Hello World!

Elementary Statistics in the Cognitive Sciences

Christer Johansson

January 9, 2017

Chapter 1

Introduction

1.1 Unlearning

Why is it so hard to grasp elementary statistics? After all, the basic calculations are not very hard, and statistics are presented everywhere. Especially the media seems very keen on presenting statistics, but often the wrong kind of statistics or misleading statistics. In my experience, it is sometimes necessary to unlearn what people have learned before taking a statistics class.

One thing many students struggle with is that numbers are not just numbers. You cannot meaningfully calculate averages of all kind of numbers. The average of my hotel number and your hotel number is not typically the room number in-between our rooms, as room numbers code and labels the rooms but they do not necessarily count them. The average number of children might be 1.2 children, but no family has that number of children. What is twice as cold as 0 degrees? What is twice as loud as a 100-decibel sound? We may even have to invent a scale of comparison, or operationalize our question such that a question is available for measurement. What is the difference in attitude between two groups of people? Such a question would demand that we first figure out what measurements would be relevant. We very often assume properties of numbers that might not be valid. We assume that we can add, subtract, multiply and divide all kinds of numbers in a meaningful way, which means that we assume a linear scale of measurement. This book will start with a discussion of data qualities and the operations that we could perform on different kinds of data. Another difficulty is the idea of how large a difference is, if we are interested in differences. Our language allows us to discuss anything as large — a proton is larger than an electron but they are both quite small compared to an elephant — but we are really interested in a comparison to expectations. Is the difference bigger than ex-

pected? What could our expectations be based on? How do we know that an observed difference is not due to measurement errors? Is the difference large enough to be relevant? We will look into expectations that are based on an assumption of statistical independence. We will also look into expectations from model assumptions, such as assuming a certain kind of distribution of measurements, for example *the normal distribution*.

How do we know that our observations are representative of the phenomenon we want to measure? This is a problem that could be relevant when we observe people. How do we know that those people that volunteer for our experiment are representative of the population we want to model and investigate? Are we aware of the difference between an individual and a model description of an idealized average individual? The latter questions could have to do with our habits of concrete thinking, whereas the statistical mode of operation is to think about generalizations and models that have some validity at a group level. That is, we would like to say something about characteristics of groups and populations, and in particular we use models to reason about these objects. We need these simplifications to reason about tendencies and differences. Remember, to notice a distance we need at least two points of reference. In order to notice a difference from expectations with more certainty we need even more data points. If we have a very large number of observations, we might notice, with significance, small differences that are not relevant t our research question. For example, if a large survey discovers, with significance, that male students relative to female students, on average get 10% more attention from their teachers, how would that translate to a class of 5 male students and 8 female students? Could the teacher detect the difference, and separate it from individual factors? Should we worry, and call for change?

This book begins with a general introduction to data levels, a measure of central tendency and some basic operations that are safe for each data level. Each data level will have tests associated with it, and the following chapters will look closer at some tests for each data level. In the start we will look into tests that assume simple contrasts. We will then extend our toolbox with tests that can handle several contrasts, and compare how factors interact with each other. For each test we are interested in how the data was sampled. Did we get repeated measures for each individual subject? Did we present multiple test items to the same subject? This book will suggest some simple experiment designs, and how to evaluate independent measures and repeated measures. The last chapters will look into mixed effect models, where we earlier have assumed fixed effects from planned contrasts.

1.2 Why do we need statistics?

Why not just trust what we see? Statistics is often thought to represent a "proof" of final verdict on an issue. Trust the statistics. The main focus on this book is the scientific use of statistics. One starting point is that we have an abstract hypothesis that states what we ought to detect as a measurable difference of a specific manipulation compared to a baseline without the manipulation, if the hypothesis is true. Observation alone will not prove our hypothesis, but if we do not observe any reliable difference then we should not claim that our hypothesis is true. However, we should bear in mind that a non-observation of a difference can happen for many reasons. One reason is simply that the difference is too small for us to observe with certainty.

Science is always very cautious with claiming proof of a hypothesis. If someone claims that observations are proof of the hypothesis then you should be very cautious. There are so many ways thinking could be flawed. You should beware of claiming *correlation as causation*, and using *appeal to authority figures*, or arguments based on *jumping on the bandwagon* such as claiming that if "most scientists" *believe* something then it must be true. This often happens when there is a new finding. For example, a new medical condition that is discussed in the media will very often be over-diagnosed, simply because staff that is inexperienced with the condition will look for positive signs of the condition, but also under-diagnosed because staff that do not know about the condition cannot look for it. That is, the precision of diagnosis will initially be bad, and the detection of real cases will simultaneously be bad.

If something is interesting, it is very often not decided. If we know something to be a fact then there would be no need for testing in the first place. This is an important lesson that will be learned by experience: there is often more uncertainty to any phenomena than is popularly believed. Many so-called facts may not be certain, or may be open to interaction with factors that are not properly controlled in our model. Sometimes the important factors are not even known.

You will very easily find lists of common logical fallacies that will make it look like all arguments are flawed. Remember then that the *fallacy-fallacy* could apply, i.e. just because you detect a logical fallacy does not mean that the conclusion is necessarily wrong, but rather that the conclusion does not follow from the flawed argumentation.

As scientists, we like to separate arguments of belief from arguments based on facts, observation and logical or physical necessity. This may sound trivial, but whenever the debate get heated we all have a tendency to defend our beliefs and our pet theories. This is another fallacy called *confirmation bias*.

