3 = 3.5$. Similarly for the other 5 cells. Note: Need expected count of at least 5 in each cell.

Note: there are more powerful methods if the variables are not both nominal scale, such as Spearman rank correlation if both are ordinal scale and Kruskal-Wallis test if one is ordinal scale and one is nominal scale.

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



Response ratio

Whole Model

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

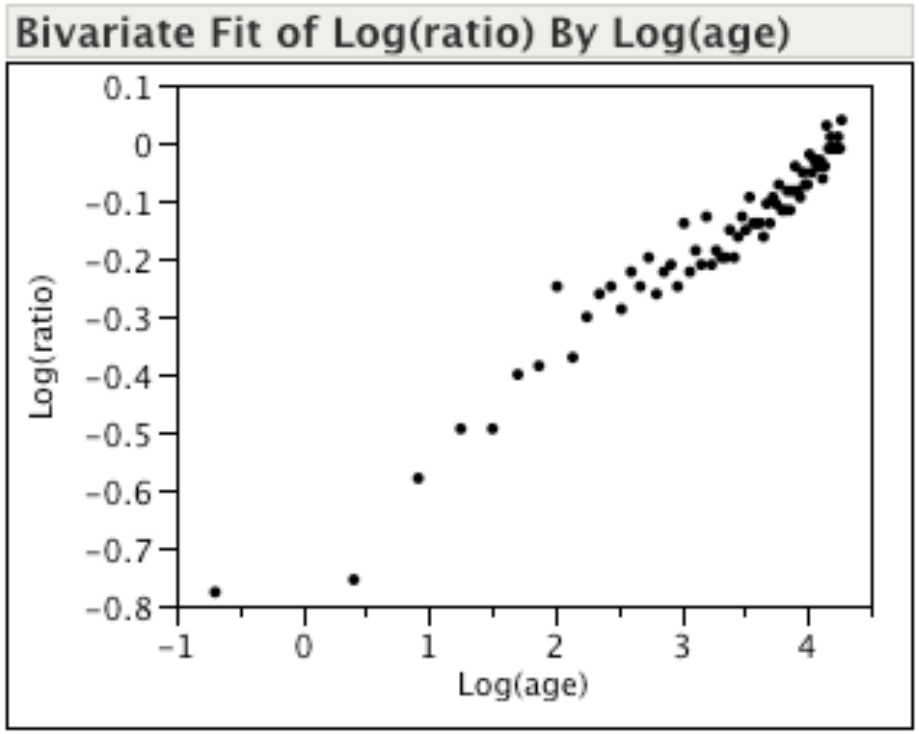
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.6656231	0.012176	54.67	<.0001*
age	0.0052759	0.000293	18.01	<.0001*

Prediction Expression

0.66562308401398

+ 0.00527590198727 * age

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



Response Log(ratio)

Whole Model

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*

Prediction Expression

-0.6959708296059

+ 0.16096040022327 * Log(age)

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 summarize 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:

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

Where:

$$\text{Covariance}(X,Y)=E((X-\text{Mean}(X))(Y-\text{Mean}(Y)); \qquad \text{Var}(X)=E((X-\text{Mean}(X))^2)$$

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:

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

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:

$\sum$(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:

$b = \sum(Y-Mean(Y))(X-Mean(X))/ \sum(X-Mean(X))(X-Mean(X))$
$= Cov(X,Y)/Var(X)$

$a = Mean(Y) - b \times Mean(X)$

(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 minimize s^2

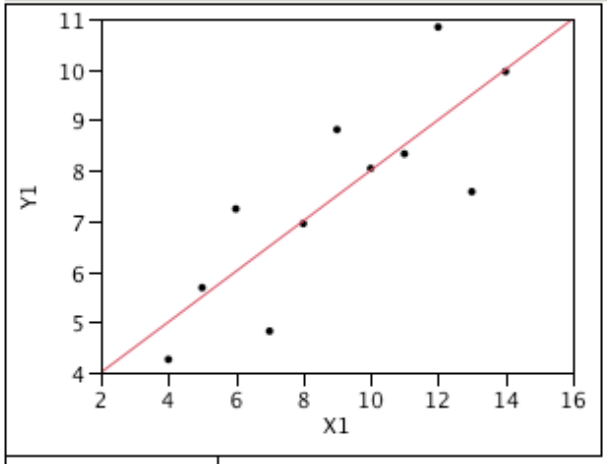
However, an important word of warning:

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.

Moderate linear relationship

Bivariate Fit of Y1 By X1



Linear Fit

Y1 = 3.0000909 + 0.5000909*X1

Summary of Fit

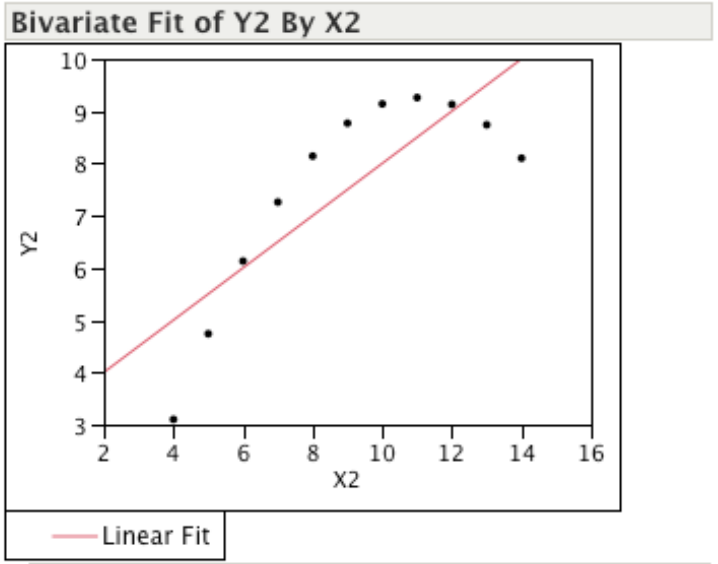
RSquare	0.666542
RSquare Adj	0.629492
Root Mean Square Error	1.236603
Mean of Response	7.500909
Observations (or Sum Wgts)	11

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*

Quadratic relationship here

Bivariate Fit of Y2 By X2



Linear Fit

Y2 = 3.0009091 + 0.5*X2

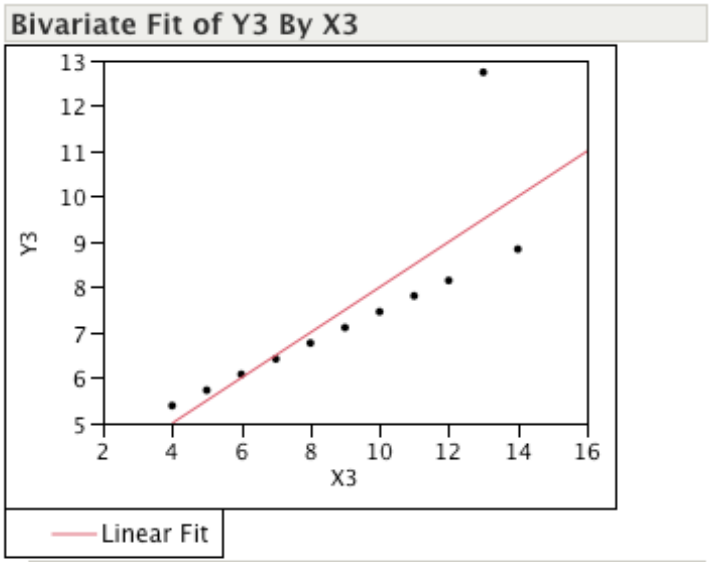
Summary of Fit

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*

The third example shows a different pattern:



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.

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

$s_b=s/\sqrt{(\sum(X-\text{Mean}(X))^2)}$ 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

$b \pm s_b \times t_{\alpha/2}(n-2)$, 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.

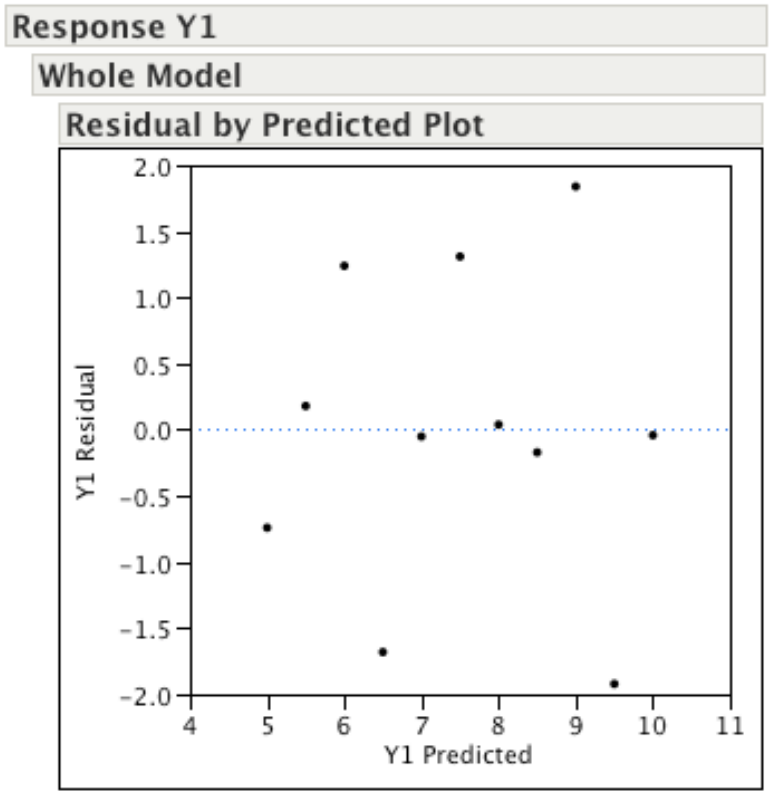
(We could also test separately for A=0 in a similar way, which would mean we are testing if Y is proportional to X, which is rarely meaningful).

Residuals

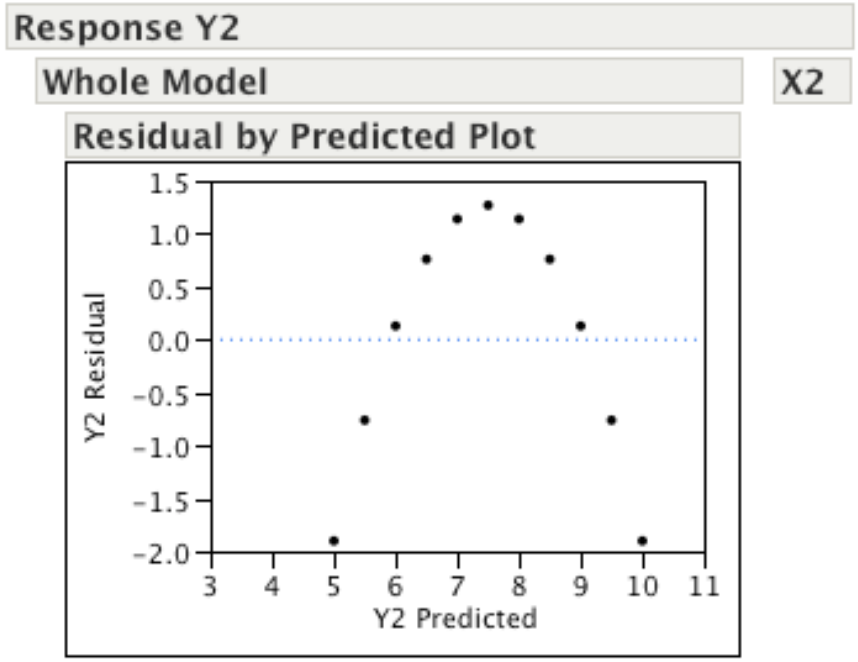
As with any tool, diagnostics are important. The simplest diagnostics for regression are the residuals.

2 key ideas: look for particular data points that look to fit badly (possible error?) and for patterns that suggest model errors

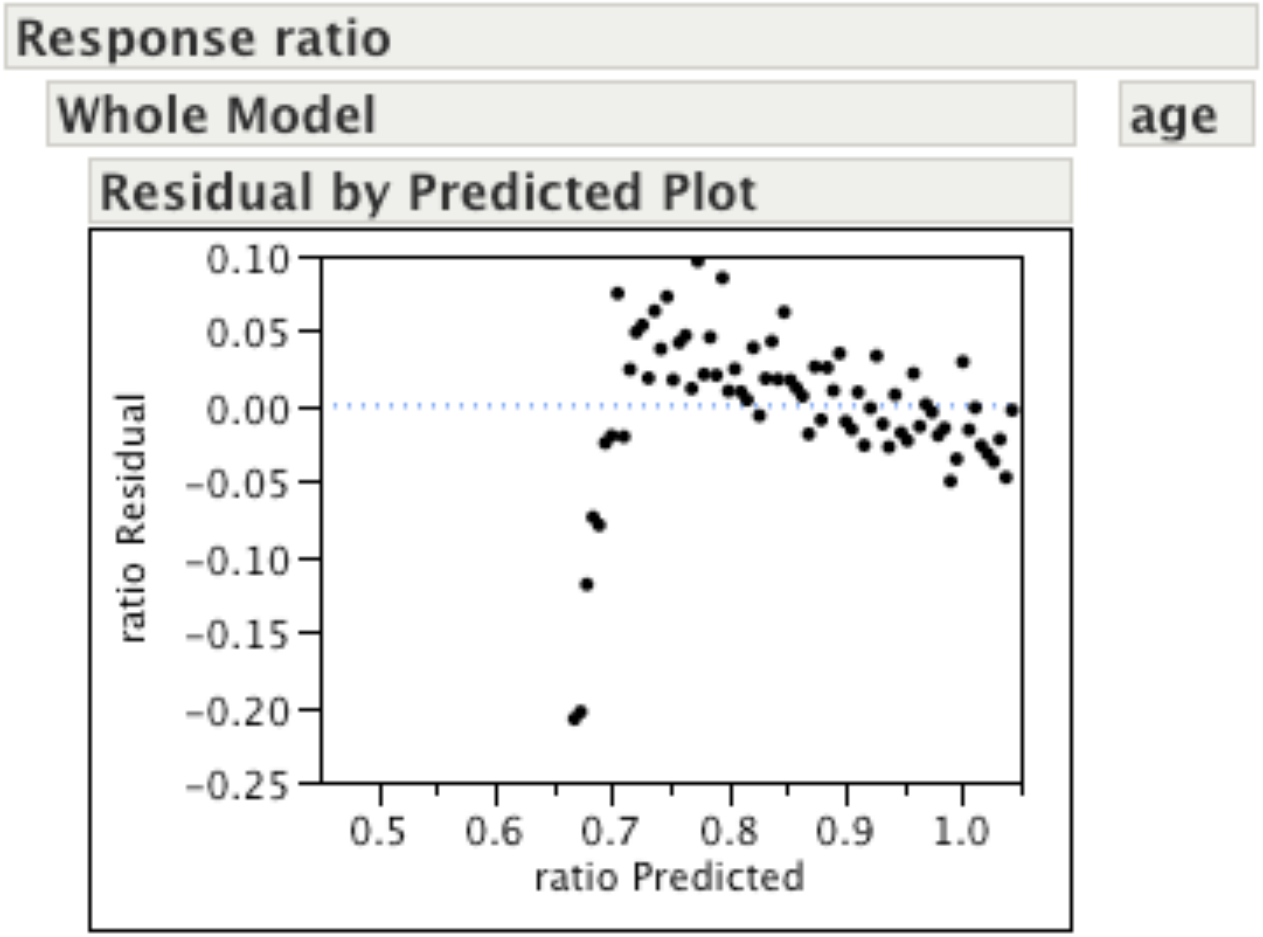
No obvious pattern



Can see quadratic pattern so clearly!



Here the growth example shows the strong non-linear relationship



Meaning for r^2

r^2 summarizes 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. $y_{fit}=\text{Mean}(y)$), 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

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

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

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

(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.

11: Multiple Regression

Learning objectives: understand statistical models for multiple continuous variables

All the ideas we have used can be easily extended to handle more than 1 independent variable.

Now our equation can be written as:

$Y=B0 \times 1 + B1 \times X1 + B2 \times X2 \ldots + B(p-1) \times X(p-1) + \varepsilon$

where B0 is now the intercept (previously=A for simple regression) and B1 to B(p-1) are the slopes (previously=B)

The advantage of writing it this way is that we can also write this in matrix and vector notation as:

$Y=X\ B + E$

where Y is an n x 1 vector and X is an n x p matrix (first column is all ones) and B is a p x 1 vector and E is the n x 1 vector of errors.

Note: a matrix is a table of numbers, while a vector is a column of numbers.

One reason for writing it in matrix notation is that we can show via calculus that the least squares solution has a very simple form in matrix notation:

$b=(X^TX)^{-1}X^TY$, where X^T is the transpose of X, meaning we swap rows and columns

$Y_{fitted} = Xb = HY$

where $H=X(X^TX)^{-1}X^T$ is a projection matrix (called the Hat matrix) that "projects" from n dimensional space onto p dimensional space to obtain the closest fit to Y (in terms of squared errors), where n is the sample size and p is the number of parameters in the model.

For example, the null model which has p=1 (i.e. Y_{fitted} =Mean(Y)) has H that projects Y onto the Mean(Y)
The simple linear model with p=2 has H that projects Y onto a 2 dimensional space, i.e. the plane that goes through Mean(Y) and X, because we already saw that the fitted line goes through Mean(Y) and has equation =a+bX

For projection, you can consider the analogy of a video projector, which casts a 2D shadow (projection) on the screen of a 3D object (e.g. your hand). This is a projection from 3D onto 2D (like a data set of 3 points when we fit a simple regression line with just one X variable)

We can show using calculus that this is the optimal solution under quite general conditions. You do not need to know this formula, it is just provided to assist understanding of what we are doing when fitting a multiple regression model using least squares.

The calculations can always be done using a spreadsheet or a statistical package.

The estimate of the error variance is: $s^2=\sum(y-y_{fitted})^2/(n-p)= \sum r_i^2/(n-p)$

The divisor is n-p to account for the p parameters that we have fitted – we can fit p points exactly (compare to 2 points fit exactly for the simple regression)

For testing $B_i=0$ or finding a confidence interval for B_i, the idea is similar to simple linear regression:

95% CI for B_i is: $B_i=b_i+/-t_{.025}\ SEB_i$

And reject Bi=0 at 5% (two-tailed test) if b_i is > $t_{.025}\ SEB_i$ or b_i<- $t_{.025}\ SEB_i$, where $SEB_i=\sqrt{(X^TX)^{-1}_{ii}}\ s$

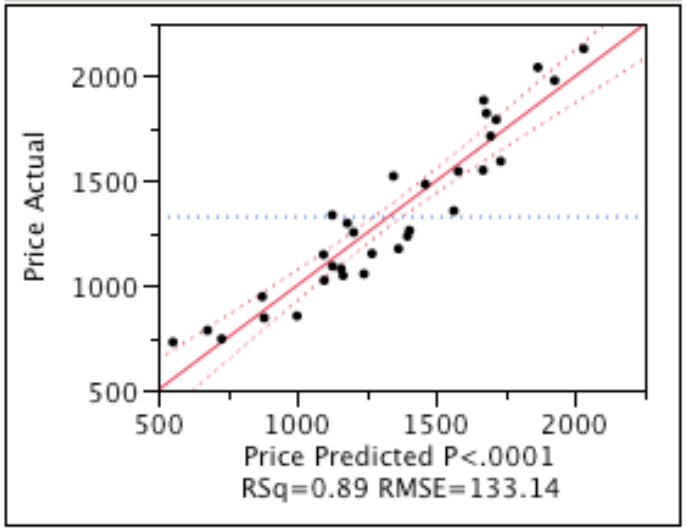
but again, we rely on the computer to calculate t and SEB.

This example is about predicting the auction price of grandfather clocks, given the age and numbers of bidders – clearly both independent variables are important.

Response Price

Whole Model

Actual by Predicted Plot



Response Price

Whole Model

Parameter Estimates

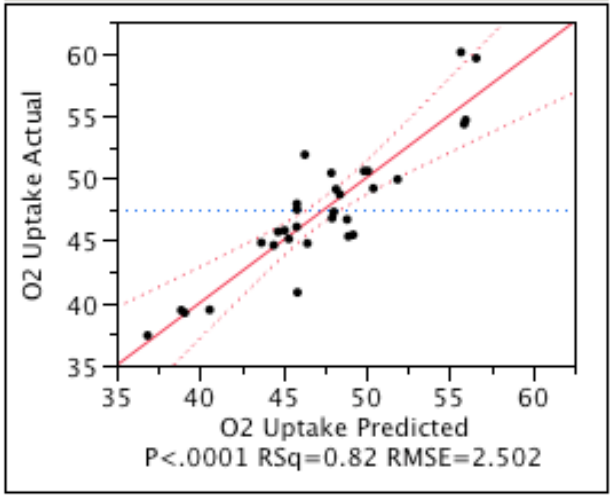
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1336.722	173.3561	-7.71	<.0001*
Age	12.736199	0.90238	14.11	<.0001*
Bidders	85.815133	8.705757	9.86	<.0001*

This example is predicting the oxygen uptake for runners, given their age, weight, run time, resting pulse, running pulse and maximum pulse. Clearly run time is very important and age and running pulse matter.

Response O2 Uptake

Whole Model

Actual by Predicted Plot



Response O2 Uptake

Whole Model

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	109.25895	14.15151	7.72	<.0001*
Age	-0.271473	0.10514	-2.58	0.0164*
Weight	-0.058402	0.0584	-1.00	0.3273
Run Time	-2.64042	0.419615	-6.29	<.0001*
Rest Pulse	-0.032821	0.071559	-0.46	0.6506
Run Pulse	-0.152575	0.060066	-2.54	0.0180*
Max Pulse	0.0635511	0.070483	0.90	0.3762

This leads to the obvious question: how to select the optimal set of independent variables in a regression?

We need to think carefully first - variables may be included in a model for 2 reasons:

a) theoretically they are important or statistical controls
b) they can be shown statistically to be important (i.e. B=0 is rejected)

Theoretically important variables that we need to statistically control for (by including them in the model) are called confounders. In this case, we need to include variables even if they are not statistically significant.

Ideally, we would randomize our experiment and avoid the problem of confounders, but this is often not possible unless we have a designed experiment (we often have observational data or retrospective data).

First, how do we test our complete model (vs. testing individual variables)?

We use a concept called Analysis of Variance (ANOVA)
The idea is to partition variability into different parts with different explanations that add up to a total that is related to overall variance.

The variability is measured in terms of sums of squared deviations (SS is the abbreviation for Sums of Squares, which means the sums of squared deviations). We try to partition (split up) the variability into different components due to different variables.

The total variability is essentially just the variance for Y times (n-1): Total SS=$\sum$(Y-Mean(Y))^2

We then divide this up - what is left after fitting the model is called the residual:

Residual SS=$\sum(Y-Y_{fitted})^2=\sum r_i^2$, (i.e. the remaining part that is assumed to be random error).

This is just (n-p) times s^2

The difference (meaning the part "explained" by our model) is: Model SS= $\sum(Y-Mean(Y))^2-\sum(Y-Y_{fitted})^2$

For each SS, we also have the degrees of freedom (d.o.f.), which is the dimension of the space spanned by this part, so d.o.f. for the Total SS is (n-1) as we have n data points, but have subtracted off the mean.
For the simple linear model, the Model d.o.f. is 1 (1 independent variable) and the Residual d.o.f. is the difference, i.e. n-2

In the multiple regression model the Model d.o.f. is p-1 (the number of independent variables) and the Residual d.o.f. is the difference, i.e. n-p

We create an intermediate stage called the Mean Square (MS), which is the SS divided by the d.o.f.

Model MS=(Model SS/Model d.o.f.)
Residual MS=(Residual SS/Residual d.o.f) (=s^2)

If the null hypothesis is true, then the Model SS should be just random variation, so that both the Model MS and Residual MS become estimates of the same variance, so the ratio will be around 1. However, if the null is false, the ratio should be large.

So this explains why our test statistic is: F=Model MS/Residual MS,

which is called the F statistic, with Model d.o.f. and Residual d.o.f. degrees of freedom and we reject the null hypothesis if F is large enough.

We can easily look up the distribution of F under the null hypothesis (i.e. that all the B_i=0)

For the simple linear model, the F statistic is mathematically equivalent to the t statistic (it is actually the square of the t statistic) for the slope of X, so it yields the same p value as the two-tailed t-test in that case.

For the multiple regression model, the F statistic is the overall test that combines all the individual t tests into one omnibus test. If we cannot reject the omnibus test, we do not need to spend time trying to identify the best model as it means that all the variables combined do not provide a better fit than just using the mean for Y.

We can also try to divide up the Model SS amongst the variables, but this is tricky, unless the Xs are uncorrelated (possible in designed experiments), as otherwise, the SS for each variable depends on the order we add it in.

We can also expand the idea of r^2 to what is sometimes often called R^2 (with a capital R) to distinguish that it is the proportion of variation "explained" by a set of variables (versus just one variable for r^2).

It relates simply to the A.o.V. table as: R^2=Model SS/Total SS

At first glance, it looks sensible to choose the model with the largest R^2. However, this is flawed, because each time you add a new variable, R^2 will increase, even if the new variable is unrelated to Y. This means R^2 is only useful in comparing models with the same number of variables. For comparing models with different numbers of variables, we should use what is called adjusted R^2

We can write R^2 and adjusted R^2 as:

R^2 = 1-(Residual SS)/(Total SS) =1-(Residual SS/(n-1))/(Total SS/(n-1))

Adjusted R^2 = 1-(Residual SS/(n-p))/(Total SS/(n-1)) = $1-s^2/s_0^2$,

where s_0^2 = Total SS/(n-1) is the variance of Y.

This adjusts for the bias in R^2 towards models with more variables – maximizing Adjusted R^2 is equivalent to minimizing s^2.

In fact, adjusted R^2 is close to R^2 if p is small compared to n, so (n-p)/(n-1) is close to 1, i.e. if the number of parameters is small compared to the sample size.

Response Price

Whole Model

Summary of Fit

RSquare	0.892713
RSquare Adj	0.885314
Root Mean Square Error	133.1365
Mean of Response	1327.156
Observations (or Sum Wgts)	32

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	4277159.7	2138580	120.6511
Error	29	514034.5	17725	**Prob > F**
C. Total	31	4791194.2		<.0001*

So, how **do** we find and select the best model (i.e. best combination of the independent (x) variables)?

It used to be popular to do stepwise regression. The idea was that you started with a null model, then added in the one variable that yielded the highest t statistic, then choose the variable that gave the highest t statistic when added to the first variable and so on, until there were no more variables to add that have a significant slope. This is called forwards stepwise regression.