We tend to more easily believe information that confirms our hypothesis, and we easily ignore information that goes against our hypothesis. We should therefore carefully design experiments that have a chance to challenge our initial belief. This is often done using *null hypothesis* testing. We try to discard the hypothesis that our experimental manipulation gives the same result as the baseline without our manipulation. The null hypothesis actually has a slightly better chance to stand, as it stands if we cannot detect a difference. As I mentioned before, one way to *not* detect a difference is to have a too small experiment to be able to detect a relevant difference (or simply not looking at all). It should be obvious that not being able to discard the null hypothesis is *not* proof of the null hypothesis, but rather that we have not detected any support to the contrary. If we do have a very large number of observations, and we cannot find evidence against the null hypothesis, then we should acknowledge that we are not very likely to find any difference, but we might look for the wrong thing, and we might look at the wrong place. It is less obvious that if we do detect a difference, this is not proof that our manipulation is responsible for the difference. First of all, differences can occur at "random" (i.e. for no obvious reason). This is the case that significance testing will help us avoid. The often cited *"p-value"* is a quantification of how often we would detect a difference as large (or larger) than the one we have evidence for in our data *if* the difference is due to the observed variability in the data. This is a common model, where we estimate two parameters: the central values and the variability. There are other models, for example based on probability of occurrence, or probability of ranking. It should be clear that there are many other possible sources of errors that are beyond the estimation of variability. For example, we need to have representative samples. Most often, we only have *convenience samples* in some sense. Much of modern psychology is built on experimental results based on *an analysis of subjects* that are college students, of a specific time, and often at large elite universities. We often use a small set of test items, and it is also questionable if these items are representative of the full range of the phenomena that is investigated. There will almost always be factors that are specific to an experiment, for example room temperature, ambient light, the news in the media of the day, the staff in the lab etc.

It may sound like it is a hopeless task to gain any new knowledge. However, the message is that it takes more than one experiment to gain knowledge. It is always better to try to get experimental evidence than to rely solely on speculation. The value of a published experiment is not so much that it will prove anything, but rather that it will challenge the scientific community to replicate the findings, and to design better versions of the experiment. The better experiments could include more relevant factors for

investigation, and/or collect much more data. Very often the first experiments are small, and uncertain. If the results are important then there is a need to replicate with more subjects. *Publication bias* is grounded in that small experiments only have a chance of publication if they show significance. However, as soon as a study is published someone else could make a better or larger experiment that should be published also if it shows no significance. Actually, it would be even more important to publish those studies if they show no significance. In an ideal case, publication bias is not a constant preference for statistical significance. The initial small studies have a higher demand on significance, but the next studies would have to argue that they are better studies and then non-significance is more interesting, although a larger confirmatory study is also desirable. You could make a whole career out of testing published smaller studies, if you find more funding.

Another caveat is that statistical significance does not mean relevant, or interesting or even easy to detect. Small and real differences could be detected in a large enough study, but if those differences really matters must be judged carefully, and often in relation to how we intend to use those differences, and of course in relation to the research question.

In conclusion, we need statistics to see if we have made an observation that cannot be explained by just random fluctuations. It is a ticket to conference presentations and publications. The study design, the research question and the quality of the study is often more important for relevance. A better and larger study that fails to show significance should be more interesting than a small and badly designed study that shows significance. However, research has to start somewhere and often the start is in small and non-perfect studies that show a significant difference that is well argued to be relevant to some interesting research question. There is likely many published results that are either non-interesting, or actually false, but if the researchers had an honest intent to find out about some phenomena that was deemed worthy of investigation and the efforts of data collection, analysis, argumentation and all the steps towards publication then we should remember that the worse alternative is not testing the hypothesis at all. We should give credit to the great efforts that go into coming up with a hypothesis, planning an experiment, and testing the hypothesis. However, since we could expect that there are many false results out there, we should remain skeptical as readers and be careful how we use new speculative findings, and see how we can challenge the results by finding ways to improve the experiments. We should not fall in the trap of dogmatic belief in "science", real science is never dogmatic. The whole point with science is that you do not need to believe in science — either it is true or it is not. Most of the time we find out that the issue is more complicated than we first thought or heard, and

this is an important step on any journey towards knowledge. Respect for complexity and alternatives do not make good sound bites. Sometimes it is easier to argue with less knowledge if you only need to persuade and not be concerned with truth. That is part of why the world is a mess. Human progress comes in small steps or big leaps – evolution rather than smooth continuity, and at any time old facts may turn out to be false and replaced by better explanations.

1.3 Failure to replicate

Lately there has been a lot of attention in psychology to the failure to replicate many classic results. It is disturbing, but maybe not so surprising. It should also be noted that no study is better than its design. It is much easier to make a design error that will make the null hypothesis stand, as it stands by default. If the study is conducted by scientists outside of the field in a non-optimal laboratory for the purpose and just following the brief instructions from published studies then it could indeed be hard to detect differences. The replication studies are not immune to experimental errors. That means that some of the studies that failed to replicate a result may in fact be false negatives. It can be quite hard to match the control group and the experimental group on all parameters. Many studies does this by using very specific groups of subjects that are matched for age, education and background in an almost automatic fashion. It is also very common to make matched repetitive measures on the same subjects, so that the subjects are their own control. It is hard to find a better match than the actual subject. Another factor is the importance of variance. The subjects would ideally understand the task in exactly the same way. Thus the instructions need to be very clear and detailed, and at the same time brief. Good instructions have a potential to lower variance in the groups, and thus make it possible to detect smaller effects with significance. The subjects could also be trained to perform the task optimally. This is typically done in two different ways. First, the typical student population might be used to taking tests, and many are in fact used to participating in experiments. Familiarity with the task lead to less hesitation and better fluency in the task. It is also common to let the subjects practice on some examples, before the actual experiment, using items that are not used in the actual experiment. One way of doing this is to let the first experimental presentations be items that lead in to the actual experiment. Sometimes, the first items will also provide some feedback so that the subject has a very clear understanding of the task both in the abstract from instructions and in practice. Yet another

factor is to make the context and environment as similar as possible between subjects, and this can be done in a professional laboratory with experienced staff. There is also a possibility to use the context and environment as an experimental factor and plan for testing the ***ecological validity*** of the experiment. In theory, a scientific experiment should be possible to conduct by anyone with access to the correct equipment. However, in real life the experience of the researchers may play a role. It would not be very hard for inexperienced, or *malicious* (i.e. researchers that are invested in the null hypothesis), researchers to increase the variance in an experiment to a level that would make it very difficult to detect differences with significance, even if the differences are real.