The problem is that this process has 2 flaws:

1) it tends to yield models that are too large because you are doing many tests, and choosing the most significant from a range of tests. This means that the true significance level of the model is much higher than claimed, based on the significance of each test (i.e. the risk of falsely rejecting the null is much higher than claimed)
2) it also may exclude some good models because some variables may only be good if included together with another variable, but we are only adding one at a time.

Another approach is to include all the possible variables to start, check if that model is significant, and then drop the least significant variables one at a time, until they are all significant (the t statistic checks for significance conditional on all the other variables already in the model). This is called backwards stepwise regression. This approach has similar problems, plus it may also have an additional problem that we call collinearity.

Collinearity means that different sets of independent variables provide almost the same information. Collinear means - in a line. It means we can predict at least one of the independent variables well from the others. This is quite common if there are a large number of variables. In the case of collinearity, there are many possible coefficients for the Xs that are very close to the best. This is often a problem with observational data. For an experiment, we can ensure that the x values are chosen such that this will not happen (we will return to this when we consider experimental designs).

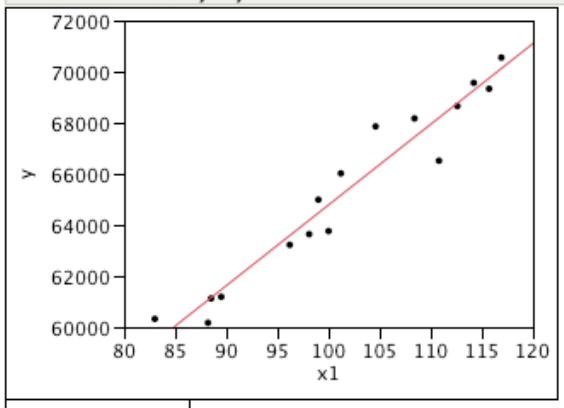
Collinearity is a problem in two senses:

1) It may be hard to estimate the coefficients accurately
2) We need to decide which variables are more important theoretically, and which we can exclude.

We look at a famous dataset due to Longley – we first look at predicting y individually from X1 to X6

Fit Y by X Group

Bivariate Fit of y By x1



Linear Fit

y = 33189.173 + 315.96609*x1

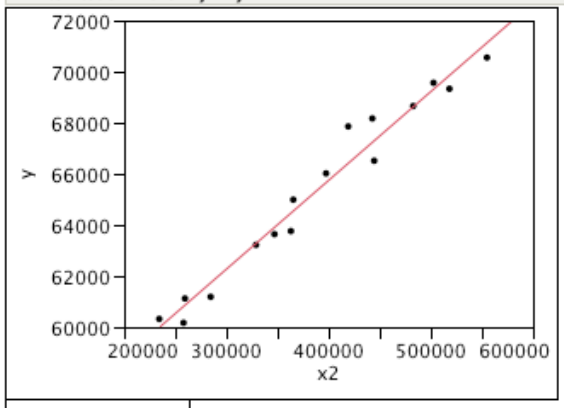
Summary of Fit

RSquare	0.942644
RSquare Adj	0.938547
Root Mean Square Error	870.6064
Mean of Response	65317
Observations (or Sum Wgts)	16

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	33189.173	2129.189	15.59	<.0001*
x1	315.96609	20.83014	15.17	<.0001*

Bivariate Fit of y By x2



Linear Fit

y = 51843.59 + 0.0347523*x2

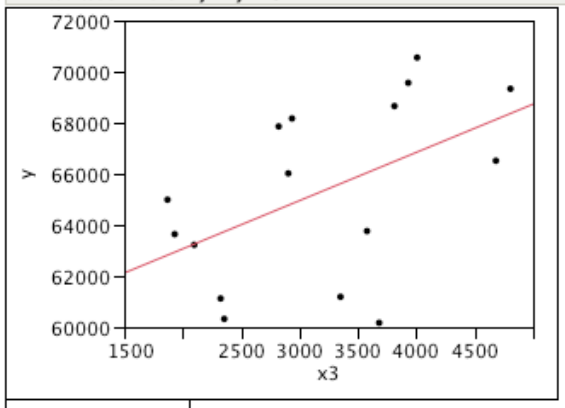
Summary of Fit

RSquare	0.967374
RSquare Adj	0.965043
Root Mean Square Error	656.6223
Mean of Response	65317
Observations (or Sum Wgts)	16

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	51843.59	681.3716	76.09	<.0001*
x2	0.0347523	0.001706	20.37	<.0001*

Bivariate Fit of y By x3



Linear Fit

y = 59286.355 + 1.8885232*x3

Summary of Fit

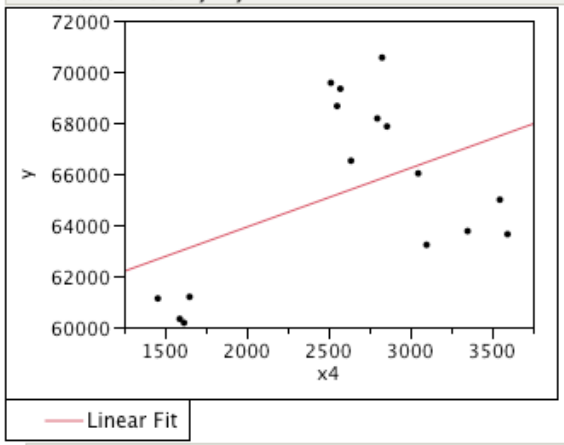
RSquare	0.252504
RSquare Adj	0.199112
Root Mean Square Error	3142.943
Mean of Response	65317
Observations (or Sum Wgts)	16

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	59286.355	2882.293	20.57	<.0001*
x3	1.8885232	0.868417	2.17	0.0473*

Fit Y by X Group

Bivariate Fit of y By x4



Linear Fit

y = 59301.265 + 2.3078084*x4

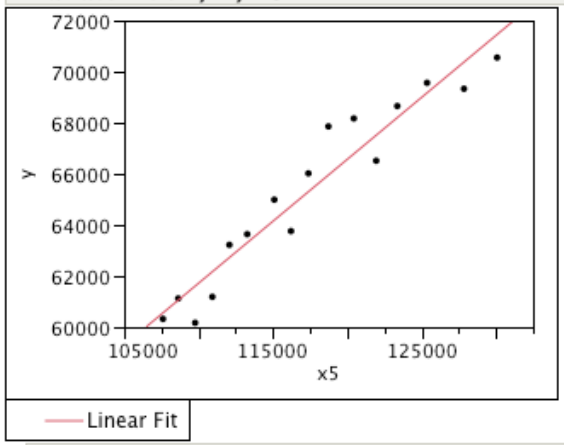
Summary of Fit

RSquare	0.20913
RSquare Adj	0.152639
Root Mean Square Error	3232.844
Mean of Response	65317
Observations (or Sum Wgts)	16

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	59301.265	3229.347	18.36	<.0001*
x4	2.3078084	1.199444	1.92	0.0749

Bivariate Fit of y By x5



Linear Fit

y = 8380.6742 + 0.4848781*x5

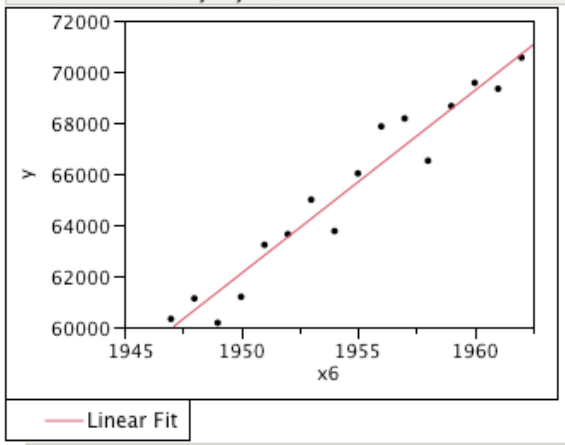
Summary of Fit

RSquare	0.92235
RSquare Adj	0.916804
Root Mean Square Error	1012.984
Mean of Response	65317
Observations (or Sum Wgts)	16

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8380.6742	4422.434	1.90	0.0789
x5	0.4848781	0.0376	12.90	<.0001*

Bivariate Fit of y By x6



Linear Fit

y = -1335105 + 716.51176*x6

Summary of Fit

RSquare	0.943481
RSquare Adj	0.939444
Root Mean Square Error	864.2308
Mean of Response	65317
Observations (or Sum Wgts)	16

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1335105	91606.69	-14.57	<.0001*
x6	716.51176	46.8695	15.29	<.0001*

However, when we try to predict from X1-X6 together, we find something strange – some of the coefficients that are statistically significant change sign and some variables (X1,X2,X5) are not useful.

Response y

Summary of Fit

RSquare	0.995479
RSquare Adj	0.992465
Root Mean Square Error	304.8541
Mean of Response	65317
Observations (or Sum Wgts)	16

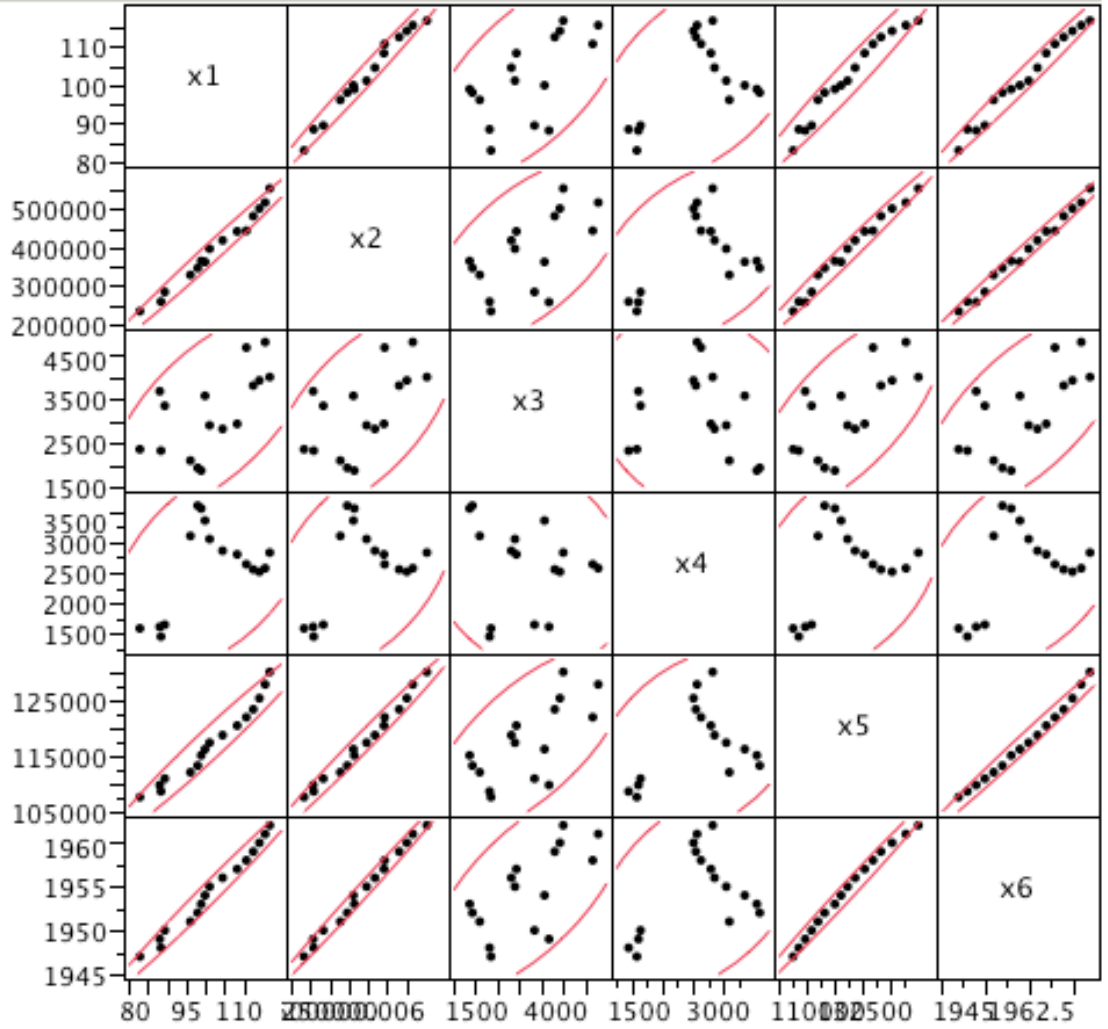
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-3482259	890420.4	-3.91	0.0036*
x1	15.061872	84.91493	0.18	0.8631
x2	-0.035819	0.033491	-1.07	0.3127
x3	-2.02023	0.4884	-4.14	0.0025*
x4	-1.033227	0.214274	-4.82	0.0009*
x5	-0.051104	0.226073	-0.23	0.8262
x6	1829.1515	455.4785	4.02	0.0030*

Why does this happen? We need to look at the correlation between the X's

Multivariate

Scatterplot Matrix



As can be seen, X1,X2,X5 and X6 are all very strongly positively correlated – this is collinearity – it means that these 4 variables have a lot of common information about Y and it turns out that once we know X6 (which is the year), the other 3 variables do not add anything useful about Y.

Another alternative is mixed stepwise, which combines backwards and forwards steps.

One of the reasons stepwise methods were popular was that the computer power available was small, which used to be an issue for large models. Nowadays, that is usually not a problem, so we may use what is called best subsets regression. This means finding the best models for each value of the dimension of our model, i.e. for any given number of variables. We can then review the models to see which ones make sense in terms of our theory.

To decide which model to choose, there are many different approaches. One approach is to start by comparing the simplest and the most complex model.

If this is not significant, then stop, as that suggests that everything we observe, other than the variability linked to variables in the simplest model, is just random, so we should just use that simplest model, which is usually the null model with no independent variables.

If it is significant, then add variables into the simplest model and choose the simplest model such that the F statistic comparing that model against the most complex model is not significant. This is called the lattice approach (construct a tree of all possible models).

Arguably, this applies the philosophy of Occam's razor. This philosophical principle (prefer simple explanations when explanations predict equally well) says that you should choose the simplest explanation that is adequate.

Mixed stepwise regression is still a useful tool, as long as we are aware of the risks, particularly of wrongly rejecting the null, so we should reduce the apparent significance level if we have a large number of tests.

Model Selection Criteria

For all linear models, common criteria are R^2, Adjusted R^2 and the F-test.

R^2 without adjustment is a very bad criterion because it will always choose the most complex model. Adjusted R^2 is better, but still tends to choose overly complex models.

The F-test is equivalent to the Log Likelihood ratio test for linear models, i.e. looking at the 2 x the difference in the maximum Log Likelihood of the 2 models being compared. However, this criterion has also been criticized for choosing models that are too complex.

Minimizing AIC or BIC are more sophisticated criteria, which are more trustworthy than repeated significance tests (the detailed statistical argument is too complex to explain here)

which means choosing s and p to be small together. Note that these criteria need adjustment if we consider models with different transformations, while adjusted R^2 can still be used.

The objective is to choose the model with lowest AIC or BIC, which both choose the highest Log Likelihood after a penalty for the number of parameters (p=number of coefficients, including the intercept). AIC is based on information theory arguments, while BIC relies on Bayesian analysis arguments, although arguably BIC is inconsistent with Bayesian inference as it takes no account of any prior information.

$AIC = 2p - 2 \text{Log}_e(\text{Lik})$, AIC has a variant called AICc which corrects for finite sample sizes.
$AICc = AIC + 2p(p+1)/(n-p-1)$ (the extra term is small if n is large compared to p^2)
$BIC = p \text{Log}_e(n) - 2 \text{Log}_e(\text{Lik})$

For linear models with independent Gaussian errors, $\text{Log}_e(\text{Lik})$ can be written as $-n/2 \text{Log}_e(\text{Residual SS})$ if we ignore constants that cancel for comparisons, so we can write

$AIC = n \text{Log}_e(\text{Residual SS}) + 2p$
$BIC = n \text{Log}_e(\text{Residual SS}) + p \text{Log}_e(n)$

BIC penalizes complexity more than AIC in all practical situations, as log(n) is usually much more than 2.

These criteria are not restricted to linear models. The only real limitation is that we need to be able to calculate the log likelihood and the number of parameters.

These criteria do not involve a statistical test, but generally differences of less than 2 in AIC or BIC are considered to be meaningless (i.e. we cannot distinguish between the models). All of these criteria except (adjusted) R2 assume that we have the same dependent variable values. AIC and BIC do not require that the models are nested (i.e. all models are special cases of the most complex model), unlike the F-test.

Problems with default settings

Note that stepwise regression does not always yield sensible answers using the default settings. We show here the results when we try to predict the oxygen uptake in the previous example from 30 randomly generated variables (so they have NO relationship with the dependent variable). This is an extreme example as the number of variables is almost as large as the sample size.

Response O2Uptake

Summary of Fit

RSquare	0.902106
RSquare Adj	0.827246
Root Mean Square Error	2.214191
Mean of Response	47.37581
Observations (or Sum Wgts)	31

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	76.194708	4.138782	18.41	<.0001*
X2	-13.74505	2.145383	-6.41	<.0001*
X3	-4.442926	1.66686	-2.67	0.0163*
X4	8.7827771	1.863743	4.71	0.0002*
X7	-12.1338	2.314868	-5.24	<.0001*
X8	5.4224362	1.740941	3.11	0.0063*
X9	4.3865304	1.844954	2.38	0.0294*
X13	-1.884296	1.454073	-1.30	0.2123
X14	-22.8847	2.628407	-8.71	<.0001*
X15	-11.4713	2.279956	-5.03	0.0001*
X20	4.8826124	1.7179	2.84	0.0113*
X23	-12.46203	2.153479	-5.79	<.0001*
X24	9.2972471	2.219653	4.19	0.0006*
X25	-9.754634	2.300322	-4.24	0.0006*

This happens because the default p-value for adding/dropping a variable is very high (0.25/0.1). If we reduce this down to 0.01/0.01 we get a null model as the most significant variable has a p-value of about 0.03, which is about 1/30 where 30 is the number of variables, so this is what we would expect by chance if no variables are associated with oxygen update.

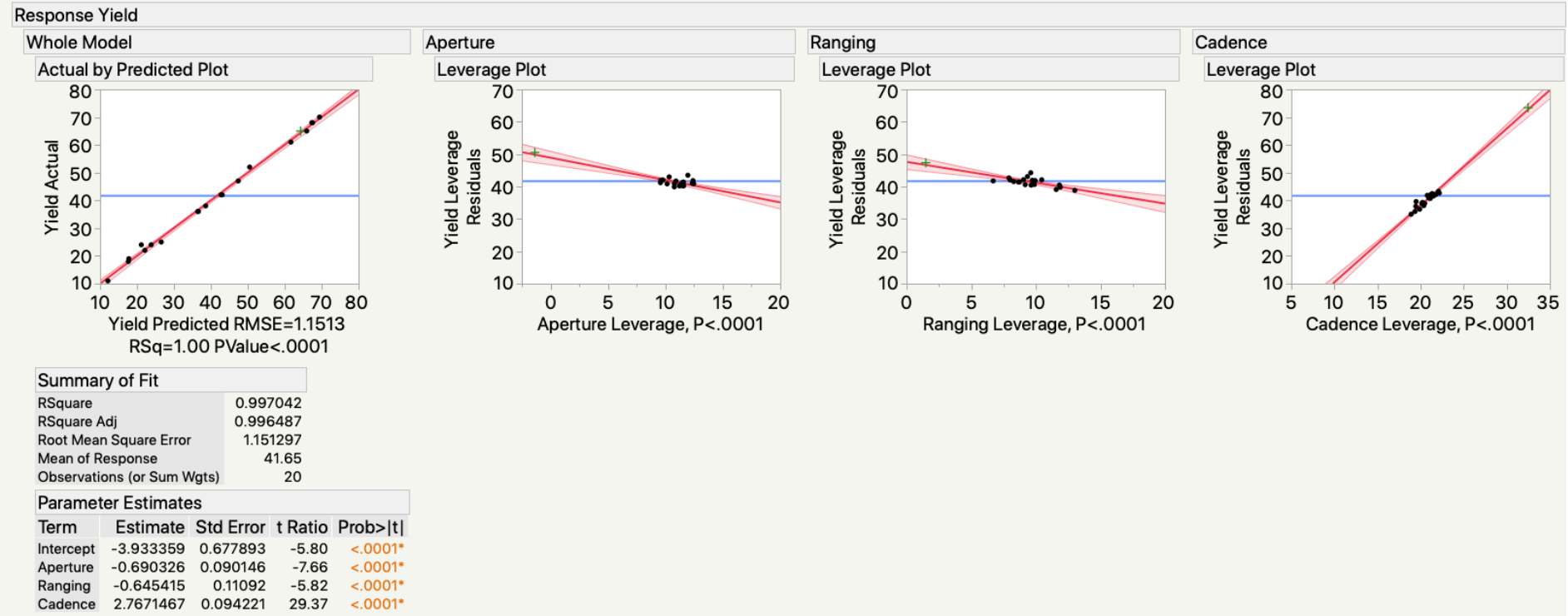
Diagnostics

The 2 core diagnostic tools are the residuals and influence or leverage. The residuals tell us what the fitting error is for each point. Influence or leverage values tell us how much the model depends on each point. They actually depend on the H hat matrix we mentioned before. If a diagonal coefficient of H is large, then this data point has a large influence on the fitted model. This is because the diagonal element of H is the coefficient of the value of y_i in estimating the predictor for y_i:

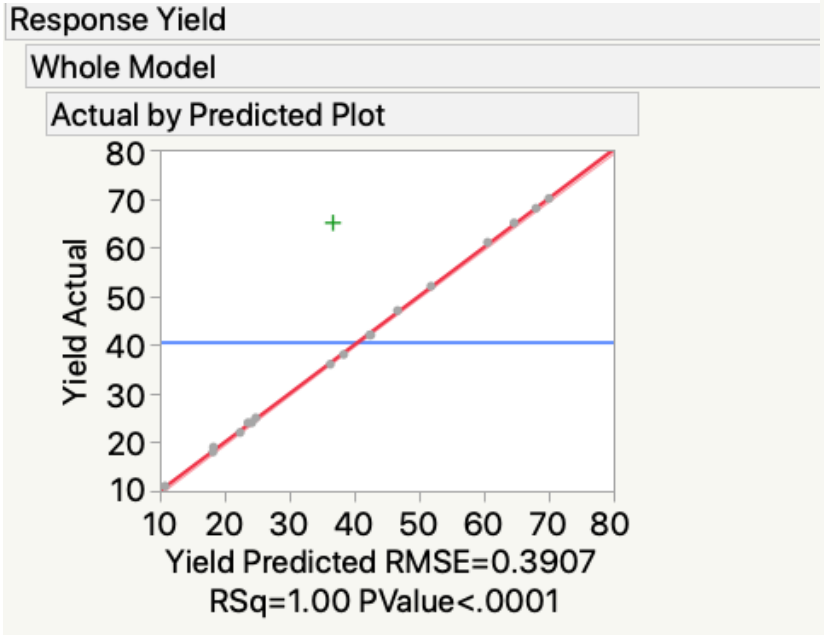
Yfitted=H times Y, so $Yfitted_i = H_{ii} \times Y_i$ + other terms

The residuals can be used in different ways. They can be plotted in sequence to see if there is a pattern in the sequence. They can be plotted against the fitted values to see if there is some non-linearity in the overall responses or they can be plotted against each independent variable to see if there is non-linearity for each variable. We can also look at the distribution of the residuals to see if they look like they follow a Normal distribution (bell-shaped curve) (see later).

Leverage plots show the influence of data points on the model, either overall, or in terms of each independent variable. For example, this dataset has 3 predictors (Aperture, Ranging and Cadence), which seem very good as the adjusted R^2 is over 99%. However, the point marked with a + is highly influential, as can be seen in the plots on the right hand side. Note that the coefficient for Cadence is positive, while for Aperture and Ranging they are both negative.



The high leverage indicates that if we now remove that one point from the fitting of our model, there may be an important change in the fitted model.



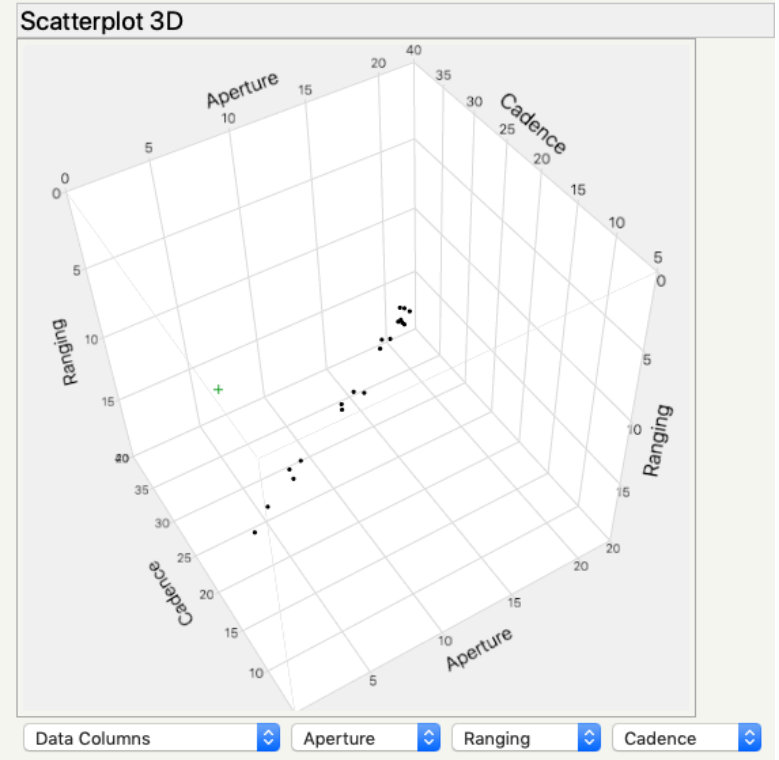
Summary of Fit

RSquare	0.999653
RSquare Adj	0.999583
Root Mean Square Error	0.390711
Mean of Response	40.42105
Observations (or Sum Wgts)	19

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.591574	0.378206	-1.56	0.1386
Aperture	1.3672357	0.187345	7.30	<.0001*
Ranging	1.392623	0.186906	7.45	<.0001*
Cadence	0.639497	0.193782	3.30	0.0049*

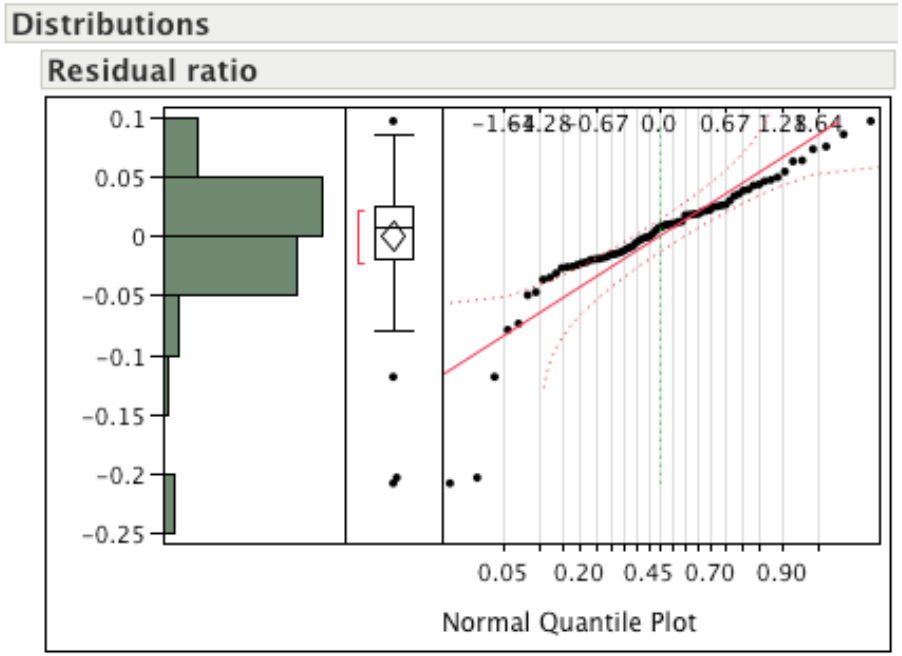
Without that point, the adjusted R^2 is over 99.9% and the parameter estimates for Aperture and Ranging change sign! How can that be? The 3D plot below of the 3 covariates, show that that point contains information separate from all other data points, as all the other data points are in a single plane.



Normality Tests

The statistical inference we do for linear models assumes that the errors independently follow a Normal distribution. Note: unlike the distribution of the sample mean when testing the population mean, a large sample does NOT make this assumption correct! There are several different formal tests of Normality, but one useful graphical method plots residuals in order of increasing value against Normal quantiles (i.e. plot residuals against their expected values if

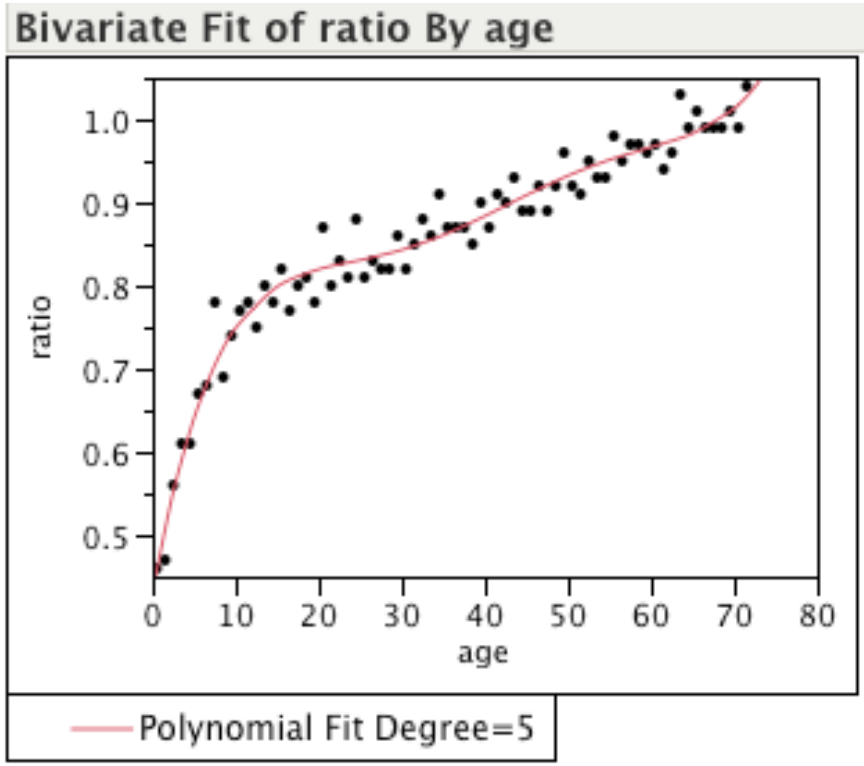
they come from a Normal distribution). If the distribution of the residuals is Normal, the plot should be roughly linear – JMP provides confidence bounds to allow an informal test of Normality. The growth example shows a failure below.



Polynomial Regression

A special type of multiple regression is polynomial regression. In this case, the question is what order of polynomial in X do we use to fit the model. Only in special situations would we use high order polynomials, because the models usually give very extreme predictions outside (or even inside) the data range. This is one reason why we would often prefer looking at

transformations instead, particularly if we can interpret the transformation sensibly. The growth example below shows that a fifth order polynomial is statistically significant but is very hard to interpret and yields an adjusted R2 value that is not that much better than the simple log-log transformation (or as we will see, rank transformation).



Polynomial Fit Degree=5

ratio = 0.7087453 + 0.0043409*age +
9.8329e-5*(age-36)^2 - 3.5057e-6*(age-36)^3 -
1.547e-7*(age-36)^4 + 5.0901e-9*(age-36)^5

Summary of Fit

RSquare	0.962915
RSquare Adj	0.960106
Root Mean Square Error	0.024317
Mean of Response	0.855556
Observations (or Sum Wgts)	72

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	1.0133508	0.202670	342.7432
Error	66	0.0390270	0.000591	Prob > F
C. Total	71	1.0523778		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.7087453	0.02241	31.63	<.0001*
age	0.0043409	0.000604	7.18	<.0001*
(age-36)^2	9.8329e-5	0.000026	3.78	0.0003*
(age-36)^3	-3.506e-6	1.835e-6	-1.91	0.0605
(age-36)^4	-1.547e-7	2.246e-8	-6.89	<.0001*
(age-36)^5	5.0901e-9	1.244e-9	4.09	0.0001*

Transformations

The most common transformation is the power transform

$z=(y^a-1)/a$, where $a=0$ is the special limiting case of $z=\log(y)$

log makes special sense when dealing with product or ratio variables, because it turns them into linear relationships:

$\text{Log}(y1 \times y2) = \text{Log}(y1) + \text{Log}(y2)$

$\text{Log}(y1/y2) = \text{Log}(y1) - \text{Log}(y2)$

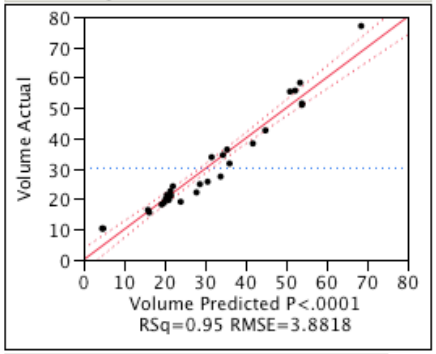
We already looked at this for the growth example, which explicitly involves ratios.

Note that transformations may simplify the model, e.g. if weight is proportional to some power of height, then log(weight) is linear in log(height). It is important to consider models that 'make sense', which often means they have a sound theoretical basis. Another nice example is predicting the volume of trees given the diameter and height. We compare regression models using the raw data and log transformed. The log model not only fits better (compare R2), but also has a very simple interpretation: V is proportional to D^2H, which makes sense if we think of the formula for volume of a cone or cylinder.

Response Volume

Whole Model

Actual by Predicted Plot



Summary of Fit

RSquare	0.94795
RSquare Adj	0.944232
Root Mean Square Error	3.881832
Mean of Response	30.17097
Observations (or Sum Wgts)	31

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	26	421.71636	16.2199	158.2425
Pure Error	2	0.20500	0.1025	Prob > F
Total Error	28	421.92136		0.0063*

Max RSq
1.0000

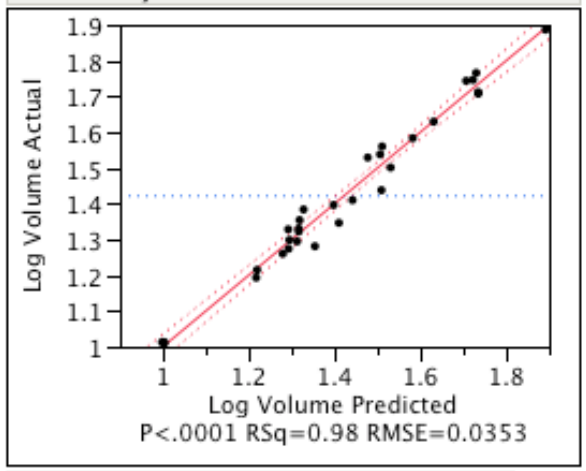
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-57.98766	8.638226	-6.71	<.0001*
Diam	4.7081605	0.264265	17.82	<.0001*
Height	0.3392512	0.130151	2.61	0.0145*

Response Log Volume

Whole Model

Actual by Predicted Plot



Summary of Fit

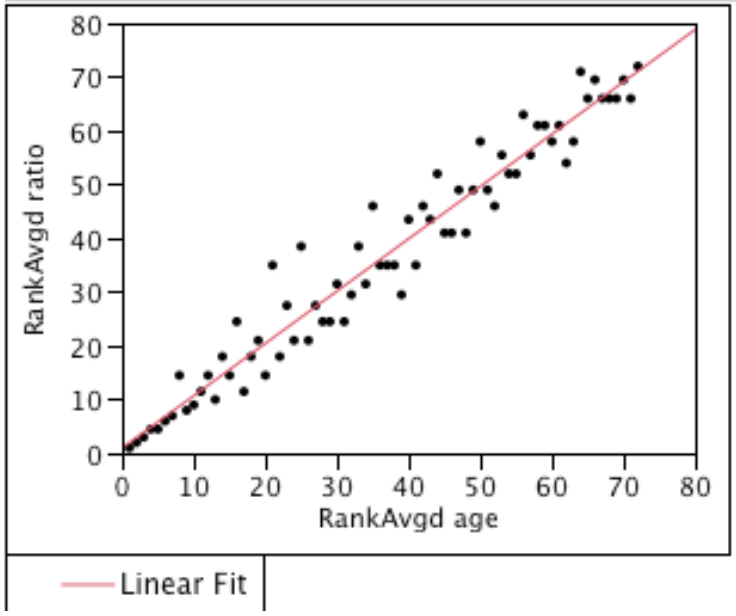
RSquare	0.977678
RSquare Adj	0.976084
Root Mean Square Error	0.035346
Mean of Response	1.421329
Observations (or Sum Wgts)	31

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2.880075	0.347344	-8.29	<.0001*
Log Diameter	1.9826499	0.075011	26.43	<.0001*
Log Height	1.1171233	0.204437	5.46	<.0001*

The other very useful transformation is ranking, i.e., we replace the data by their ranks 1 to n. Replacing the data by their ranks is usually equivalent to what we call a non-parametric procedure. This is usually a less powerful procedure, but it is based on weaker assumptions, so it is more robust. The tricky part is how to handle ties (values with the same rank). "Proper" non-parametric procedures account for the fact that ties reduce the variability and reflect less information.

We can revisit the growth example, using ranks. The simple linear model using ranks means that the ratio increases (but perhaps at a non-constant rate) with age.



Linear Fit

RankAvgd ratio = 0.9753521 + 0.973278*RankAvgd age

Summary of Fit

RSquare	0.948811
RSquare Adj	0.94808
Root Mean Square Error	4.764893
Mean of Response	36.5
Observations (or Sum Wgts)	72

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	29458.206	29458.2	1297.478
Error	70	1589.294	22.7	Prob > F
C. Total	71	31047.500		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.9753521	1.134898	0.86	0.3930
RankAvgd age	0.973278	0.02702	36.02	<.0001*

12: Analysis of Variance for categories (Factors or Groups)

Learning objectives: understand how to model effect of categorical variables on a continuous variable

What if we have more than two populations, how can we test if they have the same population mean? Analysis of Variance is also a way to simultaneously test for any difference in the means of a set of populations, assuming that they have the same population variance. There is no simple Confidence Interval idea any more, but we can still do a significance test. This time, the thing we use to test for a difference (the test statistic) is a bit more complex. The idea is as follows:

If the populations have the same mean, the sample means should be similar and hence the sample means should be close to the overall mean. We look at the ratio of the squared deviations of the sample means from the overall mean to the deviations of the individual data points from their respective sample means. If the ratio is large, then we do not believe that the sample means are similar and hence that the population means can be the same.

This should sound familiar. The idea is exactly the same ANOVA idea we used for multiple regression. For a one-way classification (one-way ANOVA), the table looks like:

Source of Variation	SS	d.o.f.	Mean Square	F Ratio
Between Groups	SS_B	k-1	$MS_B=SS_B/(k-1)$	$F=MS_B/MS_W$
Within Groups	SSw	n-k	$MSw=SS_W/(n-k)$	
Total	SS_T	n-1		

where $F=MS_B/MS_W$ and k is the number of groups. The Within Groups part is the same idea as the Residual for the multiple regression model, i.e. the part that our model does not "explain"

The precise formulae are not important, but the relative simplicity is attractive and has lead to the extension of this idea to much more complex situations, in particular where there are multiple simultaneous groupings.

For example, if there are 2 classifications, the sources become Factor 1, Factor 2, Residual and Total. This makes it relatively easy to identify the important sources (causes) of variability. If the groups are very different, then SS_B will be large. If the difference is large relative to the number of groups, MS_B will be large. If the variability left is small, then SS_W will be small. If SS_W is small given the sample size, then MS_W will be small.
Overall then, we can see why large values of F indicate that there is evidence of group differences, i.e. we reject the null hypothesis of no group differences.

Statistical software tells us whether the F value is big enough to reject the null hypothesis. Usually, it tells us the significance level at which we should reject the null - this is called the observed significance level. If there are only 2 groups, we can use the T-test, which is

equivalent to F-test in this case (k=2, means k-1=1) with the F statistic equal to the square of the t statistic.

Again, we sometimes use R^2 (R squared) as a summary of the usefulness of the grouping, where

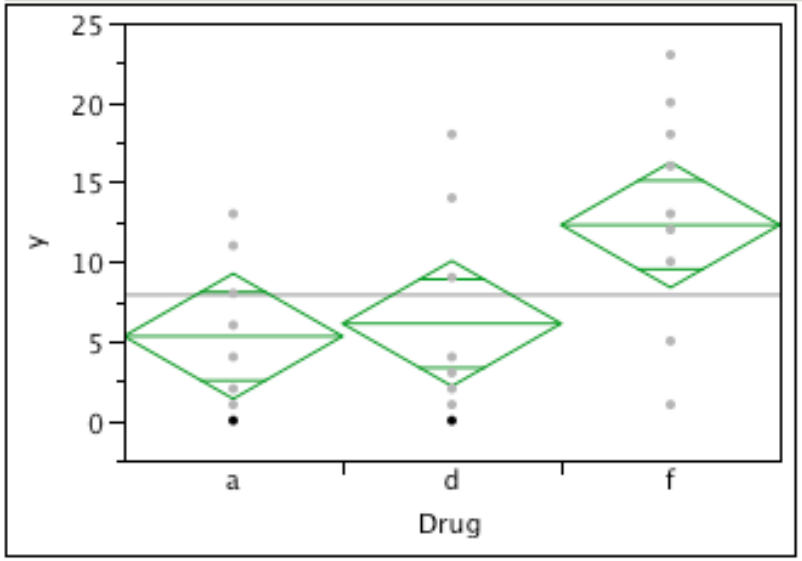
$R^2=SS_B/SS_T$, i.e., the proportion of variability 'explained' by the grouping.

Note that R^2 is scale independent, which allows easy comparisons. Again, there is a problem with R^2 in that adding an extra group will always increase it. As in the multiple regression case, we use adjusted R^2, allowing comparison of models with different numbers of groups.

$R^2 = 1-SS_W/SS_T = 1-(SS_W/(n-1))/(SS_T/(n-1))$, while Adjusted $R^2 = 1-(SS_W/(n-k))/(SS_T/(n-1))$

General strategy is to find model that is simple (few terms), good explanation (high R^2) and easy to interpret (theoretical basis and/or consistent inclusion pattern of variables and interactions). The example below illustrates a simple randomized controlled trial where f is a placebo and d,f are the active drugs.

Oneway Analysis of y By Drug



Oneway Anova

Summary of Fit

Rsquare	0.227826
Adj Rsquare	0.170628
Root Mean Square Error	6.070878
Mean of Response	7.9
Observations (or Sum Wgts)	30

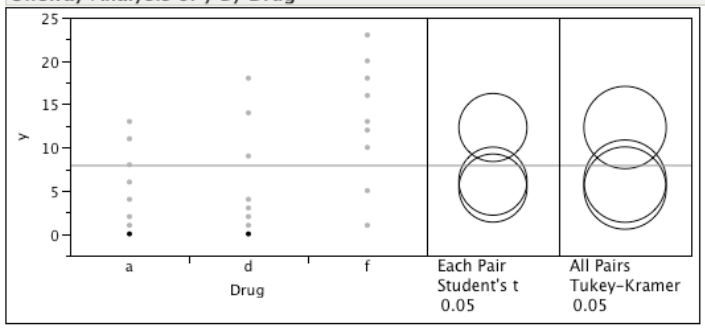
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Drug	2	293.6000	146.800	3.9831	0.0305*
Error	27	995.1000	36.856		
C. Total	29	1288.7000			

Multiple Comparisons

What if we want to compare the individual groups? For example, we might want to compare pairs of groups and see if they are significantly different. The simple approach is just to use a t-test for those two groups, but the problem is like the one for multiple regression, where there are many possible tests. One approach is to only look at pairs if the overall F test is significant. This approach is quite simple, particularly if the numbers in each group are the same (we call this a balanced design). The other better, but more complex, approach is to use a modification of the t test that takes into account the multiple comparisons being done (e.g. the Tukey-Kramer Honestly Significant Difference (HSD)). The drug example illustrates that the result is less significant once we take the multiple comparisons into account:

Oneway Analysis of y By Drug



Means Comparisons

Comparisons for each pair using Student's t

Confidence Quantile

t	Alpha
2.05183	0.05

LSD Threshold Matrix

Abs(Dif)-LSD

	f	d	a
f	-5.5707	0.6293	1.4293
d	0.6293	-5.5707	-4.7707
a	1.4293	-4.7707	-5.5707

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level		Mean
f	A	12.300000
d	B	6.100000
a	B	5.300000

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
f	a	7.000000	2.714979	1.42932	12.57068	0.0157*
f	d	6.200000	2.714979	0.62932	11.77068	0.0305*
d	a	0.800000	2.714979	-4.77068	6.37068	0.7705

Comparisons for all pairs using Tukey-Kramer HSD

Confidence Quantile

q*	Alpha
2.47942	0.05

LSD Threshold Matrix

Abs(Dif)-HSD

	f	d	a
f	-6.7316	-0.5316	0.2684
d	-0.5316	-6.7316	-5.9316
a	0.2684	-5.9316	-6.7316

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level		Mean
f	A	12.300000
d	A B	6.100000
a	B	5.300000

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
f	a	7.000000	2.714979	0.26843	13.73157	0.0403*
f	d	6.200000	2.714979	-0.53157	12.93157	0.0754
d	a	0.800000	2.714979	-5.93157	7.53157	0.9533

Contrasts

Contrasts are linear combinations of group effects – the simplest example would be the difference between any pair of groups. A more complex example is where we have 2 treatments and 1 control group and we want to test whether the treatment groups are different from the control group. In other words, this is a situation where we are testing for zero differences in specific known combinations of the groups. The drug example below tests whether the average effect of a and d is better than placebo, even if we are unclear as to which drug is better.

Response y

Whole Model

Drug

Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
a	5.300000	1.9197801	5.3000
d	6.100000	1.9197801	6.1000
f	12.300000	1.9197801	12.3000

Contrast

SS	NumDF	DenDF	F Ratio	Prob > F
290.4	1	27	7.8794	0.0092*

Nested vs. Crossed

What if there is more than one classification, e.g. Sex and Marital Status? In general, if there is more than one classification, it can work two ways: nested (hierarchical) and crossed. Nested means that the second classification works inside (nested) the first classification, whereas crossed means that both classifications are generally applicable. Sex and Marital Status will usually be crossed, whereas for School and Class Number, Class Number usually is a within school classification, i.e. it is nested inside School.

For two crossed classifications (2 way ANOVA), we have 4 possible sources of variability:

Classification 1
Classification 2
Class 1 x Class 2
Residual

where Class 1 x Class 2 (sometimes denoted Class 1.Class 2) is the interaction - this effect means that the Class 2 differences are not the same for different values of Class 1. In our example, it might mean that the differences between single and married people may be different for males and females. If this is not the case, we would say that Sex and Marital Status have independent effects and merge this source back into the Residual.

Lack of Fit & Replicates

For models of this sort, we usually check for lack of fit. This means we hope to not reject this hypothesis. This analysis is possible if we have what we call replicates. This means that for some combinations of the factors, we have multiple data points, giving us a way to estimate the "pure error" (i.e. differences between respondents with identical factor combinations) which excludes any possibility that our model does not include necessary interaction terms between the factors. We divide up the residual error into "pure error" and "lack of fit" to check whether our model looks OK (assuming we have not completely ignored important factors). Replicates are a very good idea when designing an experiment as they allow a diagnostic check that does not depend on whether the model is correct.

We look at an example of making popcorn and looking at which factors of popcorn type, oil amount and batch size and their interactions predict high yield. The results show that popcorn type, batch size and their interaction seem to matter.

Response yield

Summary of Fit

RSquare	0.892456
RSquare Adj	0.798355
Root Mean Square Error	1.417745
Mean of Response	10.75
Observations (or Sum Wgts)	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	133.44000	19.0629	9.4840
Error	8	16.08000	2.0100	Prob > F
C. Total	15	149.52000		0.0025*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10.75	0.354436	30.33	<.0001*
popcorn[gourmet]	1.475	0.354436	4.16	0.0032*
oil amt[little]	-0.525	0.354436	-1.48	0.1768
popcorn[gourmet]*oil amt[little]	-0.5	0.354436	-1.41	0.1960
batch[large]	-1.75	0.354436	-4.94	0.0011*
popcorn[gourmet]*batch[large]	-1.525	0.354436	-4.30	0.0026*
oil amt[little]*batch[large]	-0.025	0.354436	-0.07	0.9455
popcorn[gourmet]*oil amt[little]*batch[large]	0.5	0.354436	1.41	0.1960

However, it turns out that it is specifically the combination of small batches of gourmet popcorn that performs better, so we have a good simple explanation:

Response yield

Whole Model

Summary of Fit

RSquare	0.804798
RSquare Adj	0.790855
Root Mean Square Error	1.44387
Mean of Response	10.75
Observations (or Sum Wgts)	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	120.33333	120.333	57.7204
Error	14	29.18667	2.085	Prob > F
C. Total	15	149.52000		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	12.333333	0.416809	29.59	<.0001*
Small Gourmet[n]	-3.166667	0.416809	-7.60	<.0001*

Nested

For two nested classifications, we have 3 sources:

Classification 1
Classification 2 (inside Classification 1)
Residual

We would normally check whether the Class 2 differences (within Class 1) are important. In our example, this would mean checking whether classes within schools are different. The nested situation may be more complex, because Class 2 can be a sample from Class 1. In this case we say that Class 2 is a random effect, which alters the error term in our F test (too complex to explain fully here) and makes the correct statistical analysis more complex, because we need to adjust the denominator in our F tests to account for the random effects part.

Model choice objective

There is no limit to the number of classifications that we can consider in theory, although in practice, it only works well if the sample size is much bigger than the total number of groups across all classifications. We aim to find groupings such that individuals within the groups are similar, while the groups are very different. This helps us to understand what characteristics relate to our measured outcome variable. Remember that this does not prove causation, but suggests a possible causation, in other words, we use "variability explained" in a rather loose manner that does not imply causation. Also, there are often alternative explanations, so that

we can have several different groupings, each of which are useful, but where once we have used one of the groupings, the other groupings provide no further explanatory power.

Assumptions

The analysis makes 3 important assumptions:

1) Independent errors for different data points
2) The variability within groups is the same (across different values of X for regression models)
3) errors follow a Normal (bell-shaped) distribution.

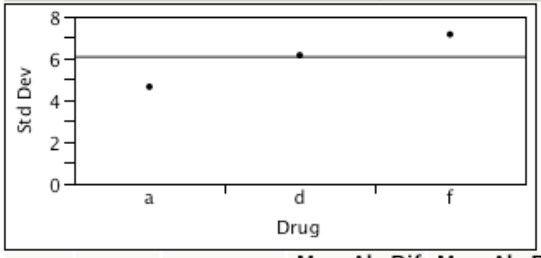
In practice, ANOVA is quite robust (not too sensitive to these assumptions), but sometimes we can reduce divergence from these assumptions by transforming the data.

Transformations again

For one-way ANOVA, we can remove the assumption of equal variance using Welch's ANOVA.

Oneway Analysis of y By Drug

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
a	10	4.643993	3.900000	3.900000
d	10	6.154492	5.120000	4.700000
f	10	7.149981	5.700000	5.700000

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	1.1395	2	27	0.3349
Brown-Forsythe	0.5998	2	27	0.5561
Levene	0.8904	2	27	0.4222
Bartlett	0.7774	2	.	0.4596

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
3.3942	2	17.406	0.0569

Random effects

If we have random effects, such as when we have repeated measurements or hierarchical effects (see student example above or animals example below), we need to use ANOVA methods that account for the random effects, which generally means that the denominator needs changing for the F-tests. These methods may be labelled as random effects ANOVA or repeated measures ANOVA.

The animals example is about trying to understand the distances that animals travel in different seasons. We have 2 animal species (Species, i.e. Fox and Coyote), 3 animals for each species (Subjects 1,2,3) and 4 seasons of one year (Season), so we have 24 data points (2 x 3 x 4). However, the animals within each species are clearly random samples within species and we have 4 measurements for each individual, so we have both repeated measurements for each animal and a hierarchy of animals within species, so we need to modify our ANOVA to account for both of these. In JMP we do this by noting that Subject is nested within Species and by noting that Subject is a random effect. We also include Season as a Factor. We then get the output below, which shows that Species and Season are both statistically significant, but that the F test has been modified to account for the fact that Species is tested relative to variation between subjects, while Season is tested relative to variation within subjects. We can see that Coyotes travel further than Foxes and animals travel furthest in Spring and least in Winter. We are implicitly assuming that we can ignore any interaction between Species & Season.

Response miles

Summary of Fit

RSquare	0.823497
RSquare Adj	0.786338
Root Mean Square Error	1.219062
Mean of Response	4.458333
Observations (or Sum Wgts)	24

Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	4.4583333	0.42287	4	10.54	0.0005*
species[COYOTE]	1.4583333	0.42287	4	3.45	0.0261*
season[fall]	-0.625	0.431003	15	-1.45	0.1676
season[spring]	1.7083333	0.431003	15	3.96	0.0012*
season[summer]	0.875	0.431003	15	2.03	0.0605

REML Variance Component Estimates

Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Wald p-Value	Pct of Total
subject[species]	0.4719626	0.7013889	0.7707006	-0.809157	2.2119344	0.3628	32.063
Residual		1.4861111	0.5426511	0.8109483	3.5597535		67.937
Total		2.1875	0.8609382	1.1475492	5.700522		100.000

-2 LogLikelihood = 78.806486054

Note: Total is the sum of the positive variance components.