1.4 Systematic errors of computation

The analysis software will want to use floating point processors for handling data. However, computers are quite bad at representing decimals exactly. In R you can set the number of decimals you want to present.

```
> options(digits=22)
```

Let us see the consequences. If you apply the inverse of a function to a function of x you get back the value of x. For example squaring a square root of x gives x, and taking the square root of a square also gives x. Thus $\sqrt{(0.1+0.2)^2}$ ought to be the same as $(\sqrt{(0.1+0.2)})^2$ which is the same as $\sqrt{(0.1+0.2)(0.1+0.2)}$. Try this in R:

```
> (sqrt((0.1 + 0.2))^2) ==  (sqrt((0.1 + 0.2)))^2
[1] TRUE
> (sqrt((0.1 + 0.2))^2) -  (sqrt((0.1 + 0.2)))^2
[1] 0

> (sqrt((0.1 + 0.2))^2) ==  (sqrt((0.1 + 0.2)*(0.1 + 0.2)))
[1] FALSE
> (sqrt((0.1 + 0.2))^2) -  sqrt((0.1 + 0.2)*(0.1 + 0.2))
[1] 5.5511151231257827021 18e-17
```

This shows that the order of operations are important, and even though there is a formal equivalence the computer's internal representation could introduce errors that are in no way random, and therefore will not cancel out. This means that small errors can accumulate if you add them up.

```
> 100000000000000000*( (sqrt((0.1 + 0.2))^2) -
                       (sqrt((0.1 + 0.2)*(0.1 + 0.2))) )
[1] 5.551115123125782702118
```

We use computers because they are very good at repetitive tasks. The fact that computers do not represent results as exactly as we would like to think should make use beware of for example aggregating data in some ways. For example, simply accumulating differences between data points a large number of times could very well introduce significant numerical errors. In the example above, the numeric error alone is summed up by a large multiplier (equivalent to summing up the same error that many times), and the result that should be 0 can in fact be any number depending on how many times we sum up.

This will typically not be a big problem for our small studies, but a complicated model that involves calculating millions of differences and solving millions of linear equations will give rise to some spurious results that stem just from the numerical errors in the computations. The programmers are aware of the problem, and often incorporate tests that warn the user if there is a suspicion that the results could be affected by numerical inaccuracies. However, programmers are people, and it is easy to make "mistakes", especially as the abstract mathematical representation indicates that there is no problem.

When you get more advanced you will probably encounter warning messages when you work with complicated models, and you will encounter messages that indicate that your model is not stable or has failed to converge. You should think of other possibilities to formulate the model that you want, and you should think about if your model could be too complicated for the data that you analyze. You should also think about if it is possible that errors accumulate in your model, or if they will cancel out. If your model is too complicated you might not have enough data to estimates the effects that you are interested in. If you have too much data, then you should be aware that accumulating data might introduce numerical errors.

1.5 Summary

The intention of this book is awake some interest in statistical analysis, and to go through step by step some fairly easy tests. After reading the book you should be more aware of the process of statistical analysis, and perhaps have an interest to learn more about the R software. R is an evolving project, where new useful modules are added every year. This functionality also

makes it essential to think about models more generally and match your understanding of the data you have, or are going to collect, with the right model. Sometimes it is not the most complicated model that is the best model for your project. Most of the time, you should consider models that you can easily motivate and that you know your audience will recognize. This book will help you with those steps, by making a step-by-step introduction to a selection of some very commonly used tests. As your experience grows and your problems get more specific you will find more useful modules, or even get interested in programming in R to customize your analyzes or build them into an automatized analysis for tasks that you often perform.

Chapter 2

Data Levels

2.1 Introduction

In order to determine which test is appropriate for the task at hand, it is often necessary to think about what kind of data the investigation will collect, or has collected. What kinds of operations are natural on that kind of data? What is a good measure of central tendency for the data?

2.1.1 Measures of Central Tendency

Some measures of central tendency are: *typical value*, *median*, and the *mean*. All of these measures are associated with different expectation on the quality of data, and which operations can be performed to compare data.

2.1.2 Typical value or most frequent value

Typical value is the most frequent value. We might also think of which item is most frequent. For example, if we want to investigate parts-of-speech in a language, we may want to find the most typical part-of-speech (POS) in a context. This can actually help produce a simple and efficient parts-of-speech tagger, simply from noting which parts-of-speech typically occur in the lexical context at hand. To find the best POS for the word 'man', in the context "Man the boat" we might look in a very large tagged corpus for the most frequent POS that occurs between a space (or a punctuation mark) and *the*, and if we find *boat* after *the* as well, we have frequency data for the longest context in the example, we might even find the exact phrase including the word *"Man"*. This strategy often produces surprisingly accurate results from just look up in a data collection.

Other examples: what is the most typical number of children in a family? My guess is around two for Scandinavia. Why would the typical value be a good statistic? The mean would suggest that the distribution is fairly symmetric around the mean, and that the distribution is fairly smooth with a tapering off on either side. This is unlikely, since the different numbers of children in families would likely be biased to just a few numbers: 0,1,2,3 and the rest will be increasingly uncommon. The mean will most likely not be an integer either, so the mean would point to a number that no family has.

The typical value is useful in the case that there are a few distinct choices. It is most often applied when we count categories (such as names, labels, etc.), i.e., a data type we might call nominal, which is the name for data based on the process of naming and counting frequencies of such labels.

2.1.3 The Median, the value in the middle of the range

The median is the middle value in a range. The operation we need to get the median is to be able to sort the data. The median is a very robust measure that is not affected by the size of the measurements we are sorting. If you suspect that there are outliers in your data. An outlier is a value that is so far from expectations that there is an increased risk that something was wrong with that measurement. Technically outliers can also be thought of as the data points that are outside of a (95%) confidence interval. If you do not want to delete your outliers, or you are unsure of the distribution of your data, then the median might be the choice even when you can motivate a data level better than ordinal data. The median is very useful when we cannot expect normally distributed data with a clear central tendency and a smooth and symmetric distribution around a central value.

2.1.4 The Mean, a constructed expectation

The mean is the value you get after summing up all the data, and dividing the sum by the number of observations. To calculate the mean you must assume that you have a scale to measure on. This means that a unit of measurement should be worth the same wherever on the scale. The mean is the measure to prefer when we expect data from a normal distribution, and we can confirm that our sample is compatible with our expectation.

2.2 Types of Data, and their uses

2.2.1 Nominal data

Uses: Categorization, naming
Observation: Difference in kind
Operations: counting frequencies of occurrence
Central tendency: typical value
Tests: *chi-square* test, *cross tables* tests

This kind of data consists of names, labels, category names etc. The labels might sometimes be numbered (indexed), but without the assumption that one index can be compared in size to the other. Collections of answers to yes/no questions can be considered gathering nominal data, and may be used to sort a population into bins. A category such as "sex" have (at least) two values: male and female; these values might in turn be coded as 1 and 2, but do not make the mistake of calculating the average of male and female. There might be even more levels to a category, apart from the common dichotomies involving just two levels. If the category levels are ordered then they have more information and belong with ordinal data. The makes of cars, assuming they are not ordered, is an example of nominal data for the car domain. Nominals are often used to structure data; to create little neat boxes to put our measurements in. For the nominals themselves, there is not much more to do than count the frequencies of occurrence. If you have nominal data, and want to make tests on the distribution of nominal data, then most likely you will use a chi-square test, or a cross tabulation test.

2.2.2 Ordinal data

Uses: Categorization, order
Observation: Difference in degree
Operations: sorting
Central tendency: median
Tests: Wilcoxon, Mann-Whitney U. These are paired and unpaired versions of the same test as implemented in the **R: wilcox.test**

This data type usually consists of positions on a hierarchical scale. The crucial property of ordinal data is that it is possible to order the data. Typical uses are data from questionaries, where the subjects are asked to rate, for example, their intuitions on some phenomenon. How likely are you to use the word "swell" : 1 never 2 very unlikely 3 somewhat 4 very likely 5 all the time.

Ordinal data may arise from races, where the positions the subjects end up are recorded. Academic grades is another often quoted example of ordinal data; it is better to get an A than a B. But is it always true that an A guarantees better quality than a B? Disregarding errors of measurement, and assignment of characters, the person with an A is expected to know more about the subject at hand, than the person with a B, at the point of giving the characters all else the same. It is very rare that all else is the same. If the characters were given for different courses, by different academic institutions, or different teachers, then there is no absolute guarantee that the order holds true for a particular individual. For ordinal data there is no information on how much difference there is between two levels, as we do not know how much each interval (e.g., between A and B) is worth. We do expect that there is a difference, and we know the direction of the expected difference.

2.2.3 Interval data

Uses: Categorization, order, equal intervals
Observation: Difference in amount
Operations: sorting, addition, subtraction
Central tendency: mean
Tests: t-test

This data level assumes a scale where the units are worth the same wherever on the scale. The temperature is one such measure. One degree Celsius is worth the same whether it is freezing or boiling. Children sometimes discover that there is something wrong with these kinds of data. What is twice as warm as 0 degrees? Interval data has an arbitrary zero value, and the zero does not indicate absolute absence of whatever is measured.

2.2.4 Quotient (Ratio) data

Uses: Categorization, order, equal intervals, absolute zero
Observation: Difference in amount, ratio to comparison
Operations: sorting, addition, subtraction, multiplication, and division
Central tendency: mean
Tests: t-test

Ratio data has everything that holds for interval data, but has an absolute zero. This means that 0 (zero) means absolute absence of what is measured.

This is a necessary property to be able to compare amounts in ratios such as *twice as much*. Typical measures are length, duration, weight etc. If it makes sense to add and subtract, and 0 means that there is complete absence of the property your are measuring, and the measurement has been achieved, then you may assign quotient data.

Alternative interpretations for a 0 are that it is a number on an arbitrary scale (0 on the centigrade scale indicates some temperature, which is different from a complete lack of heat), or 0 could indicate that there was no measurement at all, i.e. the value is unknown, which is a completely different matter. The researcher has to check the data and be careful about the meaning of the figures.

2.2.5 Type hierarchy

Note that this is a hierarchy, and everything you can do at a lower level can be done at the higher levels too. In order to meaningfully compare amounts you need at least interval data. If there are doubts about the data level, or prerequisites for a test such as approximate normal distribution (required for t-tests), then you may back down to tests that assume a less complex data level, for example from interval to ordinal; meaning you may test using a non-parametric test instead. However, you should always consider the highest possible data level, *but no higher*, as the increased information available at each higher level get more out of the data than what is available at the lower levels.

There are several advantages of models that use interval data or better. These models are often more familiar, more powerful, and more flexible and easier to handle. The researcher might be satisfied if only the most critical assumption of normally distributed data can be supported. This would demand that the study has a fair amount of observations. The assumption of normality can be formally tested provided that enough data is available. The advice is to think hard about the data level and the assumptions of each test before performing the test.

2.3 How to choose a test

There are three considerations for choosing a test. 1) What is the relevant measure of central tendency? 2) What is the data level? (I.e. can we assume approximate normal distribution? and 3) How representative of your population is your sample?

Tests that are recommended at the higher data levels (i.e., t-tests) should

be preferred if possible, as these tests can use the size of the measurements, and all the extra informativeness that comes with the more complex data level. This typically gives better chances to discover significant differences. However, if there are doubts about the mean being a good measure of central tendencies, then consider using tests relevant to ordinal data despite the formal capacity to use a potentially more powerful test.

Doubts concerning the mean may include presence of outliers (unexplained extreme values) in your data, or finding obvious deviations from the expectations of normal distribution; for example finding severely skewed distributions, or multiple peaks in the data indicating bimodal or multimodal data or mixing of different statistical populations. Such doubts are indications that your model might not be an appropriate model for your data.

When we are dealing with differences between two groups (or one group and a fixed value) there is no good excuse for not choosing a more conservative test, with fewer assumptions, if you have serious doubts. There could be serious doubts on the data level, the demands on distribution, or the presence of extreme values. Such doubts are especially serious if the number of measurements is low, as this makes it more difficult to know the distribution of data.

Non-parametric tests, such as the R wilcox.test, do not lose that much power over a t-test, and since there are no parameters to estimate there is less that can go wrong.

However, parametric tests do often provide more power and use more of the information in the data, so you should use parametric tests when they are appropriate. The test you use should be motivated, so if you use a parametric tests write explicitly that you assume an approximate normal distribution, and that the data level you have is appropriate for the test.