Total including negative estimates = 2.1875

Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
species	1	1	4	11.8932	0.0261*
season	3	3	15	10.6449	0.0005*

Effect Details

species

Least Squares Means Table

Level	Least Sq Mean	Std Error
COYOTE	5.9166667	0.59802917
FOX	3.0000000	0.59802917

season

Least Squares Means Table

Level	Least Sq Mean	Std Error
fall	3.8333333	0.60380736
spring	6.1666667	0.60380736
summer	5.3333333	0.60380736
winter	2.5000000	0.60380736

13: General Linear Model

Learning objectives: understand how to model effect of categorical and continuous variables on a continuous variable

Can integrate multiple regression with ANOVA for factors to get what is called General Linear Model or Analysis of Covariance (ANACOVA). This allows investigation of which factors (groups) and covariates (linear relations) are useful in 'explaining' the variability in the data. For individual covariates, can use T-test, but this is statistically equivalent to F-test. The T-test is a test of a null hypothesis that the slope coefficient is zero. We can also allow for interactions between Factors and Covariates (i.e. different slopes for different groups). Interactions between Covariates are cross products of covariates.

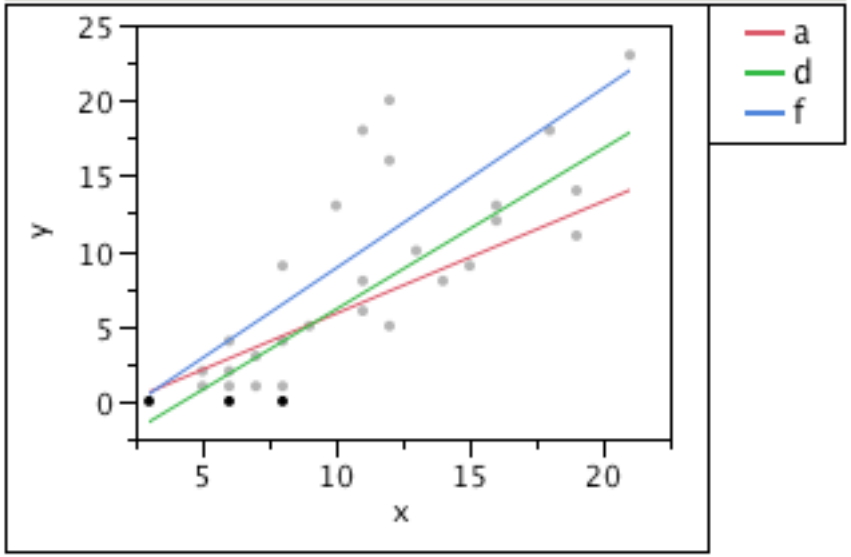
In all cases we can use diagnostic checks of the assumptions:
1) save residuals and test if they are Normal (NOT whether the dependent variable is Normal)
2) plot residuals against the predictors and look for a pattern in the center
3) plot residuals against the predictors and look for pattern in the spread
4) plot leverage for each predictor to see if certain data points are very influential

We illustrate GLM using the drug data with a covariate, x which is the baseline measurement of y (i.e. before taking the drug). This shows that there is no significant drug effect, once we control for the baseline measurement. There is also no evidence of an interaction between the drug and the baseline measurement.

Response y

Whole Model

Regression Plot



Summary of Fit

RSquare	0.691505
RSquare Adj	0.627235
Root Mean Square Error	4.070002
Mean of Response	7.9
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	891.1420	178.228	10.7594
Error	24	397.5580	16.565	Prob > F
C. Total	29	1288.7000		<.0001*

Effect Tests

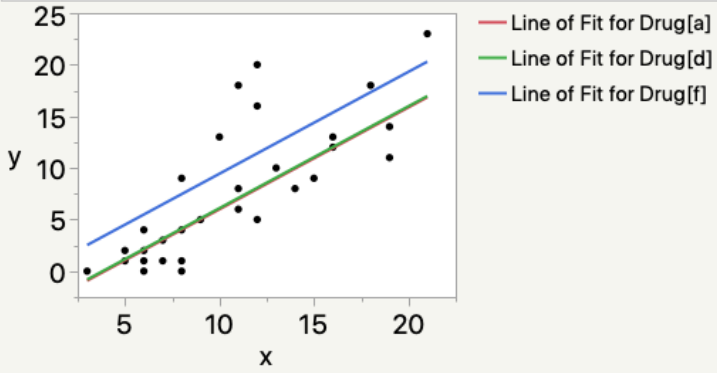
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Drug	2	2	52.05637	1.5713	0.2284
x	1	1	564.56753	34.0821	<.0001*
Drug*x	2	2	19.64465	0.5930	0.5606

Given that the interaction is not significant, we look at the results without the interaction:

Response y

Whole Model

Regression Plot



Summary of Fit

RSquare	0.676261
RSquare Adj	0.638906
Root Mean Square Error	4.005778
Mean of Response	7.9
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	871.4974	290.499	18.1039
Error	26	417.2026	16.046	Prob > F
C. Total	29	1288.7000		<.0001*

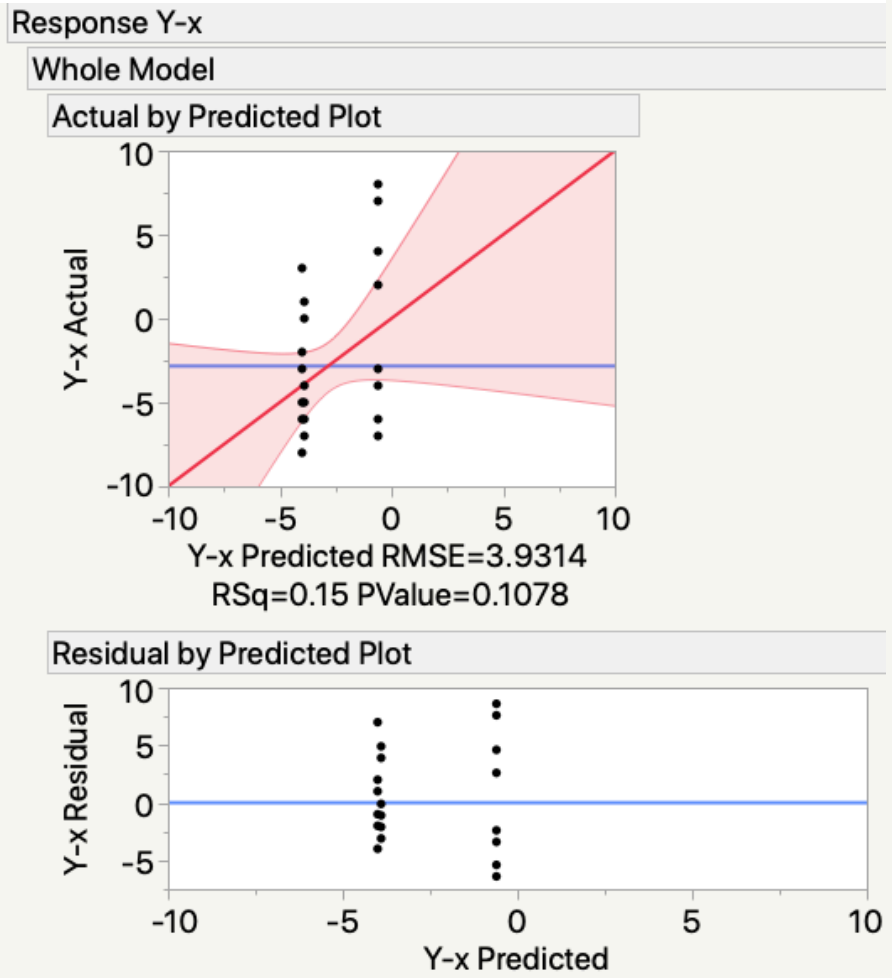
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2.695773	1.911085	-1.41	0.1702
x	0.9871838	0.164498	6.00	<.0001*
Drug[a]	-1.185037	1.060822	-1.12	0.2742
Drug[d]	-1.076065	1.041298	-1.03	0.3109

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
x	1	1	577.89740	36.0145	<.0001*
Drug	2	2	68.55371	2.1361	0.1384

This shows clearly how significant the baseline measurement is, and that the Drug effect is no longer significant, after controlling for baseline. Note that the coefficient of the baseline is close to 1, suggesting we could model the change in outcome, i.e. Y-X. We look at this below:



Summary of Fit

RSquare	0.152116
RSquare Adj	0.08931
Root Mean Square Error	3.931355
Mean of Response	-2.83333
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	74.86667	37.4333	2.4220
Error	27	417.30000	15.4556	Prob > F
C. Total	29	492.16667		0.1078

Parameter Estimates

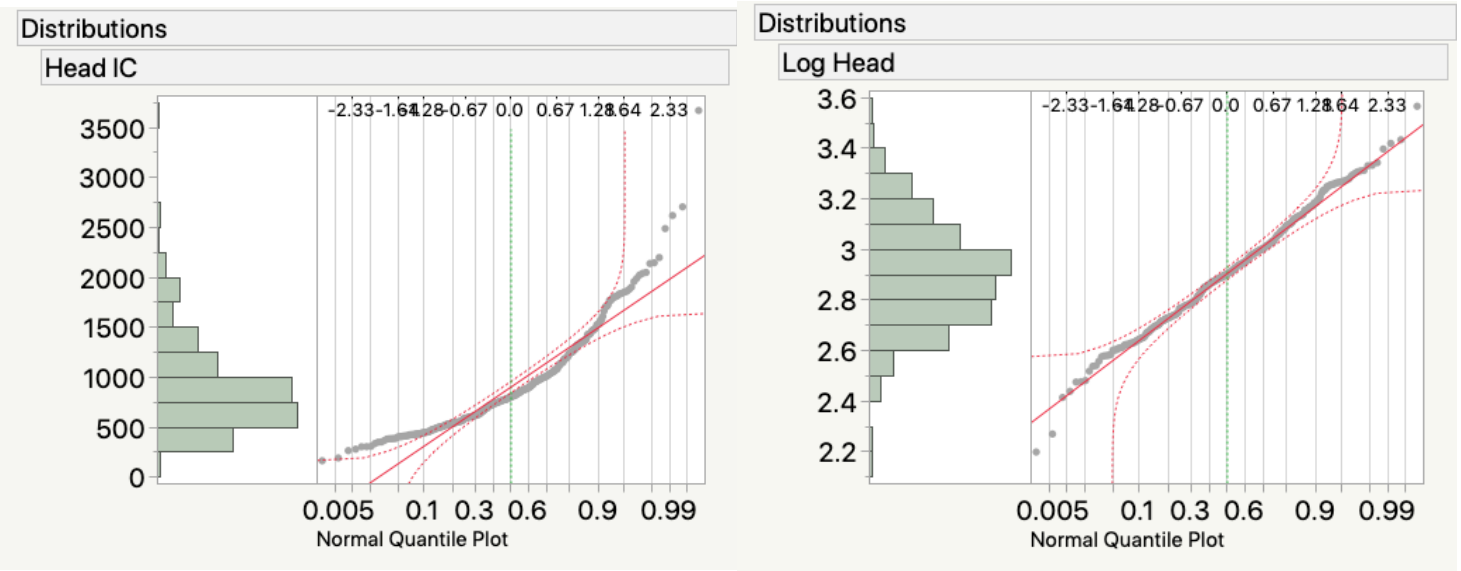
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2.833333	0.717764	-3.95	0.0005*
Drug[a]	-1.166667	1.015072	-1.15	0.2605
Drug[d]	-1.066667	1.015072	-1.05	0.3027

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Drug	2	2	74.866667	2.4220	0.1078

This still shows no significant Drug effect here, although the (a+d)/2-f contrast (average of drugs vs placebo) yields 3.7% observed significance level. The adjusted R2 is only 9%.

We now at a more complex example. This is data on safety testing for cars, where they measure the force on the head of dummies that are inside cars when they are crashed. If we look at the distribution of the force (which must be greater than zero), we can see that the distribution is very asymmetrical. While it is the residuals that must follow a Normal distribution, rather than the original variables, having a dependent variable that has a constraint like this means that the residuals cannot follow a Normal distribution, so we compare the distribution after a log transformation. Log Transformation also makes scientific sense, because we know that force has a multiplicative relationship with acceleration, which is probably what matters in terms of damage.



In this dataset, we have many potential predictors, but here we will consider three factors D/P (Driver or Passenger position of the dummy), Protection (Manual, Passive, Motorized belts, Driver only, Driver & Passenger airbags), VType (vehicle size/type, such as MPV, compact etc.) and one covariate, Log Wt (log transformed weight, using the same logic). We will look at 2 relatively simple models, one about the protection (with D/P, Protection and the interaction) and one about the vehicle characteristics (VType, Log Wt and the interaction). Both models are statistically significant, with AIC and BIC preferring the first model.

Response Log Head

Whole Model

Summary of Fit

RSquare	0.218938
RSquare Adj	0.197572
Root Mean Square Error	0.186446
Mean of Response	2.905053
Observations (or Sum Wgts)	339

AICc	BIC
-164.082	-122.803

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	9	3.205802	0.356200	10.2468
Error	329	11.436704	0.034762	Prob > F
C. Total	338	14.642507		<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
D/P	1	1	0.1086906	3.1267	0.0779
Protection	4	4	1.9671283	14.1471	<.0001*
D/P*Protection	4	4	0.4205689	3.0246	0.0180*

Response Log Head

Whole Model

Summary of Fit

RSquare	0.217884
RSquare Adj	0.181563
Root Mean Square Error	0.188296
Mean of Response	2.905053
Observations (or Sum Wgts)	339

AICc	BIC
-150.525	-87.3899

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	15	3.190368	0.212691	5.9988
Error	323	11.452139	0.035456	Prob > F
C. Total	338	14.642507		<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Log Wt	1	1	0.05063902	1.4282	0.2329
VType	7	7	0.97859079	3.9429	0.0004*
Log Wt*VType	7	7	0.79394436	3.1990	0.0027*

We then consider a more complex model that includes the factors and covariates from both models. Interestingly, AIC prefers this model, while BIC prefers the first model, because it uses a larger penalty for parameters than AIC. If we look at the Least Squares Means Table for D/P and Protection, we can see clearly that, except for the anomalous case of driver only airbags, passengers are always safer than drivers for the same protection and also that airbags are clearly safer than belts. The Residual Normal Quantile Plot shows that the residuals do look consistent with our assumed Normal distribution.

Response Log Head

Whole Model

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	303	9.3571862	0.030882	0.7470
Pure Error	11	0.4547396	0.041340	Prob > F
Total Error	314	9.8119258		0.7980
				Max RSq
				0.9689

Summary of Fit

RSquare	0.329901
RSquare Adj	0.278683
Root Mean Square Error	0.176772
Mean of Response	2.905053
Observations (or Sum Wgts)	339

AICc	BIC
-182.334	-87.3578

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	24	4.830581	0.201274	6.4412
Error	314	9.811926	0.031248	Prob > F
C. Total	338	14.642507		<.0001*

Effect Tests

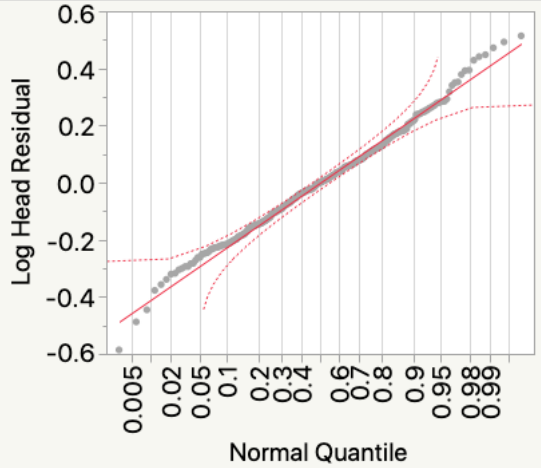
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
D/P	1	1	0.10626179	3.4006	0.0661
Protection	4	4	0.46823009	3.7461	0.0054*
D/P*Protection	4	4	0.41554684	3.3246	0.0110*
Log Wt	1	1	0.02553240	0.8171	0.3667
VType	7	7	0.57790314	2.6420	0.0115*
Log Wt*VType	7	7	0.70132749	3.2063	0.0027*

D/P*Protection

Least Squares Means Table

Level	Least Sq Mean	Std Error
Driver,d airbag	2.7448660	0.07232587
Driver,d&p airbags	2.7222092	0.16977327
Driver,manual belts	2.9566072	0.06603794
Driver,Motorized belts	2.8865003	0.07332272
Driver,passive belts	2.9369982	0.07397220
Passenger,d airbag	2.8034810	0.07503195
Passenger,d&p airbags	2.6022334	0.19426040
Passenger,manual belts	2.8384054	0.06602773
Passenger,Motorized belts	2.7947384	0.07371270
Passenger,passive belts	2.7643500	0.07401460

Residual Normal Quantile Plot



14: Generalized Linear Model

Learning objectives: understand how to model when the dependent variable follows a distribution other than the Normal distribution.

In the General Linear Model, we have that:
$E(y)=A+B\,x$ and $Y=E(y)+$ error term
Where the error term is from the Normal distribution with zero mean and constant variance.

For the Generalized Linear Model, we have that:
$g(E(y))=A+B\,x$, where g is called the link function and Y follows a specified distribution with mean E(y)

If Y is count data, then the distribution is Poisson with parameter λ, $E(y)=\lambda$ and g(y) is log(y)
If Y is binary outcomes, i.e. success=1/failure=0 then the distribution is Binomial(1,p), E(y) is p, the probability of success and g(y) is $\log(y/(1-y))$ = logit (y)
Can also model ordinal Y, using ordinal logistic model.

We can apply similar ideas to identify a good model that explains the variation in y

Please see the book "Generalized Linear Models" by Nelder and McCullagh for more technical details and examples.

For logistic models, i.e. for 2 levels,

Log(Pr(level 2)/Pr(level 1))= linear model+error,

i.e. expected log odds is linear in the factors and covariates (vs expected value for y is linear for general linear models).

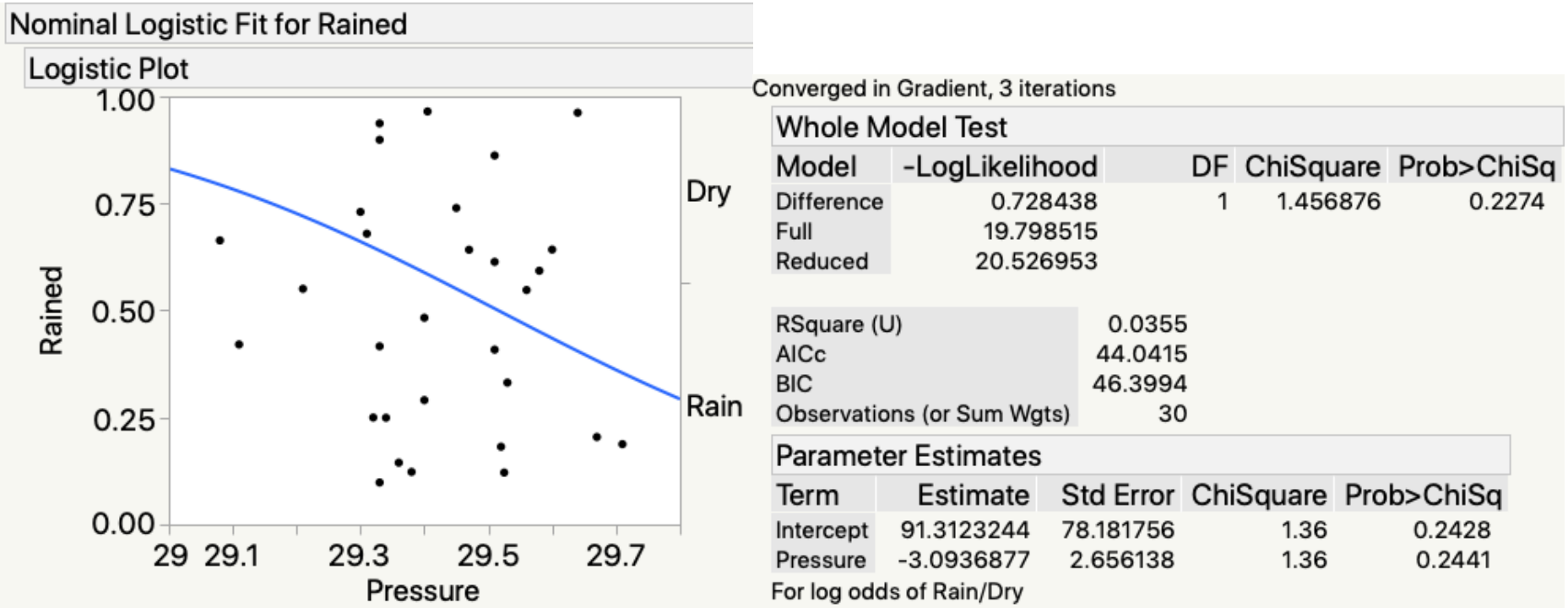
As the linear model moves from large negative to large positive, we shift from

Pr(level 1) close to 1 and Pr(level 2) close to 0

to Pr(level 2) close to 1 and Pr(level 1) close to 0.

These models need to be solved iteratively, but the computer shields us from this.

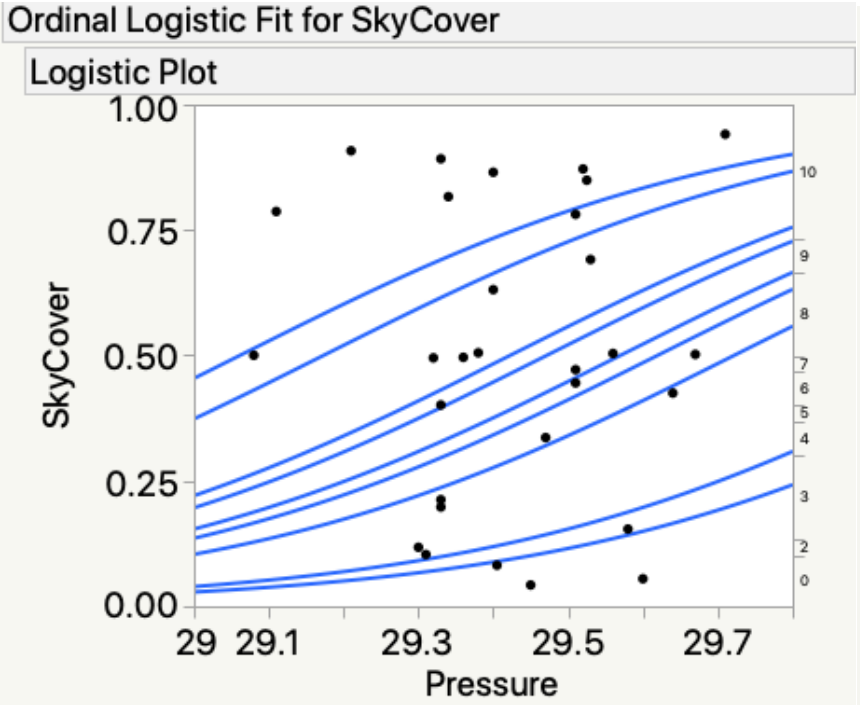
We look at an example, where the binary outcome for daily data is Rain or Dry and the covariate is the air pressure. Although the outcome is not statistically significant, we can see that as the pressure gets higher, it is less likely to rain.



We can extend these models for binary data to what is called ordinal logistic regression, which assumes a series of logistic regressions, which have the same slope but different intercepts. As this model only adds one parameter per level, it is still feasible for large numbers of levels.

We look again at the previous example, but this time we look at SkyCover (cloud cover of the sky), rather than rain, as our outcome. Clearly, there must be some SkyCover in order to get rain. SkyCover is ordinal scale (an integer from 0:no clouds to 10: complete cloud cover). There is only one parameter for pressure, which indicates how pressure affects the rise up the levels of SkyCover.

Ordinal Logistic Fit for SkyCover



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	0.925696	1	1.851391	0.1736
Full	60.925594			
Reduced	61.851290			

RSquare (U)	0.0150
AICc	153.43
BIC	155.863
Observations (or Sum Wgts)	30

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	206	52.084580	104.1692
Saturated	207	8.841014	Prob>ChiSq
Fitted	1	60.925594	1.0000

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Pressure	1	1	1.85139135	0.1736

There is also an extension for nominal variables, but we usually use an alternative formulation called log-linear models. These models assume a null hypothesis of:

Pr(Row i and Col j)=Pr(Row i) x Pr(Col j) (i.e. row and column independence)

or equivalently

Log(Pr(i,j))=Log(Pr(i))+Log(Pr(j))

Which is why they are called log-linear.

Note that this is symmetric in the variables (does not distinguish between independent and dependent variables).

This could be applied to our example of the Car Poll we looked at in the beginning, but it can also handle much more complicated situations, with many categorical variables.

We now look at an example of count data, where we are trying to understand how the count of crabs depends on crab Color and Weight. As mentioned above, for count data, the assumed distribution is Poisson and the link function is log(y). However, we need to be careful about two issues: firstly, we should use Log Weight as our covariate, to ensure that it maps onto the range of values for log(mean), and we also check for over dispersion, meaning that the variability is greater than we expect from Poisson (which assumes the variance is equal to the

mean). We can see that the variance is about three times bigger than expected, and that after accounting for this, Log Wt is statistically significant, while Color is not.

Whole Model Test

Model	-LogLikelihood	L-R ChiSquare	DF	Prob>ChiSq
Difference	14.4177935	28.8356	4	<.0001*
Full	143.007833			
Reduced	157.425626			

Goodness Of Fit Statistic	ChiSquare	DF	Prob>ChiSq	Overdispersion
Pearson	527.2302	168	<.0001*	3.1383
Deviance	542.2982	168	<.0001*	

AICc
298.5217

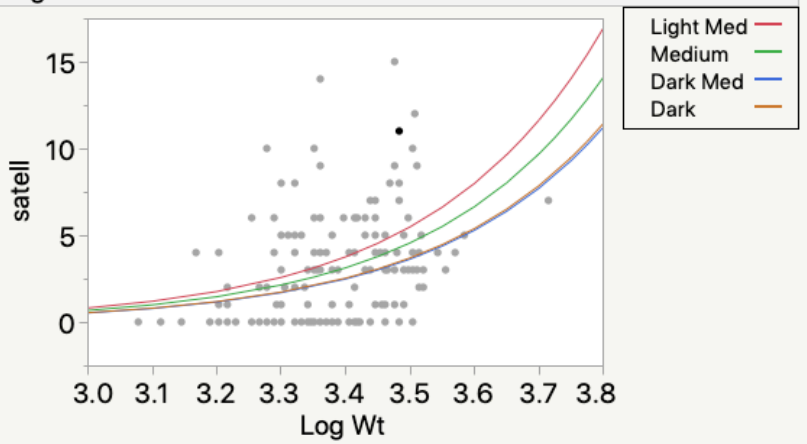
Effect Tests

Source	DF	L-R ChiSquare	Prob>ChiSq
color	3	2.3617291	0.5008
Log Wt	1	21.303066	<.0001*

Generalized Linear Model Fit

Overdispersion parameter estimated by Pearson Chisq/DF
Response: satell
Distribution: Poisson
Link: Log
Estimation Method: Firth Adjusted Maximum Likelihood
Observations (or Sum Wgts) = 173

Regression Plot



15: Experimental Design

Learning objectives: understand how best to design an experiment

We are often concerned with the analysis of data generated from an experiment. It is wise to take time and effort to organize the experiment properly to ensure that the right type of data, and enough of it, is available to answer the questions of interest as clearly and efficiently as possible. This process is called experimental design.