When you have a lot of data (close to 100 measurements and more in each group) and many variables, and levels within variables, or many groups then it is often convenient to use tests that demand at least interval data even on ordinal data. This use of testing is often explorative, and more rigorous testing will be applied, in new experiments, when the interesting variables have been identified. It might not be the best practice, but people like tests that they are familiar with, and that they know the reviewers will be familiar with. If you choose to disregard the data level, it is even more important that you check that the demands on approximate normal distribution are fulfilled.

2.4 Summary

Analyzing what kind of data you are dealing with is very important for choosing the appropriate test. There is no universal test that will fit all kinds of data. There are assumptions about the data, and about what is relevant to test, implicit in all statistical testing. You should aim to make your assumptions as clear and explicit as possible.

Chapter 3

Tests for nominal data

3.1 Nominal tests of proportion

One model for countable nominal data assumes we can compare the observed frequencies to the expected frequencies. We can test if the proportions are different from expectations. In its simplest form, we have two groups and we see if the expected proportions. The expectation might be equal proportions; but we may state explicitly any expected proportions. Is the data explained by the frequencies expected if there is no consistent effect?

3.1.1 Simple Chi-square test

When we have nominal data there is little else we can do than count the frequencies of each label. However, we can test if the frequencies deviate from expectations. One common expectation that may be our starting point is the expectation of equal proportions in each category.

The test to use is the χ^2 test (R: chisq.test). This test relies on calculating the differences between observed frequencies and expected frequencies, square each difference and divide by the expected frequency, and finally sum up these numbers and look up the critical value for significance. The critical value depends on the degrees of freedom (**df**), which is a concept that essentially tells how many of our frequency counts are free to vary given that we know how many observations in total there is to distribute. In the case of a *cross tables analysis*, the row totals and the column totals of the contingency table we want to analyze are important.

$\chi^2 = \sum_i \frac{(O-E)^2}{E}$ where E is the expected number of occurrences (frequency), and O is the observed frequency.

Let us start with a simple example. Say that we have observed that

there are slightly more male children than female children born. From experience we have estimated that the proportion is 51% boys, and 49% girls, at birth. The analysis assumes two categories that are mutually exclusive for all practical purposes and sum up to 100%, meaning that we assume no other categories for our analysis. This assumption should be motivated by the researcher, and stand in relation to the goals of the study.

If we know there is a total of 100%, then from knowing that boys are 51%, we also know that girls are 49% by simple subtraction, thus the degree of freedom (df) is 1: We only have to know one proportion, in order to calculate the other. In general, when we have a situation with N mutually exclusive categories that cover all our measurements, there are $N-1$ degrees of freedom. We will later see the calculations for a cross tables analysis. In this example we have 1 df. Note that the degrees of freedom in this test depend on the number of variables (categories) and not on how many observations we have made.

The test is based on summing up the deviance from expected frequencies, and then looking up in a table if that sum of evidence is enough to say that the observed distribution cannot be explained by random fluctuations. For example, we may look up in a table what the critical value is for the χ^2 test with one degree of freedom; it is around 3.8 if we accept significance at $p < 0.05$. This means a 5% (i.e. 1 in 20) risk of falsely rejecting our null hypothesis when the null hypothesis is actually true and all our assumptions about the statistical model are assessed correctly. Setting the critical value at 4 will give an extra margin; if the evidence is at least 4, it certainly is better than 3.8, which is the *critical value*.

Let us go through the calculations of the χ^2 test. We have a total of X observations. For our null hypothesis we expect half of these observations to be male births, and half to be female births. Our alternative hypothesis is that there are in fact 51% male births, and 49% female births.

The difference is 1% more male births, and 1% less female births, and in frequencies that means $\frac{(0.51X-0.5X)^2}{0.5X} + \frac{(0.49X-0.5X)^2}{0.5X}$. After some simplification, we get $4\frac{(0.01X)^2}{X} = 4\frac{X}{10000}$; and for this to reach the critical value of 4, X needs to be 10000. Thus to claim significance for this one percent difference from the expected frequencies we need to observe 10000 births.

This would be the analytical way of using the definition formula to calculate how many observations are necessary. We could try a numerical solution in R. You may use R, and type in the function calls to the **chisq.test**. First we guess that it could be done a thousand observations. Note that we can easily display the expected and observed frequencies, as well as performing

the actual test. R will look in the tables for us. **Hint**: you can use the arrow keys to retrieve previously entered lines of code, and then change that line. It could be very useful when you perform tests where you only change some small part in each line.

```
> chisq.test(c(male=1000*0.51,female=1000*0.49))$expected
male  female
500     500
> chisq.test(c(male=1000*0.51,female=1000*0.49))$observed
male  female
510    490
> chisq.test(c(male=1000*0.51,female=1000*0.49))
data:  c(male = 1000 * 0.51, female = 1000 * 0.49)
X-squared = 0.4, df = 1, p-value = 0.5271
```

The above illustrates how to get the expected frequencies out from the test. Since we used a formula in the input it is useful to get the observed frequencies as well, just as a sanity check that the formula has been calculated. The test shows no significance. Let us double the number of observations.

```
> chisq.test(c(male=2000*0.51,female=2000*0.49))
Chi-squared test for given probabilities
data:  c(male = 2000 * 0.51, female = 2000 * 0.49)
X-squared = 0.8, df = 1, p-value = 0.3711
```

A little better, but no cigar. Let us double again.

```
> chisq.test(c(male=4000*0.51,female=4000*0.49))
Chi-squared test for given probabilities
data:  c(male = 4000 * 0.51, female = 4000 * 0.49)
X-squared = 1.6, df = 1, p-value = 0.2059
```

Double one more time.

```
> chisq.test(c(male=8000*0.51,female=8000*0.49))
Chi-squared test for given probabilities
data:  c(male = 8000 * 0.51, female = 8000 * 0.49)
X-squared = 3.2, df = 1, p-value = 0.07364
```

Close. Let us add 1000.

```
> chisq.test(c(male=9000*0.51,female=9000*0.49))
Chi-squared test for given probabilities
data:  c(male = 9000 * 0.51, female = 9000 * 0.49)
X-squared = 3.6, df = 1, p-value = 0.05778
```

Almost there. Add another 1000.

```
> chisq.test(c(male=10000*0.51,female=10000*0.49))
Chi-squared test for given probabilities
data:  c(male = 10000 * 0.51, female = 10000 * 0.49)
X-squared = 4, df = 1, p-value = 0.0455
```

We reached significance. The expected number of observations to reach significance is somewhere between 9000 and 10000.

Note that it is possible to specify the expected frequencies, or rather the probabilities. Let us perform a thought experiment. Say that the expected proportions are 0.51 and 0.49; note that the proportions sum to 1. Imagine that we observed 500 births and 58% were boys and 42% girls; is this a significant deviance from the expected proportions? The expected proportions are 51%, and 49% respectively.

```
> prob=c(0.51,0.49)
> chisq.test(c(male=0.58*500,female=0.42*500), p=prob)
Chi-squared test for given probabilities
data:  c(male = 0.58 * 500, female = 0.42 * 500)
X-squared = 9.8039, df = 1, p-value = 0.001741
```

Yes, we can report that the observation were not likely due to random chance. $\chi^2_{(1df)} = 9.8$; $p < 0.01$.

It is helpful to always mention the number of observations as well, and we could also calculate the *effect size*.

You may wonder why the p-value was not reported with all decimals. The reason is that only some critical thresholds (e.g. 0.05, 0.01, 0.001) are important in practice. Remember that the value is only used as support for taking a decision on rejecting the null hypothesis. The results do not become much better just because they are more significant. It could simply mean that you have enough observations to determine that deviances from the expectations are not random. You should rather choose to report the *effect size* once you have determined that the results are not likely to be explained by random chance.

Note that the χ^2 test may report significance for any small difference if there are only enough number of observations. It is necessary to estimate the *effect size* to know if the observations indicate a large or small difference. Significance only tells if the differences are likely to occur by chance, or not. It is important to remember this, as it is very unlikely that the frequencies you got were actually from a truly random process.

Effect size for a χ^2 test is calculated as Φ below. $\Phi = \sqrt{\frac{\chi^2}{N}}$

with numbers inserted: $\Phi = \sqrt{\frac{9.8}{500}}$ In R:

```
> sqrt(9.8/500)
[1] 0.14
```

Some rule of thumb for effect size by Φ : small effect size 0.1, medium 0.3, large 0.5. In this case we have a small effect size, which effectively means that we would have to observe many births before we can determine if the gender rates are different. However, the effect size for the 1% difference above is much smaller. After all it took 10000% observations to find out that it was significant.

```
> sqrt(4/10000)
[1] 0.02
```

Thus we have a seven times larger effect size in our sample with 58% boys. If we state that in terms of *risk reduction* it would look more alarming.

As you may have guessed, effect size is relative to what we are observing. In some fields we would not bother with an effect size of 0.1, in other fields it would be highly relevant. In the medical field, it is common to report risk reduction; probably because this gives more impressive figures, for example prevention of something that is fairly rare. The effect size is not about significance, but how useful the information is, in this case for guessing the class. If the effect size is large then betting on the majority class would make us very likely to win. Risk reduction could be thought of as an expectation of, for example, lives saved by some manipulation compared to doing nothing. Risk reduction might be stated better in frequencies than in percentages. A risk reduction of 80% might very well be an expectation of 5 saved lives in 10000 treated subjects. If this is worth it may depend on entirely different factors, such as the cost of the treatment and what the effects would be of spending the budget on an alternative.

3.2 Cross tables

A cross table is a structure consisting of rows and columns. The most common case is probably the 2 rows by 2 columns, and it is certainly the easiest cross table to interpret. The null hypothesis is that the rows and columns are independent of each other; i.e., we do not gain any extra information from knowing which row or column we are considering. Let us consider a simple cross table and calculate the expected independent frequencies of each cell (where R_1 and R_2 are the total frequencies for row 1 and 2, and C_1 and C_2 are the total frequencies of column 1 and 2, and T is the total).

a	c	R_1
b	d	R_2
C_1	C_2	T

$$a + b = C_1$$
$$c + d = C_2$$

$$a + c = R_1$$
$$b + d = R_2$$

$$R_1 + R_2 = C_1 + C_2 = a + b + c + d = T$$

What is the probability of belonging to row 1? It could be estimated as R_1/T. Likewise R_2/T is the probability of belonging to row 2.

What is the probability of belonging to column 1? It could be estimated as C_1/T. Likewise C_2/T is the probability of belonging to column 2.

If there is no association between rows and columns, what is the probability of belonging to both row 1 and column 1 (i.e., the cell marked a)? Imagine throwing two coins; one coin determines the rows (head or tails), and the other coin determines the column (head or tails). What is the probability if the coins are absolutely fair? It would be 0.5*0.5 in each cell (head-head, head-tails, tail-head, tail-tail), and all cells would be equally probable. In life the coins might not be so fair. Say that the row coin shows 70% heads, and the column coin shows 60% heads, what would be the probability of head-head? If the coins do not affect each other, it would be $0.7 * 0.6 => 42\%$, since independent probabilities can be simply multiplied together to combine. Likewise, we can calculate the independent probabilities from the row and column frequencies.

$\frac{R_1}{T} * \frac{C_1}{T}$	$\frac{R_1}{T} * \frac{C_2}{T}$
$\frac{R_2}{T} * \frac{C_1}{T}$	$\frac{R_2}{T} * \frac{C_2}{T}$

and to get the frequencies in each cell, we multiply the expected proportion in each cell with the total number of observations to spread into the cells.

$\frac{T*R_1*C_1}{T^2}$	$\frac{T*R_1*C_2}{T^2}$
$\frac{T*R_2*C_1}{T^2}$	$\frac{T*R_2*C_2}{T^2}$

After simplification:

$\frac{R_1*C_1}{T}$	$\frac{R_1*C_2}{T}$
$\frac{R_2*C_1}{T}$	$\frac{R_2*C_2}{T}$

This gives a very simple way of remembering how to calculate the expected frequencies: just combine the appropriate row and column frequencies, and divide by the total. You just have to look at the margins of the table. Fortunately R does the job for us, we just have to specify the table of observed frequencies, and the expected frequencies will be calculated automatically along with the statistical significance. In a 2 by 2 table, we also note that we only have one degree of freedom, just as in the simple χ^2 test above. This is because the row and column frequencies are involved in the calculations, and knowing these values makes it possible to calculate all of the other frequencies from knowing the frequency of one cell. Say that cell a was known. Then b is $C_1 - a$, and c is $R_1 - a$, and then d follows from either $R_2 - b$, or $C_2 - c$.