The specific questions that the experiment is intended to answer must be clearly identified before carrying out the experiment. We should also attempt to identify known or expected sources of variability in the experimental units since one of the main aims of a designed experiment is to reduce the effect of these sources of variability on the answers to questions of interest. That is, we design the experiment in order to improve the precision of our answers.

It is important to understand that optimizing the experiment in some ways may make it less useful in others (e.g. minimizing the prediction error may not allow us to fit polynomial or other non-linear models)

Main Effect

This is the simple effect of a factor on a dependent variable. It is the effect of the factor alone averaged across the levels of other factors.

Interaction

An interaction is the variation among the differences between means for different levels of one factor over different levels of the other factor.

Example

A cholesterol reduction clinic has two diets and one exercise regime. It was found that exercise alone was effective, and diet alone was effective in reducing cholesterol levels (main effect of exercise and main effect of diet). Also, for those patients who didn't exercise, the two diets worked equally well (main effect of diet); those who followed diet A and exercised got the benefits of both (main effect of diet A and main effect of exercise). However, it was found that those patients who followed diet B and exercised got the benefits of both plus a bonus, an interaction effect (main effect of diet B, main effect of exercise plus an interaction effect).

Blocking

This is the procedure by which experimental units are grouped into homogeneous (similar) clusters in an attempt to improve the comparison of treatments by randomly allocating the treatments within each cluster or 'block'.

Randomization

Randomization is the process by which experimental units (the basic objects upon which the study or experiment is carried out) are allocated to treatments; that is, by a random process and not by any subjective and hence possibly biased approach. The treatments should be allocated to units in such a way that each treatment is equally likely to be applied to each unit.

Randomization is generally preferred because the alternatives may lead to biased results. The logic is similar to why we need random samples in order to have samples that are representative (unbiased) of the population.

The main point is that randomization tends to produce groups for study that are comparable in unknown as well as known factors likely to influence the outcome, apart from the actual treatment under study. This is critical, as in most situations there are unknown factors, certainly when dealing with social or biological experiments.

The standard analysis of variance F tests assume that the treatments are fixed, but have been allocated randomly to subjects.

Blinding

In a medical experiment, the comparison of treatments may be distorted if the patient, the person administering the treatment and those evaluating it know which treatment is being allocated. It is therefore necessary to ensure that the patient and/or the person administering

the treatment and/or the trial evaluators are 'blind to' (don't know) which treatment is allocated to whom.

Sometimes the experimental set-up of a clinical trial is referred to as double blind, that is, neither the patient nor those treating and evaluating their condition are aware (they are 'blind' as to) which treatment a particular patient is allocated. A double-blind study is the most scientifically acceptable option.

Sometimes however, a double-blind study is impossible, for example in surgery. It might still be important though to have a single blind trial in which the patient only is unaware of the treatment received, or in other instances, it may be important to have blinded evaluation.

Placebo

A placebo is an inactive treatment or procedure. It literally means 'I do nothing'. The 'placebo effect' (usually a positive or beneficial response) is attributable to the patient's expectation that the treatment will have an effect. This is sometimes linked with the Hawthorne effect, which is the effect due to attention being paid to people, i.e. people respond positively to being assessed or measured. Clearly the Hawthorne effect should occur equally for all study participants as long as the level of interaction is similar.

Ethical concerns

Clearly, it is unethical to randomize the allocation of treatments to humans, if we know that one treatment is better than another (or than the placebo), so we sometimes have to use the

current best treatment for comparison instead. This can be complex when there are risks (e.g. side effects) as well as benefits of treatments. When it is unethical to not provide the one treatment being considered, one possibility is a wait control study where we compare before (waiting), during and after treatment (to see if the effect persists) for each individual.

It may also be important to look at cost-effectiveness, which requires collecting data about the full costs of each treatment (and control).

Note that all studies that involve treatment of or personal data collection from human or animal subjects require ethical approval. For clinical human studies, this involves the *Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (IRB)*, for live animals it involves the *Committee for the Use of Live Animals in Teaching and Research (CULATR)* while for other studies with personal data or human subjects, it is the *Human Research Ethics Committee for Non-Clinical Faculties (HRECNCF).*

Completely Randomized Design

The structure of the experiment in a completely randomized design is assumed to be such that the factor combinations are allocated to the subjects completely at random.

Randomized Complete Block Design

The randomized complete block design is a design in which the subjects are matched according to a variable that the experimenter wishes to control. The subjects are put into groups (blocks) of the same size as the number of treatments. The members of each block are then randomly assigned to different treatment groups.

Example
A researcher is carrying out a study of the effectiveness of four different skin creams for the treatment of a certain skin disease. He has eighty subjects and plans to divide them into 4 treatment groups of twenty subjects each. Using a randomized blocks design, the subjects are assessed and put in blocks of four according to how severe their skin condition is; the four most severe cases are the first block, the next four most severe cases are the second block, and so on to the twentieth block. The four members of each block are then randomly assigned, one to each of the four treatment groups.

Factorial Design

A factorial design is used to evaluate two or more factors simultaneously. The treatments are combinations of levels of the factors. The advantages of factorial designs over one-factor-at-a-time experiments are that they are more efficient and they allow interactions to be detected.

Full Factorial Design

A full factorial design is a factorial design that covers all combinations of the factor levels an equal number of times (more than one times, means we have replicates).

Fractional Factorial Design

A fractional factorial design, means a design that only covers a simple fraction of a factorial design and is a full factorial design for a subset of the factors (e.g. for any pair of factors).

Principles of Optimal Design

Overall, the idea is that when the data collection can be fully (or partially controlled), we can "design" an experiment in an optimal way to answer our questions. These questions can be of different types:

a) Which factors affect the response and how?
b) What is the optimal combination of factors to maximize (or minimize) the response?
c) Which factors look like they have the greatest influence on the response?
d) What type of relationship do the factors have on the response (linear, quadratic etc.)

For a), we usually want a full factorial design, which means generating all possible combinations of the levels of the factors, which can be a large number as the number of factors and/or levels increases.

For b) we usually assume a quadratic response and then choose data points to help us find the maximum or minimum – this is called a response surface design

For c) we are interested in a screening design, which usually means assuming only 2 levels for each factor and also using a fractional factorial design. In this case, we are usually assuming that all the interactions are small (or at least that all higher order interactions are small) and we focus on the main effects of each factor, e.g. 10 factors with 2 levels each = 1024 possibilities, one eighth fraction would be 128, a sixteenth would be 64, a thirty-second would be 32, a sixty-fourth would be 16, while 5 factors with 3 levels each=243 possibilities, one third fraction would 81, one ninth would be 27

For d) we need data spread out over different levels for the factors in order to identify any non-linearity. This is only feasible for a small number of factors. We need at least 3 levels in order to have any chance of identifying non-linearity at all.

There are some general principles in choosing a design:

1) Keep the design balanced (i.e. equal number of data points for each factor level)
2) Take values at extremes
3) Take values at combinations of extremes
4) Put some points near the center to check for curvature (more for d))
5) Randomize the assignments of the data points to minimize any uncontrolled for effects and interactions.

We start with an example of a fractional factorial experiment to see which of 7 binary factors affect the speed when riding a bike: gear, dynamo, seat, tires, handlebars, breakfast and raincoat. There are only 8 observations, which is 2^3 compared to 2^7 observations needed for a full factorial experiment. This means that even for main effects, we cannot estimate the residual, so we can only order the effects, not test their significance.

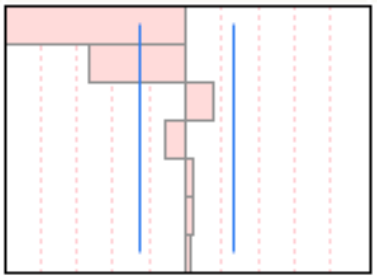
Response Y

Summary of Fit

RSquare	1
RSquare Adj	.
Root Mean Square Error	.
Mean of Response	66.5
Observations (or Sum Wgts)	8

Sorted Parameter Estimates

Term	Estimate	Relative Std Error	Pseudo t-Ratio	Pseudo p-Value
Gear[low]	-11.25	0.353553	-15.00	0.0023*
Dynamo[off]	-6	0.353553	-8.00	0.0096*
Seat[down]	1.75	0.353553	2.33	0.1266
Tires[hard]	-1.25	0.353553	-1.67	0.2195
HBars[down]	0.5	0.353553	0.67	0.5649
Brkfast[no]	0.5	0.353553	0.67	0.5649
Raincoat[off]	0.25	0.353553	0.33	0.7665



No error degrees of freedom, so ordinary tests uncomputable. Relative Std Error corresponds to residual standard error of 1. Pseudo t-Ratio and p-Value calculated using Lenth PSE = 0.75 and DFE=2.3333

Next we have a full factorial experiment with 5 binary factors (catalyst, temperature, concentration, stir rate, feed rate) but no replicates where we are trying to identify what affects reactor output, so we try fitting the main effects and 2 way interactions. The results suggest that catalyst, temperature, concentration and their interactions matter:

Response Y

Summary of Fit

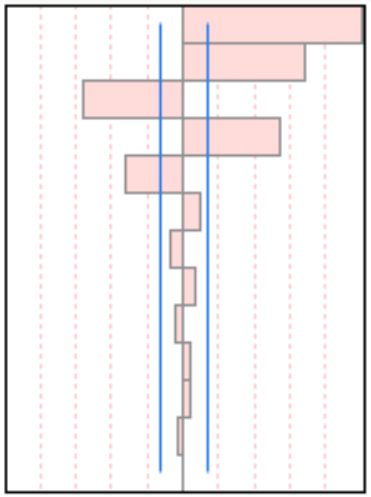
RSquare	0.970281
RSquare Adj	0.948817
Root Mean Square Error	3.385016
Mean of Response	65.5
Observations (or Sum Wgts)	32

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	13	6733.7500	517.981	45.2056
Error	18	206.2500	11.458	Prob > F
C. Total	31	6940.0000		<.0001*

Sorted Parameter Estimates

Term	Estimate	Std Error	t Ratio		Prob>|t|
Catalyst(1,2)	9.75	0.598392	16.29		<.0001*
Temperature*Catalyst	6.625	0.598392	11.07		<.0001*
Concentration*Temperature	-5.5	0.598392	-9.19		<.0001*
Temperature(140,180)	5.375	0.598392	8.98		<.0001*
Concentration(3,6)	-3.125	0.598392	-5.22		<.0001*
Concentration*Catalyst	1	0.598392	1.67		0.1120
Feed Rate(10,15)	-0.6875	0.598392	-1.15		0.2656
Catalyst*Feed Rate	0.6875	0.598392	1.15		0.2656
Temperature*Feed Rate	-0.4375	0.598392	-0.73		0.4741
Stir Rate*Catalyst	0.4375	0.598392	0.73		0.4741
Stir Rate*Feed Rate	0.375	0.598392	0.63		0.5387
Stir Rate(100,120)	-0.3125	0.598392	-0.52		0.6079
Concentration*Feed Rate	0.0625	0.598392	0.10		0.9180



16: Time Series

Learning objectives: understand how to model when the residuals are correlated over time sequence

When data are collected in a series over time, we call this a time series. Standard methods require that the data are collected at equally spaced time points (e.g. hourly, daily, monthly or yearly). The key difference from simple linear models is that the assumption that the errors are independent is usually no longer valid – the simplest generalization is that the errors are autocorrelated – in other words

Correlation of the errors for y_t and $y_{t+k} = \rho^k$, where y_t is the observed value at time t and y_{t+k} is the observed value at time k units later.

You can see that this says that errors that are closer together in time have a stronger correlation. If ρ is small, then the usual linear models will be appropriate, so we can test whether ρ is zero in order to decide whether we need to use this generalization. This test is called the Durbin-Watson test and is wise if our data has an obvious ordering. We examine this test for the Longley dataset, where it shows marginal significance for a small autocorrelation:

Response y

Whole Model

Summary of Fit

RSquare	0.992847
RSquare Adj	0.991059
Root Mean Square Error	332.0844
Mean of Response	65317
Observations (or Sum Wgts)	16

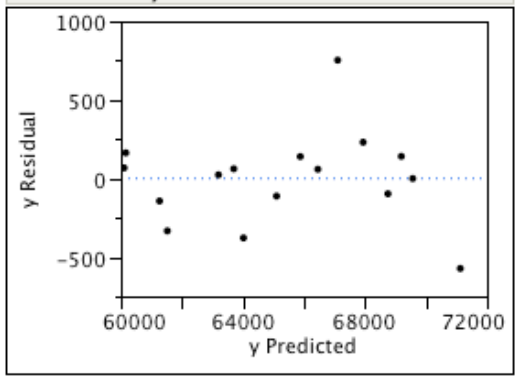
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	183685465	61228488	555.2091
Error	12	1323361	110280.06	Prob > F
C. Total	15	185008826		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1797221	68641.55	-26.18	<.0001*
x3	-1.469671	0.167137	-8.79	<.0001*
x4	-0.772281	0.183715	-4.20	0.0012*
x6	956.3798	35.52482	26.92	<.0001*

Residual by Predicted Plot



Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.4630915	16	0.1344	0.0484*

This situation (of $\rho<>0$) requires a relatively simple generalization of the linear models we have been discussing. Autocorrelation means that we may have less information than we think (i.e. the effective sample size is smaller).

Another type of time series model is a seasonal model. This means that we include an independent variable that denotes which season (e.g. spring, summer, autumn and winter). Clearly, this is just a particular factor, so it can be included in the linear models we already know how to handle, **if** we know the length of the cycle. "Seasons" are not necessarily seasons in the usual sense of the word, but can be hours of day or stages of any cycle. Note that finding the length of the cycle may be an important part of the analysis for some data sets. Special cases are where the seasonal pattern follows a sinusoidal curve (e.g. tide heights and other lunar/solar patterns).

The next stage of sophistication in time series models is differencing, moving averages and autoregressive models. The basic idea of these generalizations (usually considered together) is that by using differences of the observed y (autoregressive models) or of the errors (moving averages), possibly after initial differencing to remove trends, we can get a simple linear model. These models are known as ARIMA models (autoregressive integrated moving average)

In other words: Autoregressive (AR) models: forecasts are based on a linear combination of previous values
Moving Average (MA) models: forecasts are based on a linear combination of previous errors (i.e. a linear combination of residuals)

ARIMA models: combine differencing, AR and MA.

Differencing on y_t means= y_t- y_{t-1}

In other words, we calculate the difference between an observation and the one before it in time. This can be used to remove trends, as standard methods assume stationarity (the mean and variance are stable over time)

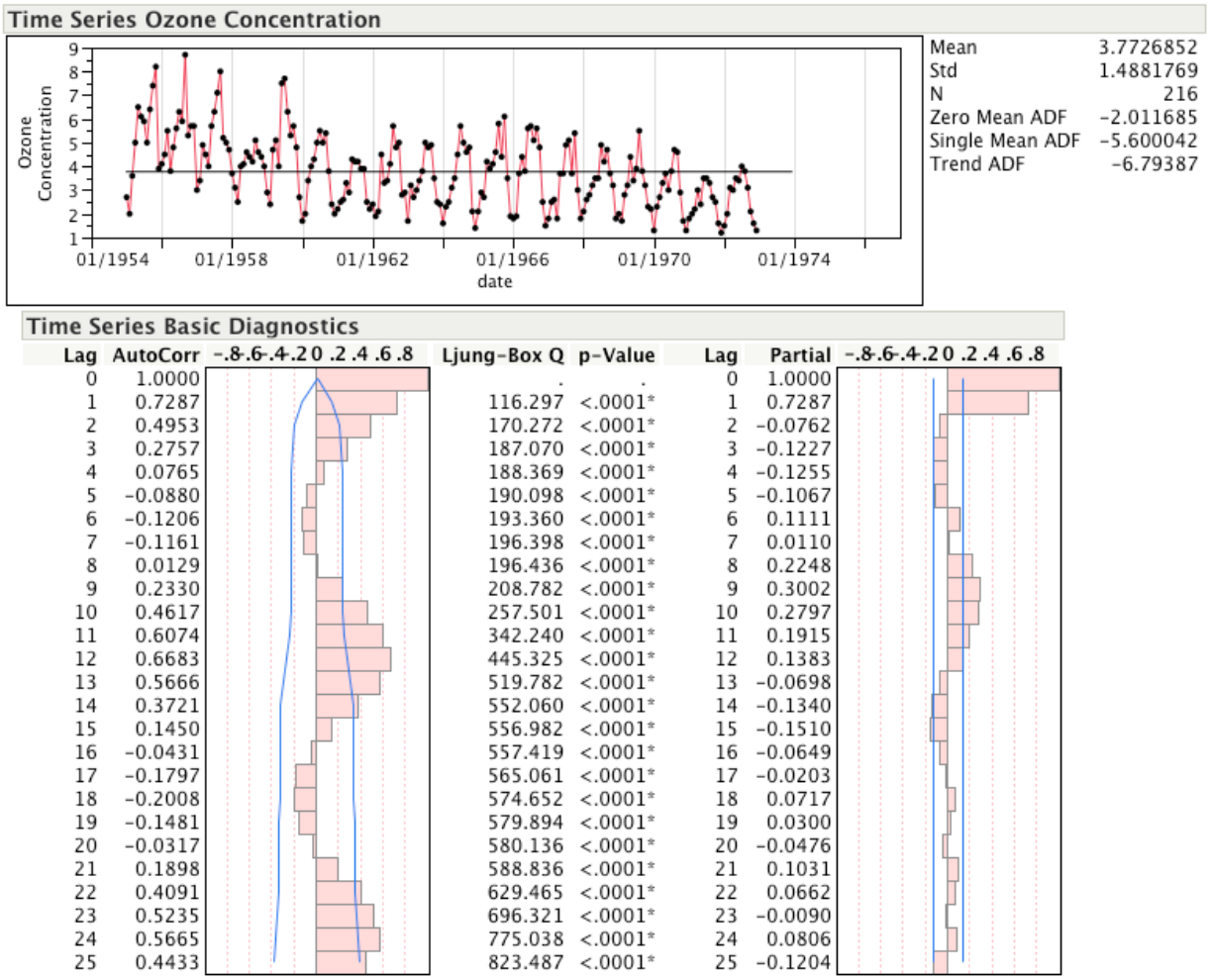
A simple example: if you have a simple linear trend
y_t=a +b t + error, then (y_t- $y_{t-1)}$=b + error, i.e. a trend becomes a constant.

If you apply differencing again, you get:
(y_t- y_{t-1})-(y_{t-1}- y_{t-2}) =0 + error, so the constant disappears.

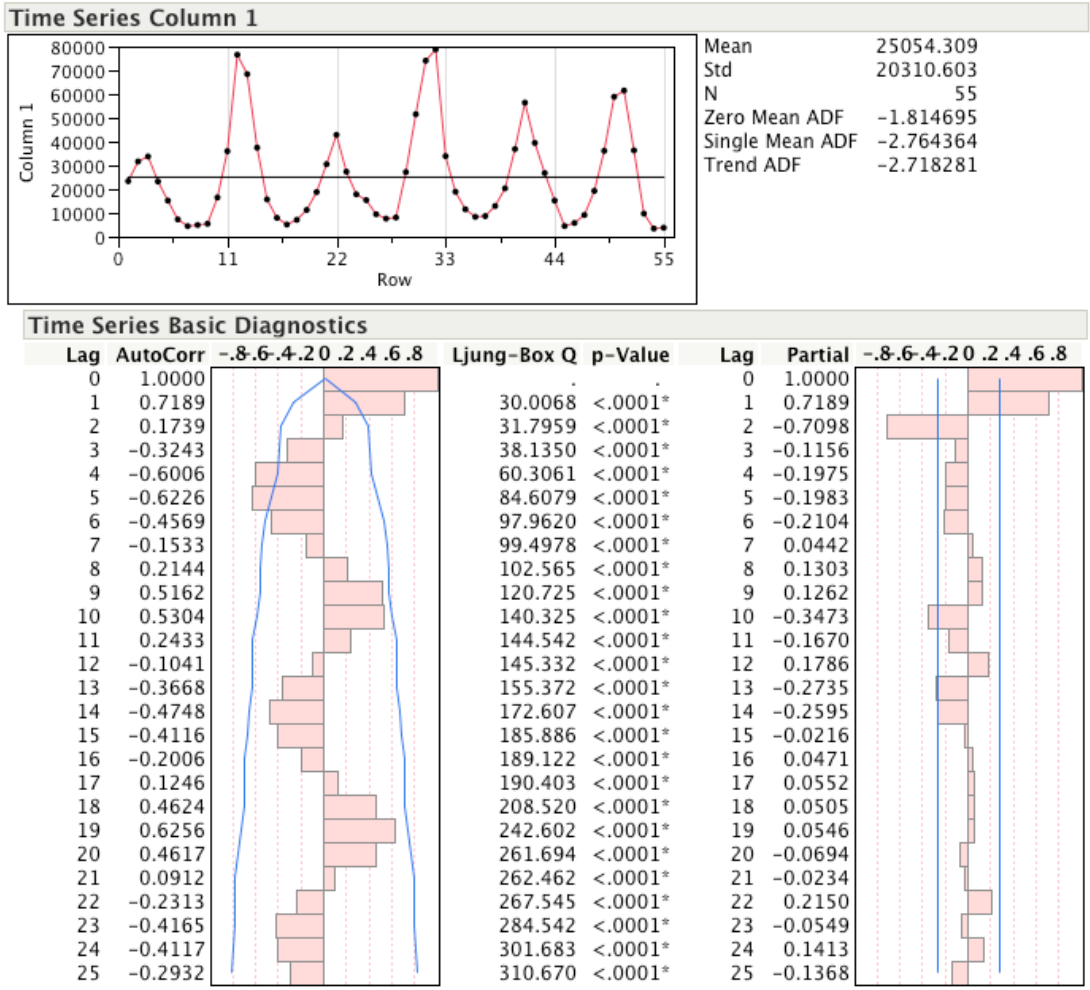
General ARIMA models are hard to identify, but special cases are easier:

For pure MA with k terms, the autocorrelation function is zero after the kth term, where autocorrelation is simply the sample correlation of y_t with y_{t-k} for different values of k.

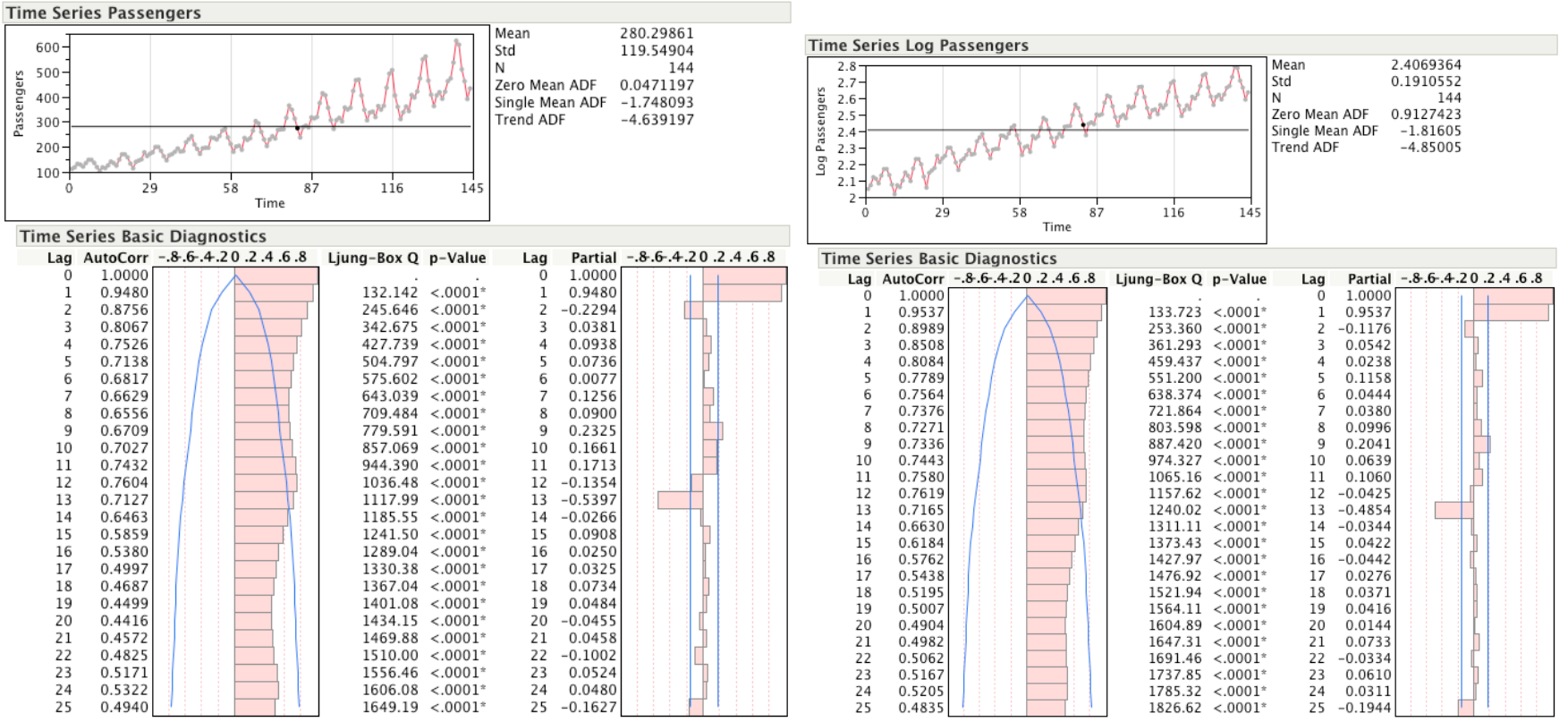
For pure AR with k terms, the partial autocorrelation function (pacf) is zero after the kth term, where the pacf is the sample correlation between y_t with y_{t-k} for different values of k, after first removing the effects of the intervening y, i.e. y_{t-1} to y_{t-k+1} on both variables by using linear regression, i.e. pacf is the autocorrelation of the residuals from the regressions.