Let us consider an example in R. Consider a problem in socio-linguistics. Is it true that males use more pronouns than females? The study used a gender-marked corpus, and extracted 8370 noun phrases, and labeled them

either as pronouns (*pro*) or full noun phrases (*np*). A full noun phrase contains a head noun such as *word*, *cat*, or *water*, or a proper name such as *Peter*. That is, we define the domain as either pronouns or non-pronouns.

We will need to represent a table of frequencies, and one way to do this is to use the matrix function. The phrases were tabulated as follows.

```
> x <- matrix(c(2330,870,3230,1940), ncol = 2,
       dimnames = list(c("male", "female"), c("pro", "np")))
> x
       pro   np
male   2330 3230
female  870 1940
```

We have more examples from the male subjects, but we are interested in the proportion of pronouns and nouns. We can test this. First graph the data using **R:assocplot** or textbfR:assoc. These graphs are sometimes called **Cohen-Friendly** graphs.

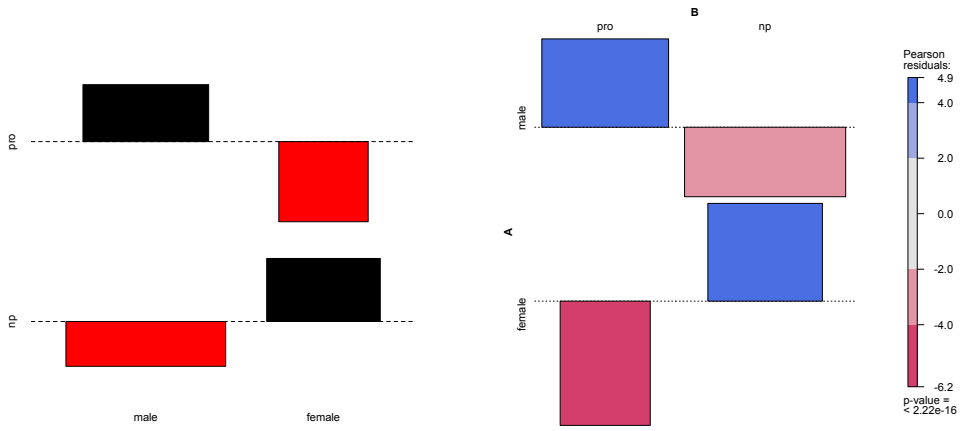
```
> assocplot(x)
```

The assoc graphing has some extra demands. First of all, you need to download the package **vcd** for Visualizing Categorial Data, using **install.packages(vcd)**, and then the command **library(vcd)** to activate the next functions. VCD contains many other useful functions. I initially hesitated to include any external packages, but the **assoc** function extends the functionality of **assocplot** in important ways. Most importantly it represents the Pearson Residuals for each cell, if you allow the option **shade=T**. The shading represents contributions of each cell, and makes it easier to see which cells contribute more to the effect and in which direction. If you want the presentation in the same format as assocplot then the rows and columns needs to be transposed, and this can easily by done with **t(x)**. The **assoc**-function also generalize to tables with more than two dimensions, but that is beyond the scope of this introduction.

```
> assoc(t(x), shade=T)
```

You will now see a graph, where the height of the bars shows contribution to significance, the direction (and color) of the bars show if the cell is over- or under-represented, and the width of the bars indicate the contribution to effect size. The graph indicate that males have a larger than expected proportion of pronouns, and females have larger proportion of noun phrases.

We can easily compute the proportion of pronouns and nouns for each gender.



(a) Graphed by assocplot (b) Graphed by assoc

Figure 3.1: Association Graphs, or Cohen-Friendly graphs

```
> round(100*x/rowSums(x))
       pro np
male    42 58
female  31 69
```

We see that males have 42% pronouns compared to only 31% for females. Is this significant?

```
> chisq.test(x)
Pearson's Chi-squared test with Yates' continuity correction
data:  x
X-squared = 94.2364, df = 1, p-value < 2.2e-16
```

Yes, this looks highly significant. We could report it as: There is a significant association between use of pronouns or nouns, and the gender of the speaker; $\chi^2_{(1)} = 94.24$, $p < 0.001$ ***. The three stars is a convention to mark high confidence in the significance. The exact number is not that important, as the chance of making a mistake for other than statistical reasons are probably more relevant. However, some journals and the APA-style guide may prefer you to report more exact values for p, and simply let the readers and reviewers judge for themselves. If the demand is for more decimals, do not argue with that, but simply provide the numbers as required (for example, you might be required to report $p < 0.05$ as $p = 0.034$ even though you feel the equal sign is not motivated. We simply state by the three stars that we

are very confident that the observed differences in proportions were not due to random chance.

Effect size. For this 2 by 2 table we can calculate $\Phi = \sqrt{\frac{\chi^2}{N}}$

In R:

```
> sqrt(chisq.test(x)$statistic/sum(x))
X-squared
0.1061076
```

This is a small effect, close to a tiny effect according to the previous rule of thumb, even though the p-value was so promising. Remember, the p-value is just a statement about the likelihood of falsely rejecting the null-hypothesis, not an indication of relevance or effect size. In this case, we get high significance because we have a lot of observations (8370 observed phrases).

Another way to estimate the effect is the odds ratio. Type x to get the table back in R.