The first example is looking at monthly Ozone concentrations, which show a strong seasonal pattern and a downward trend.

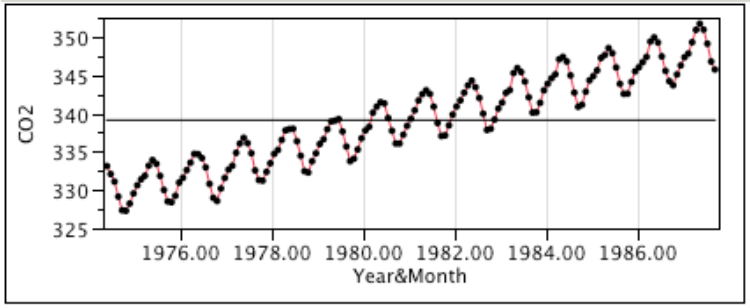


The second example looks at the yearly numbers of lynx (a wild cat) caught in Canada and shows a strong cycle of 9.5 years.



This example shows the pattern for passenger numbers – we can see that log transformation is helpful here

Time Series CO2



Mean	339.13255
Std	6.0282384
N	161
Zero Mean ADF	0.7643682
Single Mean ADF	-1.283413
Trend ADF	-3.882658

Time Series Basic Diagnostics

Lag	AutoCorr	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
0	1.0000		.	.	0	1.0000	
1	0.9714		154.760	<.0001*	1	0.9714	
2	0.9096		291.306	<.0001*	2	-0.6022	
3	0.8328		406.490	<.0001*	3	0.0596	
4	0.7601		503.071	<.0001*	4	0.2348	
5	0.7031		586.240	<.0001*	5	0.1203	
6	0.6683		661.858	<.0001*	6	0.1391	
7	0.6581		735.660	<.0001*	7	0.1787	
8	0.6728		813.311	<.0001*	8	0.2291	
9	0.7087		900.032	<.0001*	9	0.2005	
10	0.7552		999.140	<.0001*	10	0.0670	
11	0.7935		1109.31	<.0001*	11	-0.1702	
12	0.8049		1223.42	<.0001*	12	-0.2880	
13	0.7781		1330.76	<.0001*	13	-0.3110	
14	0.7201		1423.33	<.0001*	14	-0.0243	
15	0.6480		1498.80	<.0001*	15	0.1466	
16	0.5794		1559.57	<.0001*	16	0.1741	
17	0.5255		1609.89	<.0001*	17	0.0833	
18	0.4913		1654.19	<.0001*	18	-0.0111	
19	0.4805		1696.86	<.0001*	19	0.0083	
20	0.4926		1742.03	<.0001*	20	-0.0733	
21	0.5245		1793.60	<.0001*	21	-0.0496	
22	0.5662		1854.14	<.0001*	22	-0.0287	
23	0.6012		1922.87	<.0001*	23	-0.0263	
24	0.6112		1994.43	<.0001*	24	-0.0761	
25	0.5858		2060.65	<.0001*	25	-0.0954	

Finally, this time series plot for CO_2 shows a strong trend and seasonal pattern.

17: JMP Basics

Learning objectives: understand basics of using JMP

Data Tables

We will assume that we can represent data as cases=rows and variables=columns, i.e. basic spreadsheet concept. The main variation on this is that we may use one column as a count or frequency column. This allows us to avoid entering every case, when some of the cases are identical. (Illustrate simple 2x2 table in JMP)

Variable Type

One key issue when doing statistics is the type of variables that we have, in particular, whether they are nominal, ordinal or continuous. In JMP, we have to select the type for each column and this helps in that JMP can be smart in identifying appropriate analysis as long as the type is correct. (Illustrate in JMP)

Entering Data

Data can come from multiple sources, including import from Excel or a text table, keyboard entry or randomly generated. For importing, you need to know what the delimiter is and whether the names of variables are in the first row. JMP tables can also be exported similarly. (Illustrate import/export of animals from Sample Import folder and keyboard entry of new data)

Row Selection

One of the beauties of an interactive package is that we can select rows in one place and they will be highlighted in other places.
The rows can be selected manually (by selecting in the spreadsheet view), programmatically (will show), or by lassoing points in a graph (will show). Selected rows can then be coloured, hidden or excluded. This is invaluable if trying to understand whether there is a problem with some data points. More detailed selections can be made using repeated use of row selection to narrow down or expand the selected rows. (illustrate different ways of selecting rows and then excluding or coloring using students)

Graphs and Reports

To keep results, you need to either (Mac: export, Windows: save as) the output as a pdf, rtf, text or html file or copy and paste results into another file (e.g. Word) (illustrate)

Manipulating Tables

Sometimes you need to reorganize tables, using sorting, summarizing or stacking. You can also join tables together. (summarise Companies, stack Cheese taste, subset(rows), sort(rows), transpose(swap rows and columns), concatenate(rows), join (columns), split(reverse of stack))

Formulae for creating variables or generating random data

If we want to adjust data (e.g. standardize by subtracting the mean and dividing by the S.D), recode or create random data, the formula editor is our friend (like a more sophisticated formula in Excel applied to columns rather than cells). (illustrate using students to standardize, hotdogs to recode, pendulum to transform and create randdist to generate)

18: Data display with JMP

Learning objectives: understand how to display data in JMP

Univariate distributions

We start by looking at what JMP does when we ask for distributions for different types of variable.

Key questions include whether the data is unimodal (if not, perhaps there is more than one process generating the data) and whether it might follow a Normal distribution (possibly after transformation). We can use alternative plots that may offer advantages in some situations (e.g. box plots and stem-and-leaf) and Normal quantile plots (which plots the data against what we would expect if the data came from a Normal distribution with matching mean and variance and provides confidence limits to illustrate if the data is non-Normal).

It is important to remember that statistical significance may be different to practical significance

Graphics can help us identify when the data is truncated or there are atypical points.

Sometimes, the test of normality is rejected for very minor deviations from normality, because there is a very large dataset. We also need to be careful to look at data within each group, rather than the overall dataset.

Non-normal data

Sometimes a power transformation turns the data into Normal data (i.e. the residuals are Normal). We can write this as:

$Z=(y^{\alpha}-1)/\alpha$

So that the limiting case of α goes to zero becomes $Z=Ln(y)$ (which is the natural logarithm). Log transforms often make sense because they turn multiplicative or ratio processes in additive and subtractive ones (which underlie the Normal distribution). They can work as long as y is always positive (examples later). If data includes zeros, may add a small amount before doing log transform to account for rounding (add ¼ if rounded to integer) or truncation (add ½ if truncated to integer)

If power transform does not work, can use non-parametric methods – simplest idea is to use ranks as a transformation – i.e. analyse ranks rather than raw data – t-test and F-test become Mann-Whitney and Kruskal-Wallis tests. Only difference is that proper nonparametric analysis takes accounts of ties (equal values) when calculating significance levels, which is missed if just using ranks in standard analysis. Without accounting for ties, we think we have more information than we really do. (Ranks can be added through the save option in Distribution)

Multivariate data

Can do multiple bivariate plots or use principal component approach
Bivariate plots means plotting all the pairs of dependent variables, which does not always reveal all the structure of the data.
Principal components is based on the idea that linear combinations of the dependent variables may be able to summarise most of the variability in a relatively small number of dimensions. The extreme case would be where all the variables are perfectly correlated (+1 or -1), when one variable can summarise all the variability. The disadvantage is that the linear combinations are not always easy to interpret.

19: Model Building with JMP

Learning objectives: understand how to build models in JMP

Two groups

Testing for 2 groups is easy, as it is still based on the t-test.
We need to distinguish between paired and unpaired data - paired data turns back into 1 group data by looking at differences.

Multiple Groups (One-way ANOVA)

Simplest case is balanced design (equal sample size for each level, i.e. each group) with constant variance (same for different groups), but can adjust for both of these.
Multiple comparisons: often interested in comparisons between levels, e.g. which pairs are significant or best versus rest. Need to adjust significance level for multiple tests. Usually start by requiring that overall F-test is significant.

Multiple Regression

Model selection – dangers of "data mining" leading to models that are too complex (linnerand example) (average number of variables added by chance will be significance level times the number of variables tested, if we ignore correlations).
Diagnostics – residuals and leverage

Residuals help us identify model deficiencies (are the residuals Normal, is the variance constant, is there are a missing variable, do we need a transformation)
Look for patterns in residuals against row, against covariates for mean or variance
Row (sequence) pattern for residual suggests missing variable (time?) or a change over time.
Pattern for residual against covariate suggests need for products or powers of covariates.
Pattern in variance suggests need for transformation to stabilize variance (assumed constant in standard models)
Transformation – can try log, rank etc.
Leverage helps us identify high dependence of model fit on a few key data points (model not necessarily "wrong", but some risk and need to check).
Collinearity – problem with covariates that are collinear (i.e. can predict one covariate from others), so parameter estimates not well defined, or may change considerably depending on which other variables included.

General Linear Models with JMP (Multiple Regression + Multi-way ANOVA combined) (Analysis of Covariance)

Model selection is now much more complex as we need to select from a wide range of models. Aim at simple answers (small degrees of freedom for model) that fit well and that can be explained in terms of a sensible theory.
Notation:
Y=A,X,A*X

Means Factor A, Covariate X (A and X are both called Main Effects) and the Interaction between the 2, where Factor means a categorical independent variable and Covariate is a continuous independent variables. (Factor: ANOVA, Covariate: Regression model)
For multiple factors, can have many interactions, but we often assume that higher level interactions are likely to be (relatively) unimportant.
Look at lack of fit to check whether interactions are missing – only useful if we have replicates (i.e. multiple data points with same factor levels and covariate values). Replicates are very useful as they enable us to estimate the error variance independent of any model errors (at least for those factors and covariates included)
Least Square Means – adjusted means for levels of a factor, after adjusting for covariates (e.g. if looking at lung cancer rates for smokers and non-smokers of different ages, then we can adjust the means of smokers and non-smokers so as to remove any differences due to the different average ages of smokers and non-smokers).
Nested (hierarchical) classification

Multivariate Linear Model with JMP

The difference from the general linear model is that we have multiple dependent variables, which are correlated (If uncorrelated, can look at them separately). Generally assume that the same independent variables for all the dependent variables (i.e. they have similar causal processes). In order to choose the appropriate tests, need to know whether the measurements are all repeated measurements of the same construct (repeated measures) or not. If not,

usually use identity matrix formulation, which means focus primarily on the separate variables, rather than some specific combination of them.

Generalized Linear Model with JMP

Extend to data that does not follow the Normal distribution, but some other distribution.

In the General Linear Model, we have that:
$E(y)=A+B\,x$ and $Y=E(y)+$ error term
Where the error term is from the Normal distribution with zero mean and constant variance.

For the Generalized Linear Model, we have that:
$g(E(y))=A+B\,x$, where g is called the link function and Y follows a specified distribution with mean E(y)
e.g. Y is count data, then the distribution is Poisson with parameter λ, $E(y)=\lambda$ and g(y) is log(y)
e.g. Y is binary outcomes, i.e. success=1/failure=0 then the distribution is Binomial(1,p), E(y) is p, the probability of success and g(y) is log(y/(1-y)) = logit (y)

We can apply similar ideas to identify a good model that explains the variation in y

Please see the book "Generalized Linear Models" by Nelder and McCullagh for more technical details and examples.

Ordinal Data with JMP

Use logistic models, i.e. for 2 levels,
Log(Pr(level 2)/Pr(level 1))= linear model+error,
Instead of y=linear model+error
i.e. expected log odds is linear in the factors and covariates (vs expected value for y is linear for general linear models).

As the linear model moves from large negative to large positive, we shift from
Pr(level 1) close to 1 and Pr(level 2) close to 0 to Pr(level 2) close to 1 and Pr(level 1) close to 0.
These models need to be solved iteratively, but the computer shields us from this.
Can extend these models for ordinal data to what is called ordinal logistic regression, which assumes a series of logistic regressions, which have the same slope but different intercepts.
As this model only adds one parameter per level, it is still feasible for large numbers of levels.
There is also an extension for nominal variables, but we usually use an alternative formulation called log-linear models.
These models assume a null hypothesis of:
Pr(Row i and Col j)=Pr(Row i) x Pr(Col j)
Or Log(Pr(i,j))=Log(Pr(i))+Log(Pr(j))
Which is why they are called log-linear.
Note that this is symmetric in the variables (does not distinguish between independent and dependent variables) and corresponds to independence of row and column in Chisquared goodness of fit.

Compositional data with JMP

This is data where non-negative variables sum to a known constant – usually 100% or 1, such as when we have a set of proportions. In this case, correlations do not have the usual meaning because of the constraint. If the variables are all within the range (0,1), i.e. there are no zeroes or ones, we can use a log ratio transformation that is meaningful if we believe that relative ratios are what matter. For example, if we have 3 components: (x1, x2, x3) that sum to 1, then we can look at y1=log(x1/x3), and y2=log(x2/x3) using the standard General Linear Model approach.

Note that we can map these log-ratios back to the original x space and that it does not matter which of the x's we use as a divisor for General Linear models. Please see John Aitchison's book: "The Statistical Analysis of Compositional Data" for more details of this approach.

Time Series with JMP

For time series data, we need to account for dependence of data on data collected at previous time points. The basic theory is introduced in XV. The key practical skill is to identify the structure of time dependence (combinations of autocorrelation, moving average, seasonality, time trends on different time scales).

20: R Basics

Learning objectives: understand basics of using R, including install R; install packages; import & manipulate data

Installation of R

How to install R depends on which operating system you use and even on which version of that operating system. As R is open source, you can install it from scratch, i.e. compiling the source code, but that is not recommended unless you are already familiar with compiling code from scratch on your system, as it can be hard to get it to work properly. Instead, visit the following repository where the latest version of R can always be downloaded, together with installation instructions. If you have an older version of R, I recommend installing the latest version.
https://cran.r-project.org

However, there is a HK mirror of CRAN that I recommend you use:
https://mirror-hk.koddos.net/CRAN/

There are pre-built installers for
Windows (one installer that installs 32bit or 64bit versions as appropriate)
OSX (different installers for 10.15, 10.11-10.14)
Linux (installers for Debian, Redhat, Suse and Ubuntu)

Important Notes:

1. You need to turn off anti-virus software, which otherwise may block R from moving files and hence completing the installation of packages.
2. Packages installation may have problems if you have not set your location/region to an English language location (e.g. Australia, Canada, UK, USA) so I strongly encourage making this change before installing packages - you can change it back after installation. If you get warnings like: Setting LC_CTYPE failed, using "C", then switch to US or British locale, or type into R: On Mac:
   system("defaults write org.R-project.R force.LANG en_US.UTF-8")
   On Windows/LINUX:
   Sys.setlocale("LC_ALL","English")
   In both cases, restart R after this.
3. If the files do not download, check if your firewall is blocking the download
4. Check that all dependencies are installed, if not install them manually until there are no errors when loading the library.

Mac specific

You must install XQuartz first, which provides a UNIX based windowing system known as X11, and you must log out and log in again before installing R. If you have problems, you can try to run R from Terminal instead of double clicking on the R.app Directory names in OSX use : as the separator, but R expects /

If having problems, check the FAQ: https://cran.r-project.org/bin/macosx/RMacOSX-FAQ.html

Windows specific:

Please choose the "single-document "R interface (SDI). In the Startup options screen of the R installer, select Yes (customized startup). Then select the SDI (single-document interface) in preference to the default MDI (multiple-document interface). You need access rights to the folder you have installed R in, or you can right click the R icon and select "Run as administrator." Directory names for Windows inside R need double slash \\ instead of single \

If having problems, check the FAQ: https://cran.r-project.org/bin/windows/base/rw-FAQ.html

There is a free 100 page R manual called Introduction to R, in the repository.

However, these installations provide command line interfaces rather than menu-based systems. So, we will also install R Commander, which adds menus for many common statistical procedures. Details can be found here:

https://socialsciences.mcmaster.ca/jfox/Misc/Rcmdr/index.html

including details of plug-ins which extend the capability.

Before installing packages, it is good practice to change:
R Preferences : Startup: Default CRAN mirror to be the HK mirror

It is also good practice to change the Directory to the folder containing your R datafiles.

R Commander is an add-on to R called a package, so we can install it from inside R, with this instruction:

```
install.packages("Rcmdr", dependencies=TRUE)
```

```
(allow R to install any dependencies, meaning other software needed to make it work, if
it asks which mirror to use, select the HK mirror)
```

and then load it using this instruction: `library(Rcmdr)`

On the Mac, you can either start R by double-clicking the R GUI program installed, or by typing R in the Terminal (which can be found in the Utilities folder). Using Terminal works better if you intend to use R Commander, which sometimes conflicts with the R GUI, so I will use this approach. Terminal can be found inside Utilities folder inside Applications folder.

Note: One very good thing about R Commander is that it shows you which R commands it used in the R Script window. This allows you to save the script for what you did or even save a markdown version for writing reports, as well as saving your output. R Commander has help available for the menu choices.

A more advanced system than R Commander, called R Studio, which includes a debugger and visualisation tools and can operate as a browser interface to a server, can be found here (the basic version is free): https://www.rstudio.com

More information about markdown can be found at the R Studio site:
https://rmarkdown.rstudio.com

Data Tables

We will assume that we can represent data as cases=rows and variables=columns, i.e. basic spreadsheet concept. Note that weights in R usually mean precision weights, not frequency weights as in SPSS and will usually give different results as precision weights do not affect degrees of freedom calculations.

Variable Type

One key issue when doing statistics is the type of variables that we have, in particular, whether they are nominal, ordinal, interval or ratio scale. R only knows categorical (caused by entering anything non-numeric as data, which it then assumes is nominal scale and hence creates a factor) and numeric data. It is your responsibility to know whether your numeric data are nominal, ordinal, interval or ratio scale and hence whether the analysis is appropriate. Note that factors in R can be identified as ordinal. R Commander assumes that all numeric data is interval scale, so if you have numeric coding of any categorical variables, it is wise to use R Commander to create new factors to use instead and hence reduce the risk of mistakes (illustrate creating a factor from a numeric variable, including ordering).

Entering Data

Data can come from multiple sources, including import direct from Excel, SPSS, SAS, Minitab, STAT, URL or a text table, keyboard entry or randomly generated. Some R packages already include datasets, which can easily be loaded. We will start with the iris dataset, which is inside the package datasets. In iris, there are 4 interval scale variables and 1 categorical (Species)

For importing from text, you need to know what the delimiter (field separator) is and whether the names of variables are in the first row. R data tables (known as data frames in R) can also be exported similarly. (Illustrate import of body.txt, USPres92.xlsx and keyboard entry of new data). R uses NA to mean missing and if you enter anything that is not a number or NA, it assumes the variable is a categorical variable (factor).

Saving Data

Saving data means that the data is saved as an R data frame and can be easily reloaded.

Manipulating Tables

Sometimes you need to reorganize tables (see Data/Active data set), using sorting, aggregating, subsetting or stacking.

Creating variables or generating random data

If we want to create new variables (e.g. standardize by subtracting the mean and dividing by the S.D), recode, create random data, convert numeric variables to factors, the data editor (see Data/Manage variables in active data set) is our friend (like a more sophisticated formula in Excel applied to columns rather than cells).

21: Data display with R

Learning objectives: understand how to display data in R

Univariate distributions

We start by looking at what R Commander does when we ask for (statistics) summaries for different types of variable.

For numeric, it offers: min, max, median, mean, SD, quartiles.
For factors it offers: frequency distributions

It also offers graphs such as stem and leaf plots, histograms, box plots for numeric data

Illustrate using data from package datasets

iris(interval and factor), trees(interval), airquality (create factor for Month)

Key questions for interval scale data include whether the data is unimodal (if not, perhaps there is more than one process generating the data) and whether it might follow a Normal distribution (possibly after transformation). We can also use Normal quantile plots (which plots the data against what we would expect if the data came from a Normal distribution with

matching mean and variance and provides confidence limits to illustrate if the data is non-Normal).

It is important to remember that statistical significance may be different to practical significance. Graphics can help us identify when the data is truncated or there are atypical points. Sometimes, the test of normality is rejected for very minor deviations from normality, because there is a very large dataset. We also need to be careful to look at data within each group, rather than the overall dataset.

Sometimes a power transformation can turn the data into Normal distributed data (i.e. the residuals are Normal, not the raw data). We can write this as:

$Z=(y^{\alpha}-1)/\alpha$
So that the limiting case of α goes to zero becomes $Z=\log(y)$

Log transforms often make sense because they turn multiplicative or ratio processes into additive and subtractive ones (which underlie the Normal distribution). They can work as long as y is always positive (examples later). If data includes zeros, may add a small amount before doing log transform to account for rounding (add ¼ if rounded to integer) or truncation (add ½ if truncated to integer)

Note: R has many functions including abs(), sqrt(), log(), log10(), exp(), min(), max(), y^x, rank(). However, inside R formulae, I() means the operators inside the bracket, such as ^ or +, are used in the arithmetic sense, otherwise + means include as a factor or covariate, * means cross and ^ means cross to a degree.

If power transform does not work, we can use non-parametric methods – simplest idea is to use ranks as a transformation – i.e. analyse ranks rather than raw data – t-test and F-test become Mann-Whitney/Wilcoxon and Kruskal-Wallis tests. Only difference is that proper nonparametric analysis takes accounts of ties (equal values) when calculating significance levels, which is missed if just using ranks in standard analysis. Without accounting for ties, we think we have more information than we really do. (Ranks can be added through the rank function)

Multivariate data

Can do scatterplot matrix (under Graphs) or use principal component approach (under Statistics: dimensional analysis in R commander).

Scatterplot matrix means plotting all the pairs of dependent variables, which does not always reveal all the structure of the data. Principal components is based on the idea that linear combinations of the dependent variables may be able to summarise most of the variability in a relatively small number of dimensions. The extreme case would be where all the variables are perfectly correlated (+1 or -1), when one variable can summarise all the variability. The disadvantage is that the linear combinations are not always easy to interpret.

22: Model Building with R

Learning objectives: understand how to build models in R

Two groups

Testing for 2 groups is easy, as it is based on the t-test.
We need to distinguish between paired and unpaired data - paired data turns back into 1 group data by looking at differences.

Multiple Groups (One-way ANOVA)

Simplest case is balanced design (equal sample size for each level, i.e. each group) with constant variance (same for different groups), but can adjust for both of these.
Multiple comparisons: often interested in comparisons between levels, e.g. which pairs are significant or best versus rest. Need to adjust significance level for multiple tests. Usually start by requiring that overall F-test is significant.

Multiple Regression

Model selection – dangers of "data mining" leading to models that are too complex (linnerand example) (average number of variables added by chance will be significance level times the number of variables tested, if we ignore correlations).

Diagnostics – residuals and leverage
Residuals help us identify model deficiencies (are the residuals Normal, is the variance constant, is there are a missing variable, do we need a transformation)

Look for patterns in residuals against row, against covariates for mean or variance
Row (sequence) pattern for residual suggests missing variable (time?) or a change over time.
Pattern for residual against covariate suggests need for products or powers of covariates.
Pattern in variance suggests need for transformation to stabilize variance (assumed constant in standard models)
Transformation – can try log, sqrt etc.

Leverage helps us identify high dependence of model fit on a few key data points (model not necessarily "wrong", but some risk and need to check).

Collinearity – problem with covariates that are collinear (i.e. can predict one covariate from others), so parameter estimates not well defined, or may change considerably depending on which other variables included.

R commander allows us to include polynomials and splines (smoothed polynomials) and offers numerical diagnostics and graphs to help check model assumptions.

Can use best subset to help identify which is the best model, given different criteria (AIC, BIC, Mallows Cp are good criteria that penalise complexity in different ways, adjusted R2 is not as good, R2 is useless unless all models have same complexity). Do not use stepwise, always prefer best subset, because stepwise may miss the best fitting models.

General Linear Models with R

GLM means Multiple Regression + Multi-way ANOVA combined, sometimes called Analysis of Covariance

Model selection is now much more complex as we need to select from a wide range of models. Aim at simple answers (small degrees of freedom for model) that fit well and that can be explained in terms of a sensible theory.

For multiple factors, can have many interactions, but we often assume that higher level interactions are likely to be (relatively) unimportant.

Generalized Linear Model with R

Sometimes called GLIM, to distinguish from GLM.

Extend GLM to data that does not follow the Normal distribution, but some other distribution, like Poisson, Gamma, Binomial

In the General Linear Model, we have that:
$E(y)=A+B\,x$ and $Y=E(y)+$ error term
Where the error term is from the Normal distribution with zero mean and constant variance.

For the Generalized Linear Model, we have that:
g(E(y))=A+B x, where g is called the link function and Y follows a specified distribution with mean E(y)

e.g. Y is count data, then the distribution is Poisson with parameter λ, E(y)=λ and g(y) is log(y)
e.g. Y is binary outcomes, i.e. success=1/failure=0 then the distribution is Binomial(1,p), E(y) is p, the probability of success and g(y) is log(y/(1-y)) = logit (y) or can use probit.

We can apply similar ideas to identify a good model that explains the variation in y

Please see the book "Generalized Linear Models" by Nelder and McCullagh for more technical details and examples.

Binary, Ordinal and Multinomial Data with R

Use logistic models, i.e. for 2 levels,
Log(Pr(level 2)/Pr(level 1))= linear model+error,

instead of y=linear model+error

i.e. expected log odds is linear in the factors and covariates (vs expected value for y is linear for general linear models).

As the linear model moves from large negative to large positive, we shift from
Pr(level 1) close to 1 and Pr(level 2) close to 0 to Pr(level 2) close to 1 and Pr(level 1) close to 0.
These models need to be solved iteratively, but the computer shields us from this.

We can extend these models for ordinal data to what is called ordinal logistic regression, which assumes a series of logistic regressions, which have the same slope but different intercepts.
As this model only adds one parameter per level, it is still feasible for large numbers of levels.

There is also an extension for nominal data called multinomial logit, which has many more parameters, because the slope is no longer assumed to be common across levels.

Compositional data with R

This is data where non-negative variables sum to a known constant – usually 100% or 1, such as when we have a set of proportions. In this case, correlations do not have the usual meaning because of the constraint (in the trivial case, if p and q add to 1, they must have a correlation of -1). If the variables are all within the range (0,1), i.e. there are no zeroes or ones, we can use a log ratio transformation that is meaningful as relative ratios are what matter. For example, if we have 3 components: (x1, x2, x3) that sum to 1, then we can look at

y1=log(x1/x3), and y2=log(x2/x3)

using the standard General Linear Model approach.

Note that we can map these log-ratios back to the original x space and that it does not matter which of the x's we use as a divisor for General Linear models. Please see John Aitchison's book: "The Statistical Analysis of Compositional Data" for more details of this approach.

Beyond R Commander

R offers many packages beyond those included in R Commander. They can be installed just like R Commander, inside R. However, some caution is advised as not all packages are well maintained. Usually, if there is an up-to-date package at the R repository, that is a good sign that this package is well maintained, but the documentation may not always be very clear.

23: Big Data

Learning objective: understand how to handle datasets which are too large to fit into memory on a typical computer including some idea of the risks when analysing large datasets

What is Big Data?

Big Data is often assumed to involve:

Volume

- Tall data (many records, N is large)
- Wide data (many variables, P is large)
- Model complexity (calculation usually increases linear in N, but may increase faster with P, for example may increase with NP^2 for some algorithms)

Variety

- Including non-standard data coding (such as text, audio or video – we have some text in our example datafiles)

Velocity

- Real-time data (may need to process data continuously, some of our datafiles are generated every day)

Veracity

- Can we trust the data (Reliability, validity, representativeness, see chapters 3 and 5 of the course book)

but we will mainly consider Volume, primarily N is large or NP is large.

Open Data

ODI-HKU Open Data documents
http://www.ssrc.hku.hk/open_data_main.php

3rd ODI conference (webinar)
Panel 1 on Spatial data and COVID:
https://youtube.com/playlist?list=PL23J327M1nExVRJWz_aGQEnsbHozXi_Hi

Panel 2 on Digital literacy and employment:
https://youtube.com/playlist?list=PL23J327M1nExp9AAP3KXo3eZZQFczBVME

Stand alone videos on Open Data
https://youtube.com/playlist?list=PL23J327M1nEy9gUwgv6gwNfGCvktwD0_r

Includes my video on "Safely Maximising Value from Publicly Funded Data" as part of the stand alone videos, which we will watch now so you understand the issues of obtaining access to datasets in Hong Kong.
https://youtu.be/s-WUpT3U0SQ

Kaggle open data (need to register)
https://www.kaggle.com/datasets
Includes lots of virus data now!

Hong Kong Government data portal

https://data.gov.hk/en/

This was not very good in the past, but it continues to improve, with many datasets now available using documented APIs, JSON or CSV. The help is quite useful:
https://data.gov.hk/en/help

JSON: means the data is JavaScript text-based objects
See https://www.digitalocean.com/community/tutorials/how-to-work-with-json-in-javascript

CSV means comma separated values (which can be read easily into almost any program). Good CSV files contain variables names in the first record – need to use quotes for any commas inside a field. Wikipedia is your friend for full details:

https://en.wikipedia.org/wiki/Comma-separated_values

Previously much of the data was in PDF files, in some cases, even image-based PDF files!

Hospital Authority

(very little Open Data, as there are privacy problems with making most hospital data open)
https://data.gov.hk/en-datasets/provider/hospital?order=name&file-content=no

They are now working on a Data Collaboration Lab (DCL), but you need an HA collaborator, you must attend a workshop with limited places, only government, university and NGO employees can be PI, but this is already better than the initial plan:

https://www3.ha.org.hk/data/DCL/Index/

HK Census 2016 data (build your own tables)

Major constraint is that you cannot produce very detailed tables (it will auto merge categories as you add variables), actually covers 2006, 2011, 2016 in theory, although much of the data is only for 2016, this is arguably not Big Data other than in terms of the total sample of about 0.7M (based on 10% sample of 7M) as you cannot download the individual records for privacy reasons.
<https://www.bycensus2016.gov.hk/en/bc-own_tbl.html>
[Will illustrate generating a table]

HK Census & Statistics microdata

This is secure enclave processing for academics to access all 0.7M records for 2016 Census and similarly for earlier census datasets, plus now added Household Expenditure Survey, General Household Survey, Birth/death/marriage records, Thematic Household Survey.
https://www.censtatd.gov.hk/service_desk/list/microdata/index.jsp

HK Language maps

Maps generated from 2011 microdata (which of 27 languages people claim to speak) linked to land boundary maps for 412 District Council Constituency Areas (DCCA). Two levels: districts and DCCA within districts. 2016 maps (covering oral and written languages) out soon
http://www.ssrc.hku.hk/hklangmaps/

HKU research data management policy

http://www.rss.hku.hk/integrity/research-data-records-management
https://hub.hku.hk/researchdata/rds.htm
This policy means that any datasets that you collect or generate will be archived in HKU, but will not be shared, unless you agree (unlike your thesis, which WILL be made public, possibly after a delay). I encourage you to make your datasets available in the HKU Scholar's Hub
http://hub.hku.hk/advanced-search?location=crisdataset
using the new HKU data hub:
https://datahub.hku.hk

UK Data management Policy

In contrast, we have the UK Research Council common principles, which start with:

Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner.

Institutional and project specific data management policies and plans should be in accordance with relevant standards and community best practice. Data with acknowledged long-term value should be preserved and remain accessible and usable for future research.

https://www.ukri.org/apply-for-funding/before-you-apply/your-responsibilities-if-you-get-funding/making-research-data-open/

Datasets for this chapter

The first dataset we will explore can be found here:
http://stat-computing.org/dataexpo/2009/the-data.html
It contains about 7M records per year on 29 core variables (from 1987-2019). I have downloaded the 2008 data (airline2008, zip file is 169MB, CSV file is 622MB) to illustrate. Integers unless stated, times are HHMM

V1: Year
V2: Month
V3: DayofMonth
V4: DayOfWeek
V5: DepTime
V6: CRSDepTime
V7: ArrTime
V8: CRSArrTime
V9: UniqueCarrier (text)
V10: FlightNum
V11: TailNum (text)
V12: ActualElapsedTime
V13: CRSElapsedTime
V14: AirTime
V15: ArrDelay
V16: DepDelay
V17: Origin (3 letter airport code)
V18: Dest
V19: Distance
V20: TaxiIn
V21: TaxiOut
V22: Cancelled
V23: CancellationCode (all empty)
V24: Diverted
V25: CarrierDelay
V26: WeatherDelay
V27: NASDelay
V28: SecurityDelay
V29:LateAircraftDelay

The second dataset we will explore (commodity_trade, can be found here: http://data.un.org/Explorer.aspx (it is also in Kaggle)
It is Commodity Trade Statistics – 10 variables and 7,874,347 rows

V1: country_or_area (text)
V2: year (integer)
V3: comm_code (integer)
V4: commodity (text)
V5: flow (Import/Export)
V6: trade_usd (integer)
V7: weight_kg (integer)
V8: quantity_name (text for units)
V9: quantity Integer
V10: category (text)

The third dataset about gun violence incidents in the US can be found here: https://www.gunviolencearchive.org and in Kaggle, the Kaggle version contains 29 columns and 231,753 records:
Text unless stated otherwise

V1: incident_id (integer)
V2: date (YYYY-MM-DD)
V3: state
V4: city_or_county
V5: address
V6: n_killed (integer)
V7: n_injured (integer)
V8: incident_url
V9: source_url
V10: incident_url_fields_missing
V11: congressional_district (integer)
V12: gun_stolen
V13: gun_type
V14: incident_characteristics
V15: latitude (float number)
V16: location_description
V17: longitude (float number)
V18: n_guns_involved (integer)
V19: notes
V20: participant_age
V21: participant_age_group
V22: participant_gender
V23: participant_name
V24: participant_relationship
V25: participant_status
V26: participant_type
V27: sources
V28: state_house_district (integer)
V29: state_senate_district (integer)

I also have a cleaned version, where the numbers in n_killed and n_injured have been extracted as integers.

Materials

I have put many pdf files explaining different elements of handling Big Data in R in the Google Drive folder.

Statistical tools

Primarily R plus add-ons (we will use Chapter 21,22 of the coursebook for the basics
Warning: bigglm.ffdf currently fails https://github.com/edwindj/ffbase/issues/61
but will hopefully be fixed soon

Conceptual arguments

Do we need to build models, or is correlation sufficient if we have enough data?

Google Flu Trends

Optimism in 2008/9:
https://www.youtube.com/watch?v=243TQ8zEs8A (news?)
https://www.youtube.com/watch?v=6111nS66Dpk (Google promo!)

Even had a paper in Nature

https://www.nature.com/articles/nature07634
with nearly 4,000 citations

2016 Still claiming, whoops!
https://www.youtube.com/watch?v=lEDt89eQ64o

Reality:
https://www.youtube.com/watch?v=X0XqnAqvyIk

Arguably, NOT just about overfitting, but also lack of validation and lack of understanding of correlation versus causation (review chapter 2 of coursebook on association and causation)

Statistical problems with tall and wide datasets

Tall datasets – almost everything becomes statistically significant (as we can detect very small effects when the sample size is large), need to distinguish what is practically significant. Can do equivalence test, where you do not look for zero difference, but a difference that is big enough to matter (practical significance, not just statistical significance).

How to select models? (look at discussion in p154 of coursebook, including AIC).

Wide datasets – selection bias where if you do enough tests, always find something significant, otherwise known as p-hacking. (look at example in p156 of coursebook)

Can use simple approach of Bonferroni adjustment (multiply significance by the number of tests to control the risk of any false rejections of the null hypothesis), but it is then very hard to detect effects when K is large (we say the power is low, meaning we are unlikely to detect that the null hypothesis is false). However, we can instead control the false discovery rate (proportion of discoveries which are false, where discovery means the null hypothesis is rejected) – this assumes that we care about the proportion of false rejections, rather than the number of false rejections.

The Benjamini–Hochberg procedure (BH step-up procedure) controls the False Discovery Rate at level alpha. It works as follows:

4. We have m null hypotheses tested and we list the m p-values in ascending order and denote them by $P_{(1)}$-$P_{(m)}$.
5. For a given alpha, find the largest k such that $P_{(k)}$ is less than k x alpha/m
6. Reject the null hypothesis (i.e., declare discoveries) for i=1 to k

Geometrically, this corresponds to plotting $P_{(k)}$ vs. k (on the y and x axes respectively), drawing the line through the origin with slope alpha/m, and declaring discoveries for all points on the left up to and including the last point that is below the line. This procedure is valid when the m tests are independent, and also in various scenarios of dependence, but is not universally valid.

(look briefly at example in the BHprocedure.pdf, to see how this works, see p104)

Lots of missing data – need to check if missing data is informative (e.g. mortgage data, records with missing income data may be a good indication of default), not just delete all record with missing data or replace all missing data with the average non-missing value, which may both introduce large bias (let's stop and think why large bias and why is it a problem?)

What is the non-statistical problem with large volume?

Large volume can mean

- Takes up too much disk space (many laptops only have a few hundred GB space on the SSD drive in total)
- Needs too much memory to process (some laptops only allow 2GB of memory, or may swap the memory out to the hard disk, which is very slow if not an SSD drive, R normally places the full dataset in memory, whereas SAS only keeps summaries in memory).
- Too slow to process (depends on the sophistication of the analysis, descriptive analysis is always quick, non-linear model fitting may take considerable time to converge in an iterative estimation process, similarly simulations may take a long time to generate sufficient estimates, i.e. this depends on model complexity)
- Too many records to process (R has a 2 billion record limit)

Simple possible solutions

Select variables that are needed from the full set of variables

Sample rows (systematic sample is easy, but may bias the results if the data order is informative and we do not estimate using samples that encompass the full dataset, may be problematic if the number of variables is large as most statistical procedures assume n > > p)

Intermediate difficulty

Use R packages (e.g. ffdf object ff package) that can keep the data on disk and read it in a chunk at a time, but still give access to most statistical tools (e.g. biglm in ffbase package).

Use R packages (e.g. data.table) that reorganize the data in a more structured way that speeds up calculations (may not save memory), but still give access to most statistical tools.

Buy/borrow a more powerful computer (expensive if just for one-off task, but may be worth it otherwise) that has more memory (up to 1TB) and/or large SSD drive (up to 10TB).

Using ff/ffbase/biglm

We need to start by installing R first (see Chapter 20 of the coursebook)

Windows Security:

1. C:\Program Files\R - and this is the folder that contains "R" and not RStudio folder!
2. Right click for properties.
3. Security.
4. In "Group or user names" select your name
5. click "Edit"
6. select "Full control"
7. Apply and click

Mac – may need to switch to English during installation.
Need to add ffbase (includes ff) and biglm libraries.

```
install.packages("ffbase", dependencies = TRUE)
install.packages("biglm", dependencies = TRUE)
```

We will download the 3 large datafiles we want to use and unzip it into a folder (Note: the trade file is about 1.3GB when unzipped!):
Commodity_trade.csv.zip
Airline2008.csv.zip
Gun_violence.csv.zip

Then open and store the data in a binary file format that R can access and that the tools in ffbase can use. Ffbase needs a directory where it can store a copy of your datafile, as it is not stored in memory, unlike R.

It is easiest to change the working directory in R (Menu: Misc/Change Working Directory) to use a directory that contains your datafiles and that you have permissions to write files.

#look at pdf on using ffbase

Note: If you get this error:

Error in setwd(dfile): cannot change working directory

You must type this in R, where the path is to a folder that you have permissions to use:
options(fftempdir = "path/to/your/folder")

```
#load the libraries we need
library("ffbase")      #includes ff
library("biglm")


#setwd - go to the correct directory if we do not want to type paths, best to use R menu
```

```
#trade data 7.9M x 29
#country_or_area,year,comm_code,commodity,flow,trade_usd,weight_kg,category etc.
trade<-read.csv.ffdf(file="commodity_trade.csv",header=TRUE,
        first.rows=10000, next.rows=50000, colClasses=NA)
```

In windows, if your intention is to simply open the file then you might as well directly drag & drop the file onto the open app window.
Otherwise if what you want to do is navigate in the Open File dialog to the required folder, what I do is simply Alt+Tab to Explorer, Alt+D and Ctrl+C to focus the address bar and copy the path, Alt+Tab back to the dialog and Ctrl+V.

In Windows any \ in the path needs to become \\

On Mac, drag the file into the R Window to get the path. Any ":" in the path becomes "/"

```
#save ffdf file in zip format
pack.ffdf(file="trade.zip",trade)
# unpack does the reverse

#check we have an ffdf object
class(trade)

#how many rows and columns?
dim(trade)

#what have we got?
```

```
names(trade)

#look at distribution
tabulate.ff(trade$year)
quantile.ff(trade$year)

# histogram
hist(trade$year)

#similarly for other numerical variables in trade

#load model fit for big data
#includes linear and generalised linear models

library(biglm)

#now try and fit a simple model - normally use lm()
model1<-biglm(quantity ~ year, data = trade)
summary(model1)
coef(model1)
deviance(model1) #residual SS
AIC(model1)
model2<-biglm(trade_usd ~ year, data = trade)
summary(model2)
coef(model2)
deviance(model2) #residual SS
AIC(model2)
```

```
#2008 airline data 7M x 29
#Year,Month,DayOfWeek,DepTime,UniqueCarrier,AirTime,ArrDelay,DepDelay,Distance etc.

air08<-read.csv.ffdf(file="airline2008.csv",header=TRUE,
        first.rows=10000, next.rows=50000, colClasses=NA)

#save ffdf file in zip format
pack.ffdf(file="air08.zip",air08)

modela<-biglm(ArrTime~as.character(Month),data=air08)
summary(modela)

modelb<-biglm(ArrTime~as.character(Month)+Distance,data=air08)
summary(modelb)

modelc<-biglm(ArrTime~as.character(Month)+Distance+as.character(DayOfWeek),
        data=air08)
summary(modelc)

modeld<-biglm(CRSElapsedTime~as.character(Month)+Distance
              +as.character(DayOfWeek),data=air08)
summary(modeld)
AIC(modeld)
```

```
#gun_violence data 230k x 29
#data about gun violence in the US in 2013, including date, state, latitude, longitude,n_killed,n_injured

gun<-read.csv.ffdf(file="gun_violence.csv",header=TRUE,
        first.rows=10000, next.rows=50000, colClasses=NA)

#save ffdf file in zip format
pack.ffdf(file="gun.zip",gun)

tabulate.ff(gun$n_killed)

#bigger class of models - normally use glm
#can model poisson, Gamma, binomial, gaussian
#can specify link e.g. family = binomial(link=probit)
```
Warning: bigglm does not work properly in R 4.x at all, see https://github.com/edwindj/ffbase/issues/61
And it seems not to work in R 3.x currently for ffdf, as the example here:
https://rdrr.io/cran/ffbase/man/bigglm.ffdf.html
does not work with 3.x or 4.x!

```
#model3<-bigglm(n_killed~longitude+latitude, family=poisson(), data = gun)
#model4<-bigglm(n_killed~longitude, family=poisson(), data = gun)
#model5<-bigglm(n_killed~longitude+latitude+longitude*latitude
        +I(longitude^2)+I(latitude^2), family=poisson(), data = gun)
#model6<-bigglm(n_killed~state, family=poisson(), data = gun)
#model7<-bigglm(n_injured ~ n_killed, family = poisson(), data = gun)
#model8<-bigglm(n_injured ~ log(n_killed+1), family = poisson(), data = gun)
```

So we illustrate using biglm, without the Poisson model, for now, using log(count+1) as the dependent variable instead (like using log Normal instead of Poisson, as an approximation)

```
#linear in x and y coordinates
model3<-biglm(log(n_killed+1)~longitude+latitude, data = gun)
AIC(model3)
#linear in x corrdinate only
model4<-biglm(log(n_killed+1)~longitude, data = gun)
AIC(model4)
#quadratic in x and y coordinates
model5<-biglm(log(n_killed+1)~longitude+latitude+longitude*latitude
        +I(longitude^2)+I(latitude^2), data = gun)
AIC(model5)
#depend on State instead (which have different gun laws)
model6<-biglm(log(n_killed+1)~state, data = gun)
AIC(model6)
#model injured depends on killed
model7j<-biglm(log(n_injured+1) ~ n_killed, data = gun)
AIC(model7j)
# for Poisson model, should normally model using log(n_killed+1) instead as this is model for log of mean
model8j<-biglm(log(n_injured+1) ~ log(n_killed+1), data = gun)
summary(model8j)
AIC(model8j)
```

More difficult solutions

Can we break down the analysis by variables or rows? Store the data in an SQL database and only extract the rows and columns needed for each analysis (assumes that this reduces the size sufficiently). Can use dplyr which has compiled code to speed up the break apart and join work, if you have the data stored in a database.
https://www.youtube.com/watch?v=aywFompr1F4
Switch from R to SAS (expensive, time consuming and may not have the capabilities we need)

What if the dataset is still too big?

Split the work across a cluster of computers (needs new software tools to map the work across the cluster and reduce it back together which may use Python instead of R, access to a cluster such as HPC in HKU or Amazon or Google). See ITS tools for using Python (pdf from ITS website). Python is also useful for data cleaning (such as the module Pandas)

Hadoop in 5 minutes:
https://www.youtube.com/watch?v=aReuLtY0YMI

Map, filter and reduce using Python
https://www.youtube.com/watch?v=hUes6y2b--0

Data science using Python
https://www.youtube.com/watch?v=cUw3DsDpQCE

24: Hierarchical Linear Models

Learning Objective: how to use Bayesian Hierarchical Linear Models to model data that vary at more than one level.

Why HLM:

Hierarchical Linear Models (HLM), otherwise known as multilevel models, mixed models or random-effects models are used to model data that vary at more than one level.

For example, data from cluster sampling, such as students in classes in schools, where students, classes and schools would be the three levels. These models allow us to understand variation at the different levels and model the impact of factors and covariates at the different levels.

Simple linear models assume that all observations are independent and have fixed variance, which is not true in our school example, as students in the same class or in the same school share common effects. Using simple linear models on the students or class averages is flawed because it does not account for the common variation and will be biased. Using simple linear models on school averages does not allow us to account for characteristics of the students, such as age, baseline ability at entry and parental characteristics, which may vary greatly across students and schools, distorting the school level relationships.

HLM means that the model for the expected value may have factors and covariates at each level (which is simple to account for) and the error term in our model may have elements from each level of our hierarchy, which is more complex to account for in the estimation and hypothesis testing process.

HLM can also be used for longitudinal data, where we have repeated measurements on the same individuals, so those repeated measurements are not independent because of common individual variation, so we again distinguish between the individual level and the measurement occasion for that individual level.

For example, a simple linear regression model assuming Gaussian errors and repeated measurements on individuals (a two level model) where the regression coefficients may be different for different individuals, would look like this:

$Y_{ij} = beta_{0j} + beta_{1j} X_{ij} + e_j + e_{ij}$
j is the jth individual and ij means the ith measurement on the jth individual
e_j is the error due to variation across individuals
e_{ij} is the error due to variation across measurements of an individual
e_j and e_{ij} both follow independent Normal distributions with different variances (so e_j is in common for the errors of measurements on the same individual)

Statistical inference approaches for linear hierarchical models:

Traditional statistical approaches, based on MVUE (minimum variance unbiased statistical estimators), are problematic as they model variance components (from different sources) and the estimated variance components can be negative, which is clearly impossible.

The two approaches we will consider are

Maximum Likelihood (ML) and Bayesian

In both cases, we need first to be able to state what the probability density function is (if our dependent variable is categorical, this changes to needing the probability mass function).

We will start with a simple model with constant mean and independent Gaussian (or Normal) errors:

The probability density for each data point (Gaussian density) is proportional to

$$\frac{\exp(-1/2((y-\theta)/\sigma)^2)}{\sigma}$$

where theta is the mean and sigma is the standard deviation

Maximum Likelihood (ML)

Likelihood is the overall probability for the whole data set, so

$$L(\text{theta} \mid \text{data}) = p(\text{data} \mid \text{theta}) \text{ is proportional to } \frac{\exp(-1/2\sum((y-\text{theta})/\text{sigma})^2)}{\text{sigma}^n}$$

We prefer to look at the log likelihood, which is quadratic in theta for Gaussian errors.

$$\text{LogL}(\text{theta} \mid \text{data}) = \text{constant} - n \log(\text{sigma}) - 1/2\sum((y-\text{theta})/\text{sigma})^2$$

ML requires finding the maximum, often found using simple calculus (assuming there is a unique maximum inside the boundary), by setting the first derivative to zero and solve:

$\frac{d\ \text{LogL}}{d\ \text{theta}} = \sum(y-\text{theta})/\text{sigma}^2 = 0$, means the estimate of theta must be the mean of y.

$\frac{d\ \text{LogL}}{d\ \text{sigma}} = -n/\text{sigma} + \sum(y-\text{theta})^2/\text{sigma}^3 = 0$, means the estimate of sigma must be RSS/n, which

is biased - the unbiased estimate is RSS/(n-1), where RSS is $\sum(y-\text{mean of } y)^2$

We can also find the approximate variance of the estimate by looking at the inverse of minus the second derivative of LogL:

$$\frac{d2\ LogL}{d2\ theta} = -n/sigma^2$$

so the estimated variance for theta = $sigma^2/n$

which is exact for this case.

Unfortunately, for small samples, ML can be highly biased, especially if the number of parameters is large (RSS/n versus RSS/(n-p)) or the maximum is at a limit (e.g. sigma=0). This can be improved using restricted maximum likelihood (REML) or bootstrap adjustment, but that is beyond our scope here.

Also, for hierarchical models, ML does not work very well, as shown by the James-Stein estimator:

We extend our simple case above to one where Y follows a Normal distribution with constant variance, with n measurements from each of m groups, each of which has a different mean.

ML would suggest using the group sample means as the best estimate of the group population means.

However, shrinkage of the group estimates towards the overall mean is superior, i.e.

group mean estimate = overall sample mean + k (group sample mean - overall sample mean)

where $k = 1 - \frac{(m-3)\,\text{sigma}^2}{\sum (\text{group mean} - \text{overall mean})^2}$

is provably superior (in terms of lower mean squared error) to the group sample mean if m>=4!

This is called a shrinkage estimator. As we will see next, the Bayesian approach naturally yields something similar to this estimator.

Bayes rule and inference

Bayesian inference starts from Bayes rule:

$$p(\text{theta} \mid \text{data}) = \frac{p(\text{data} \mid \text{theta})\, p(\text{theta})}{p(\text{data})}$$

as p(data) is a constant and p(data | theta) is the likelihood L(theta | data) (notice that we have switched from a function of the data conditional on theta to a function of theta conditional on

the data, this depends on the Likelihood principle, which states that inference about theta should only depend on L(theta | data)), we get

p(theta | data) is proportional to L(theta | data) p(theta)

Notice that this assumes we can make probability statements about theta, which cannot be interpreted as long-run relative frequencies, but only as subjective degree of belief, which has been criticized as unscientific, but this approach enables us to quantify all the elements of uncertainty. In other words, we can apply probabilities not only to observable outcomes, but also to parameters in a model, where there are no long-run relative frequencies.

We talk about this as Bayesian updating from the prior distribution (before seeing the new evidence) using the likelihood to get the posterior distribution, which incorporates the new evidence.

The conceptual challenge is how to construct the prior distribution before seeing any evidence, but in practice, we can use a widely spread out distribution, such as Gaussian with large variance.

Go back to our simple model:

$\text{LogL}(\text{theta} \mid \text{data}) = \text{constant} - n \log(\text{sigma}) - 1/2 \sum((y-\text{theta})/\text{sigma})^2$

If we use the term precision to mean the inverse of the variance, then the precision for each data point is $1/\text{sigma}^2$ and the sample precision, Pn is n/sigma^2

If our prior distribution for theta is Gaussian with mean theta_0 and precision $P0=1/\ \text{sigma}_0^2$ ($\text{variance}=\text{sigma}_0^2$)

Then we easily show that the posterior distribution for theta is also Gaussian with

$$\text{posterior mean} = \frac{P0\ \text{theta}_0 + Pn\ \text{sample mean}}{P0 + Pn}$$

posterior precision=P0+Pn,

so if P0 is very small, the answer is similar to ML, but note that the posterior mean shrinks the sample mean towards theta_0

If we build a Bayesian HLM for the group example above and assume that

the group means are a Gaussian sample from the prior mean,

$\text{theta}_j \sim N(\text{theta}_0, \text{sigma}_0^2)$

then the group means are shrunk towards the overall sample mean and these estimates outperform ML estimates (and provide a rationale for James-Stein type estimates).

Bayesian methods are naturally hierarchical as regards the parameters, so we can have second order priors for $theta_0$ and $sigma_0^2$

In short, there is a compelling empirical case that Bayes methods are superior for HLM in terms of providing better estimates (in fact, all admissible procedures are either Bayes methods or limits of Bayes methods, where admissible means I cannot find a procedure that is always better).

As we shall see later, Bayesian methods also allow us to easily replace the Gaussian assumption with other distributions, including discrete distributions, allowing many different sensitivity analyses to be considered.

Comparison of classical, ML and Bayes in simple situation

Let's look at a simple coin tossing example (Heads = Success, Tails=Failure) of the effect of different inference approaches (not HLM yet):

Our statistical model is Binomial data – observe r successes out of n trials with chance theta of success, so

LogL(theta | r,n) = constant (does not depend on theta) + r log(theta) + (n-r) log(1-theta)

If we have a prior for theta of Beta(a,b), i.e. proportional to $\text{theta}^{(a-1)}(1-\text{theta})^{(b-1)}$

then the posterior for theta is Beta(a+r,b+n-r)

This is what we call a conjugate prior as the posterior is in the same class of distributions as the prior. This is simple and does not need any sophisticated software, but if I sample successes from k different individuals, who have different theta, then I have a hierarchical model and the analysis gets more complex.

What is our best estimator for theta, the chance of success?

Bayes posterior mean is (a+r)/(n+a+b) if we use the Beta(a,b) prior for theta - which must be >0 and <1 as long as a,b>0. This shrinks the estimate towards a/(a+b), the prior mean.

ML estimate is r/n, even though that may estimate theta as 0 or 1 if r=0 or n, an estimate which is wrong unless we have a 2 headed or 2 tailed coin.

Classical (MVUE) cannot answer the question without knowing what the stopping rule was for the experiment we did (this is irrelevant for ML and Bayes).

If n is fixed, we have Binomial (r is random, n is fixed), the sufficient statistic is r and the best estimator is r/n, same as MLE.

If r is fixed, we have Negative Binomial (r is fixed, n is random), the sufficient statistic is k (# of failures) and the best estimator is (r-1)/(r+k-1) = (r-1)/(n-1), so ML estimate is biased in this case.

Marginal Posterior Distribution

There is one important practical problem with Bayesian methods that we have not yet discussed:

Bayes formula provides the joint posterior distribution for all parameters that we are estimating. How do we obtain the marginal posterior distribution for specified parameters? In some simple cases, we can do exact symbolic integration. For example, if we use an inverse gamma prior distribution for the variance (i.e. gamma prior for the precision) and Gaussian prior for the mean, we can show that the marginal posterior for the mean is a Student-T distribution and the marginal posterior for the variance is an inverse gamma (i.e. gamma posterior for the precision).

Markov Chain Monte Carlo (MCMC)

However, in many cases, we cannot do exact symbolic integration and must use numeric integration. Unfortunately, precise numerical integration takes a long time if the number of parameters is large (i.e. much greater than 10) as the amount of calculation increases in

general exponentially with the number of parameters, so we need to rely on Monte Carlo integration (i.e. sampling). Monte Carlo sampling works by allowing us to calculate

$$E(g(X)) = \frac{\sum g(x_i)}{N},$$

where x_i is the ith sample out of N from g(x), and this is not very precise unless we generate a very large number of points to ensure that the sampling error is small.

Fortunately, there was a major advance in the 1980s known as Markov Chain Monte Carlo (MCMC), which uses sampling in a much more efficient way to generate marginal posterior distributions and any expectations we might be interested in.

What is a Markov Chain? A Markov chain (MC) is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. For example, if we want to model the chance of rain tomorrow given the history of rain on previous days, it assumes the chance depends only whether it rained today. MC is crucial to the mathematics behind MCMC)

The key requirement is that we can write the statistical model as a full joint probability distribution for all parameters and observable data. This is always possible for parametric models and distributions, but often there is conditional independence between many elements of the model, which greatly simplifies the calculations. MCMC involves sampling in turn from

each conditional distribution (this is called using the Gibbs sampler, for which there are alternatives that we will not discuss here).[3]

One very helpful aspect is that sensitivity analysis is often simple to implement by generalizing the model, for example, if we replace a Gaussian by a T distribution, which includes Gaussian as a special case, where the degrees of freedom are large.

This MCMC sampling takes advantage of the fact that if we know the conditional posterior densities for each parameter (this is the Markov Chain part) and use them to generate sequential samples, then we converge (under quite weak conditions) to the correct posterior density, yielding all the marginal posterior distributions as an outcome. In many situations, we can easily find those conditional densities if we have a parametric model.

The biggest challenge is arguably to confirm convergence – have we converged yet, or are we still moving very slowly across the parameter space? The technical term is that our Markov chain should be stationary in order that we obtain the correct marginal posterior distributions. One basic approach is to run parallel simulations (chains) starting from different initial values and see if they all converge to the same region and compare the variability within chains to

---

[3] Alan E. Gelfand & Adrian F. M. Smith (1990) Sampling-Based Approaches to Calculating Marginal Densities, Journal of the American Statistical Association,85:410, 398-409, DOI: 10.1080/01621459.1990.10476213

variability across chains. This fits well with using a cluster of high performance computers. We can often improve convergence by re-parameterizing our model such that the parameters have low correlation.

This was originally implemented in software called WinBUGS, which only runs under Windows and is no longer developed, so we will use open source software called Just Another Gibbs Sampler (JAGS), that was completely rewritten but uses the same model description language (known as BUGS), so it can solve all the many examples developed for WinBUGS, including many hierarchical models.

The sampler needs to be run long enough to ensure that we have converged on the posterior distribution (so we have completed what is called "burn-in") first. Each new iteration then produces a set of estimates for all the parameters, taken from the posterior distribution, so we will run for sufficient iterations to generate sufficient sampling precision in the posterior summaries of interest, such as posterior means, posterior standard deviations or marginal posterior distribution densities.

JAGS has several R interfaces called rjags, runjags, R2jags, jagsUI but to save the time ensuring everyone has R correctly installed, we will instead work through some examples in JAGS directly. The R interfaces include a package called Coda to check convergence, which is probably the most challenging element of MCMC.

MCMC References

JAGS manuals:
https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/

JAGS examples:
https://sourceforge.net/projects/mcmc-jags/files/Examples/4.x/

JAGS software:
https://sourceforge.net/projects/mcmc-jags/files/JAGS/4.x/

BUGS book:
https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-the-bugs-book/

Windows installer hint: The path to the installer must contain only latin letters, no non-English symbols, such as Chinese.

BUGS language

BUGS language understands scalars and arrays, including subsets of arrays and simple for loop structures

~ is used for a stochastic (random) equation and <- for a deterministic equation

Most distributions have a d prefix to their name, e.g. dnorm, dgamma

Normal distributions are parameterised using mean and precision (=1/variance)

There are 2 contexts, model and data

BUGS Linear Regression example

A simple model example for simple linear regression is:
$Y_i \sim \text{Normal}(\alpha + \beta (x_i - \text{xbar}), \tau)$

Priors are:
$\alpha \sim \text{Normal}(0, 0.0001)$
$\beta \sim \text{Normal}(0, 0.0001)$
$\tau \sim \text{Gamma}(0.001, 0.001)$

```
model {
        for (i in 1:N) {
                        Y[i] ~ dnorm(mu[i], tau)
                        mu[i] <- alpha + beta * (x[i] - x.bar)
                        }
        x.bar <- mean(x)
        alpha ~ dnorm(0.0, 1.0E-4)
        beta ~ dnorm(0.0, 1.0E-4)
        sigma <- 1.0/sqrt(tau)
        tau ~ dgamma(1.0E-3, 1.0E-3)
}
```

Notes: x is centered to reduce the correlation in the estimates for alpha and beta, sigma is there purely for understanding the results

BUGS data

The data part can be very simple and data is stored in R format.

This means that scalars are written as:

"theta" <- 0.1243

vectors are written as:

"x" <- c(2, 3.5, 4.0e-3)

matrices are more tricky as we need to tell BUGS the structure, such as for a 3 x 2 matrix A

"A" <- structure(c(1,2,3,4,5,6), .Dim=c(3,2))

Our example of a simple linear model would need a data section like:

"N"<- 8
"x"<c(1,2,3,4,5,6,7,8)
"Y"<-c(2,5,7,9,11,13,15,16)

It is also possible to simulate data inside BUGS like this:

```
data {
        for (i in 1:N) { y[i] ~ dnorm(0,1)}
}
```

or transform data

```
data {
        for (i in 1:N) { z[i] <- sqrt(y[i])}
}
```

My script for this in linreg.cmd is:

```
model in "linreg.bug"
data in "linreg-data.R"
compile, nchains(2)
inits in "linreg-init.R"
initialize
update 1000
monitor alpha, thin(10)
monitor beta, thin(10)
monitor tau, thin(10)
update 1000
coda *
```

JAGS output in CODA format

The standard CODA format for monitors is to create an
index file and one or more output files.

The index file is has three columns with on each line:
1. A string giving the name of the (scalar) value being recorded
2. The row number of the first line for this value in the output file(s)
3. The row number of the last line for this value in the output file(s)

The output file(s) contain two columns:
1. The iteration number
2. The value at that iteration

If we do not have R (and the package coda) we can look at these values in Excel to check for convergence. If we run at least 2 chains, we can check that they converge to similar values.

Let's now compare the results from this simple Bayesian model with the results from a standard statistical package (JMP) (I do not expect you to do this example yourselves, I will show you this example first).

HLM example with repeated measurements and random effects.

This is an example that can be found in the ANOVA chapter of this coursebook (see Chapter 12). There are 2 species, each of which has 3 animals, which are each measured in the 4 seasons to see how far they travel. Clearly the animals are a random effect within species, and we have repeated measurements of each animal. We want to understand whether the species travel different distances and whether the distances travelled vary by season. For simplicity here, we will ignore the possibility of interactions between species and season, although it is straightforward to add that element.

$Y_{ijk} = \text{Normal}(\beta_{ij} + \gamma_k, \tau)$

where β_{ij} is the average distance for animal j of species i and γ_k is the effect due to season k.

The model can be expressed as

```
model {
        for (i in 1:N) {
         for (j in 1:M) {
         for (k in 1:S) {
                        Y[k,j,i] ~ dnorm(mu[k,j,i], tau);
                        mu[k,j,i] <- beta[i,j] + gamma[k];
                        }
                }
        }
```

```
        for (i in 1:N){alpha[i] ~ dnorm(mu0,taua);}
        for (k in 1:S) {gamma[k] ~ dnorm(0,taug);}
        for (i in 1:N) {for (j in 1:M) {beta[i,j] ~ dnorm(alpha[i],taub);}}
        mu0 ~ dnorm(5,1.0e-3)
        tau ~ dgamma(1.0E-3, 1.0E-3)
        taua ~ dgamma(1.0E-3, 1.0E-3)
        taub ~ dgamma(1.0E-3, 1.0E-3)
        taug ~ dgamma(1.0E-3, 1.0E-3)
    }
```

Our data is:

```
    "N"<-2
    "M"<-3
    "S"<-4
    "Y"<-structure(c(5,3,0,0,5,4,3,1,6,2,4,3,7,8,4,2,6,6,5,4,8,9,7,5),.Dim=c(4,3,2))
```

Our script is (animals1.cmd)

```
model in "animals.bug"
data in "animals-data.R"
compile, nchains(2)
inits in "animals-init.R"
initialize
update 5000
```

```
monitor alpha, thin (100)
monitor gamma, thin(100)
monitor tau, thin(100)
update 10000
coda *
```

which outputs data on the species means, the seasonal effects and the residual variability. However, we can generate new comparisons, such as compare the species means and ask the question what is the average difference in distance between the species.

Before running the JAGS/WinBUGS examples, we first change our directory of the folder containing the example files (cd changes directory (use double quotes for the pathname), dir lists files from inside JAGS, pwd to show current directory)

For my examples:

cd “Documents/Course materials/HLM workshop/my examples”

For where JAGS installed standard examples, so I can easily access the example files:

cd "/Volumes/Macintosh HD/Applications/JAGS/examples/classic-bugs"

HLM Gaussian regression example

This is the rats example originally used for WinBUGS in the Volume 1 set but now also included in JAGS, from Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling by Alan E. Gelfand, Susan E. Hills, Amy Racine-Poon and Adrian F. M. Smith, in Journal of the American Statistical Association, Vol. 85, No. 412 (Dec., 1990), pp. 972-985. This example is taken from section 6 of the paper and concerns 30 young rats whose weights were measured weekly for five weeks.

The model is a random effects linear growth curve where the intercept and slope depend on the rat and x_i is the day (8,15,22,29,36) of the measurement.

$Y_{ij} \sim \text{Normal}(\alpha_i + \beta_i (x_j - \text{xbar}), \tau_c)$
$\alpha_i \sim \text{Normal}(\alpha_c, \tau_\alpha)$
$\beta_i \sim \text{Normal}(\beta_c, \tau_\beta)$

where xbar = 22, and τ represents the precision (1/variance) of a normal distribution.

In BUGS language, this is expressed as:

```
model {
   for (i in 1:N) {
      for (j in 1:T) {
         mu[i,j] <- alpha[i] + beta[i]*(x[j] - x.bar);
         Y[i,j]   ~ dnorm(mu[i,j],tau.c)
      }
      alpha[i] ~ dnorm(alpha.c,tau.alpha);
      beta[i]  ~ dnorm(beta.c,tau.beta);
   }
   alpha.c   ~ dnorm(0,1.0E-4);
   beta.c    ~ dnorm(0,1.0E-4);
   tau.c     ~ dgamma(1.0E-3,1.0E-3);
   tau.alpha ~ dgamma(1.0E-3,1.0E-3);
   tau.beta  ~ dgamma(1.0E-3,1.0E-3);
   sigma    <- 1.0/sqrt(tau.c);
   x.bar    <- mean(x[]);
   alpha0   <- alpha.c - beta.c*x.bar;
}
```

JAGS script example

We will run this for 1,000 updates to converge and then 10,000 more to yield estimates using this JAGS script (test1.cmd):

```
model in "rats.bug"
data in "rats-data.R"
compile, nchains(2)
inits in "rats-init.R"
initialize
update 1000
monitor alpha0
monitor beta.c
update 10000
coda *
```

HLM logistic regression example

This is the seeds example from Volume 1, using data from Table 3 of Crowder, M. J. 1978. Beta-binomial Anova for proportions, Journal of the Royal Statistical Society (C, Applied Statistics) 27(1), 34–37.

This example concerns the proportion of seeds that germinated on each of 21 plates arranged according to a 2 by 2 factorial layout by seed and type of root extract.

The model is a random effects logistic, allowing for over-dispersion. If p_i is the probability of germination on the ith plate, we assume

$r_i \sim \text{Binomial}(p_i , n_i)$

$\text{logit}(p_i) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_{12} x_{1i} x_{2i} + b_i$

where x_{1i} , x_{2i} are the seed type and root extract of the ith plate, and an interaction term

$\alpha_{12} x_{1i} x_{2i}$ is included.

$b_i \sim \text{Normal}(0, \tau)$ is the random effect unrelated to the seed and root

α_0 , α_1 , α_2 , α_{12} , τ are given independent priors with small precision.

This yields this model in BUGS:

```
model {
   alpha0  ~ dnorm(0.0,1.0E-6);  # intercept
   alpha1  ~ dnorm(0.0,1.0E-6);  # seed coeff
   alpha2  ~ dnorm(0.0,1.0E-6);  # extract coeff
   alpha12 ~ dnorm(0.0,1.0E-6); # interaction coeff
   tau     ~ dgamma(1.0E-3,1.0E-3);    # 1/sigma^2
   sigma  <- 1.0/sqrt(tau);
   for (i in 1:N) {
      b[i]  ~ dnorm(0.0,tau);
      logit(p[i]) <- alpha0 + alpha1*x1[i] + alpha2*x2[i] + alpha12*x1[i]*x2[i] + b[i];
      r[i]  ~ dbin(p[i],n[i]);
   }
}
```

Again we will run 1000 times to "burn-in", then 10,000 times to get estimates

School hierarchical model with pupil and school covariates

This is the schools example from Volume 2, using data from “A Multilevel Analysis of School Examination Results”, by Harvey Goldstein; Jon Rasbash; Min Yang; Geoffrey Woodhouse; Huiqi Pan; Desmond Nuttall; Sally Thomas, Oxford Review of Education, Vol. 19, No. 4. (1993), pp. 425-433 (pdf in folder).

They present an analysis of examination results from inner London schools. They use hierarchical or multilevel models to study the between-school variation and calculate school-level residuals in an attempt to differentiate between 'good' and 'bad' schools.

Data:
Standardized mean examination scores (Y) were available for 1978 pupils from 38 different schools. The median number of pupils per school was 48, with a range of 1--198. Pupil-level covariates included gender plus a standardized London Reading Test (LRT) score and a verbal reasoning (VR) test category (1, 2 or 3, where 1 represents the highest ability group) measured when each child was aged 11. Each school was classified by gender intake (all girls, all boys or mixed) and denomination (Church of England, Roman Catholic, State school or other); these were used as categorical school-level covariates.

Model:
We consider the following model, which essentially corresponds to Goldstein et al.'s model 1.

$Y_{ij} \sim \text{Normal}(\mu_{ij}, \tau_{ij})$

$\mu_{ij}=\alpha_{1j}+\alpha_{2j}\,\text{LRT}_{ij}+\alpha_{3j}\text{VR1}_{ij}+\beta_1\text{LRT}_{ij}^{\,2}+\beta_2\,\text{VR}^2_{ij}+\beta_3\,\text{Girl}_{ij}+\beta_4\,\text{Girls'school}_j+\beta_5\,\text{Boys'school}_j+\beta_6\,\text{CEschool}_j+\beta_7\,\text{RCschool}_j+\beta_8\,\text{other school}_j$

$\log\tau_{ij} = \theta + \varphi\,\text{LRT ij}$

for the ith pupil in the jth school.

We wish to specify a regression model for the variance components, and here we model the logarithm of τ ij (the inverse of the between-pupil variance) as a linear function of each pupil's LRT score. This differs from Goldstein et al.'s model which allows the variance σ 2 ij to depend linearly on LRT. However, such a parameterization may lead to negative estimates of $\sigma^2_{\ ij}$.

This yields this model:

```
model {

for (p in 1 : N) {
Y[p] ~ dnorm(mu[p], tau[p])
mu[p] <- alpha[school[p], 1] + alpha[school[p], 2] * LRT[p] + alpha[school[p], 3] * VR[p, 1] +
beta[1] * LRT2[p] + beta[2] * VR[p, 2] +       beta[3] * Gender[p] + beta[4] * School.gender[p, 1]
+ beta[5] * School.gender[p, 2] + beta[6] * School.denom[p, 1] +
       beta[7] * School.denom[p, 2] + beta[8] * School.denom[p, 3]
log(tau[p]) <- theta + phi * LRT[p]
sigma2[p] <- 1 / tau[p]
LRT2[p] <- LRT[p] * LRT[p]

}
min.var <- exp(-(theta + phi * (-34.6193))) # lowest LRT score = -34.6193
max.var <- exp(-(theta + phi * (37.3807))) # highest LRT score = 37.3807
```

Priors for fixed effects:

```
for (k in 1 : 8) {
beta[k] ~ dnorm(0.0, 0.0001) }
theta ~ dnorm(0.0, 0.0001)
phi ~ dnorm(0.0, 0.0001)
```

```
# Priors for random coefficients:
for (j in 1 : M) {
       alpha[j, 1 : 3] ~ dmnorm(gamma[1:3 ], T[1:3 ,1:3 ]);
       alpha1[j] <- alpha[j,1] }

# Hyper-priors:
gamma[1 : 3] ~ dmnorm(mn[1:3 ], prec[1:3 ,1:3 ]);
T[1 : 3, 1 : 3 ] ~ dwish(R[1:3 ,1:3 ], 3)}

#   Compute ranks:
  for (j in 1:M) {
     for (k in 1:M) {
        greater.than[j,k] <- step(alpha[k,1] - alpha[j,1]);
     }
     rank[j] <- sum(greater.than[j,]);  # rank of school j
  }
}
```

Note that alpha1[] measures the 'residual effect' for school j after adjusting for pupil- and school-level covariates. This might represent an appropriate quantity by which to rank schools' performance in terms of how good the results are after accounting for the language ability of the students. We can calculate the ranks for each complete iteration and hence find the mean and s.d. of the rank for each school.

Inhalers (ordered categorical repeated measures)

This example is in Volume 1, from Ezzet and Whitehead (1993), who analyse data from a two-treatment, two-period crossover trial to compare 2 inhalation devices for delivering the drug salbutamol in 286 asthma patients. Patients were asked to rate the clarity of leaflet instructions accompanying each device, using a 4-point ordinal scale.

The response R_{it} from the i_{th} subject (i = 1 to 286) in the t_{th} period (t = 1,2) thus assumes integer values between 1 and 4. It may be expressed in terms of a continuous latent variable Y is taking values on (-inf, inf) as follows:

$$R_{it} = j \text{ if } Y \text{ is in } [a_{j-1} , a_j), j = 1,..,4$$

where a_0 = -inf and a_4 = inf. Assuming a logistic distribution with mean m_{it} for Yit, then the cumulative probability Q_{itj} of subject i rating the treatment in period t as category j or worse (i.e. Prob(Y it >= a j) is given by

$$\text{logit } Q_{itj} = -(a_j + m_{sit} + b_i)$$

where b_i represents the random effect for subject i. Here, m_{sit} depends only on the period t and the sequence s_i = 1,2 to which patient i belongs. It is defined as

$$m_{11} = b / 2 + p / 2$$

$m_{12} = - b / 2 - p / 2 - k$
$m_{21} = - b / 2 + p / 2$
$m_{22} = b / 2 - p / 2 + k$

where b represents the treatment effect, p represents the period effect and k represents the carryover effect. The probability of subject i giving response j in period t is thus given by:

$p_{itj} = Q_{itj} - 1 - Q_{itj}$, where $Q_{it0} = 1$ and $Q_{it4} = 0$

The BUGS language for this model is shown below. We assume the b_i 's to be normally distributed with zero mean and common precision t. The fixed effects b, p and k are given vague normal priors, as are the unknown cut points a_1, a_2 and a_3. We also impose order constraints on the latter using the I(,) notation in BUGS (T() in JAGS as this is truncation, not censoring), to ensure that $a_1 < a_2 < a_3$.

```
model
{
#
# Construct individual response data from contingency table
#
  for (i in 1 : Ncum[1, 1]) {
    group[i] <- 1
    for (t in 1 : T) { response[i, t] <- pattern[1, t] }
```

```
  }
  for (i in (Ncum[1,1] + 1) : Ncum[1, 2]) {
    group[i] <- 2 for (t in 1 : T) { response[i, t] <- pattern[1, t] }
  }

  for (k in 2 : Npattern) {
    for(i in (Ncum[k - 1, 2] + 1) : Ncum[k, 1]) {
      group[i] <- 1 for (t in 1 : T) { response[i, t] <- pattern[k, t] }
    }
    for(i in (Ncum[k, 1] + 1) : Ncum[k, 2]) {
      group[i] <- 2 for (t in 1 : T) { response[i, t] <- pattern[k, t] }
    }
  }
#
# Model
#
  for (i in 1 : N) {
    for (t in 1 : T) {
      for (j in 1 : Ncut) {
#
# Cumulative probability of worse response than j
#
        logit(Q[i, t, j]) <- -(a[j] + mu[group[i], t] + b[i])
```

```
        }
#
# Probability of response = j
#
        p[i, t, 1] <- 1 - Q[i, t, 1]
        for (j in 2 : Ncut) { p[i, t, j] <- Q[i, t, j - 1] - Q[i, t, j] }
        p[i, t, (Ncut+1)] <- Q[i, t, Ncut]

        response[i, t] ~ dcat(p[i, t, ])
        culmative.response[i, t] <- culmative(response[i, t], response[i, t])
      }
#
# Subject (random) effects
#
      b[i] ~ dnorm(0.0, tau)
}

#
# Fixed effects
#
   for (g in 1 : G) {
      for(t in 1 : T) {
# logistic mean for group i in period t
```

```
        mu[g, t] <- beta * treat[g, t] / 2 + pi * period[g, t] / 2 + kappa * carry[g, t]
      }
    }
    beta ~ dnorm(0, 1.0E-06)
    pi ~ dnorm(0, 1.0E-06)
    kappa ~ dnorm(0, 1.0E-06)

  # ordered cut points for underlying continuous latent variable
    a[1] ~ dflat()T(-1000, a[2])
    a[2] ~ dflat()T(a[1], a[3])
    a[3] ~ dflat()T(a[2], 1000)

    tau ~ dgamma(0.001, 0.001)
    sigma <- sqrt(1 / tau)
    log.sigma <- log(sigma)

  }
```

Note that the data is read into BUGS in the original contingency table format to economize on space and effort. The individual responses for each of the 286 patients are then constructed within BUGS.

Other BUGS examples

There are many other examples in the 3 volumes of WinBUGS example pdfs and full JAGS sample files for Volumes 1 and 2.

Bayesian Model Choice

We can always look at the posterior distribution for any parameter or combination (e.g. contrast) of parameters, to see the evidence that it might be zero or small.

Existing criteria, which link to -2 x max log Likelihood (see Chapter 11)

$AIC = -2 \log\{p(y|\theta)\} + 2k$ (evaluated at the ML estimate of θ)

$BIC = -2 \log\{p(y|\theta)\} + k \log n$

By loose analogy (this is still being argued in the literature),
$DIC = D(\theta) + 2\, p_d$

Where $D(\theta)$ is the deviance, i.e. $-2 \log\{p(y|\theta)\}$ evaluated at the posterior mean (instead of the ML estimate of θ) and p_d is the effective number of parameters (this is tricky as in a hierarchical model, the effective number of parameters is less than the actual number). p_d can be estimated as the difference between the posterior mean for $D(\theta)$ and $D(\theta)$ evaluated at the posterior mean for θ. JAGS lets us monitor DIC (see the pdfs for more information).

Related tools

Instead of using rjags, runjags or R2jags, you can also consider NIMBLE, which is like JAGS operating completely inside R

Stan

If you are trying to model something beyond what JAGS can manage (e.g. JAGS is too slow or you need a distribution not included in JAGS), the best tool is probably the open source software called Stan, which is more flexible, but requires stronger programming skills. Stan is more efficient as it uses automatic differentiation of the log density to use a more efficient sampler (Hamiltonian sampler instead of the Gibbs sampler), but cannot handle categorical or missing data easily.

URL: http://mc-stan.org

For example, my PhD student was modelling compositional data, i.e. p1+p2+p3=1 and p1,p2,p3>0. This is modelled using log ratios, i.e. log(p1/p3) and log(p2/p3) follow a bivariate Gaussian distribution. She wanted to examine censoring of p1,p2,p3 and this cannot be done in JAGS as the censoring option in JAGS can only be applied to the sampled distributions (i.e. the log ratios), not on p1,p2,p3 directly. In Stan we could find the censored log-density by integrating numerically over the censoring regions. Convergence was slow, but this was the correctly formulated problem (versus censoring of the log-ratios).

25 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,

https://www.gradsch.hku.hk/gradsch/current-students/courses-workshops-dialogues-career-preparation/supporting-courses-services/statistical-advice-centre-for-students-stacs#

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