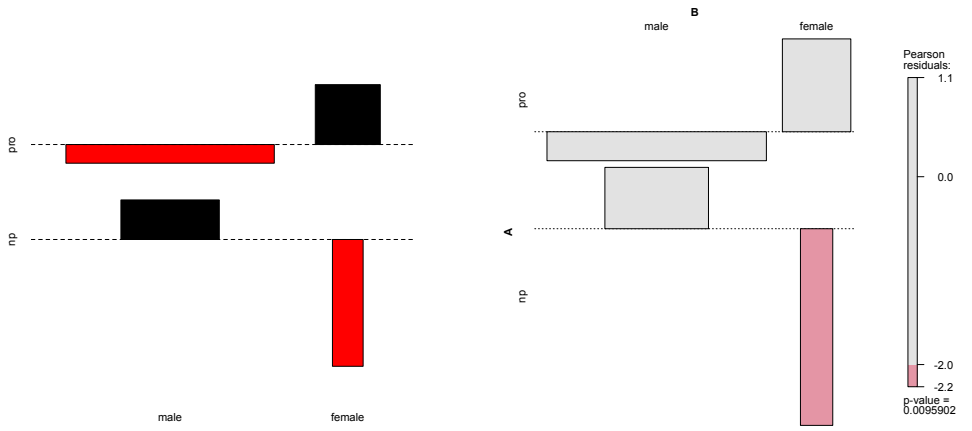
```
        pro   np
male   2330 3230
female  870 1940
```

The odds of male subjects using a pronoun is 2330 pronouns to 3230 nouns (0.7214), and for females it is 870 to 1940 (0.4485). The odds ratio: 0.7214/0.4485 = 1.61. This looks more impressive. It shows a clear male preference to use more pronouns.

The researcher that planned the experiment then came up with an idea. What if pronoun use depends on the sentence complexity? What if the phrases were divided up in *independent* (main clauses) and *dependent* (sub-clauses)? The main point is that a clause is *either* dependent *or* independent, the linguistics of such definitions tend to be more complicated. All the material could be divided up in these two exclusive categories. Let us first see the result for dependent clauses.

```
> x <- matrix(c(1500,160,350,20), ncol = 2,
       dimnames = list(c("male", "female"), c("pro", "np")))
        pro  np
male   1500 350
female  160  20

> assocplot(x)
```



(a) Graphed by assocplot (b) Graphed by assoc

Figure 3.2: Association Graphs, or Cohen-Friendly graphs for dependent clauses

The association plot for dependent clauses reveals that females seem to have a larger proportion of pronouns than males, and a smaller proportion of full noun phrases. The shaded plot reveals that the effect of each cell is not as obvious as before, and this gives an advantage to the **assoc**-function for presenting a clearer presentation of the effects per cell.

```
> round(100*x/rowSums(x))
       pro np
male    81 19
female  89 11
```

Females used 89% pronouns, compared to 81% for males. Are the association between gender and linguistic use significant?

```
> chisq.test(x)
Pearson's Chi-squared test with Yates' continuity correction
data:  x
X-squared = 6.1958, df = 1, p-value = 0.01281
```

Yes, significance shows up with $\chi^2_{(1)} = 6.2$; $p < 0.05$, so the observed association is likely not due to random chance. However, the effect size is tiny.

```
> sqrt(chisq.test(x)$statistic/sum(x))
 X-squared
0.05524619
```

Calculating the odds ratio:

```
male   1500 350
female  160  20
```

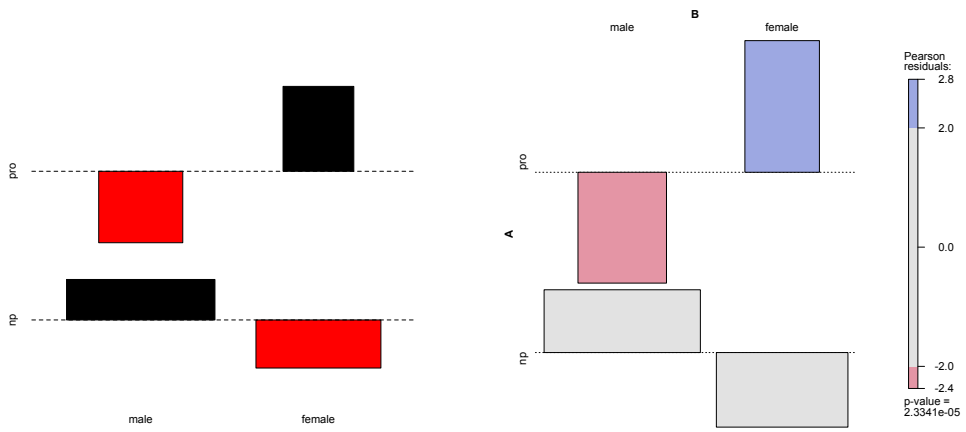
Odds for pronouns over np

```
Male: 1500/350 (4.2857)
Female 160/20 (8)
```

Odds ratio, female over men: $8/4.2857 = 1.87$.

Since we got an effect for male overuse of pronouns, surely men must use many more pronouns in the remaining main clauses?

```
> x <- matrix(c(830,710,2880,1920), ncol = 2,
       dimnames = list(c("male", "female"), c("pro", "np")))
       pro   np
male   830 2880
female 710 1920
```



(a) Graphed by assocplot (b) Graphed by assoc

Figure 3.3: Association Graphs, or Cohen-Friendly graphs for main clauses

The association plot again shows a positive association between pronoun use and females. This is confirmed by calculating the proportions.

```
> round(100*x/rowSums(x))
       pro np
male    22 78
female  27 73
```

Females show 27% pronouns, compared to 22% for males. Can this be significant?

```
> chisq.test(x)
Pearson's Chi-squared test with Yates' continuity correction
data:  x
X-squared = 17.6446, df = 1, p-value = 2.663e-05
```

Yes, it is highly significant. $\chi^2_{(1)} = 17.6$ $p < 0.001$ *** But again we see a tiny effect size.

```
> Phi = sqrt(chisq.test(x)$statistic/sum(x))
 X-squared
0.05275472
```

Calculating the odds ratio:

```
male   830 2880
female 710 1920
```

Odds for pronouns over np

```
Male: 830/2880 (0.2882)
Female 710/1920 (0.3698)
```

Odds ratio, female over men: $0.3698/0.2882 = 1.28$.

Adding information about the status of the clause, turned the results from males using a larger proportion of pronouns overall, to females using a higher proportion of pronouns for both dependent and independent clauses. This looks like a paradox: If females use more pronouns in both dependent and independent clauses, surely they use more pronouns overall?

When we looked at the effect size, we saw that the effect is tiny for dependent and independent clauses, but small for the overall results. Which result should we trust?

This is not an easy question, and that is why it is *a paradox*. The paradox even has a name: **Simpson's Paradox**.

Maybe it is simply so that we cannot know for sure, given the available data. Significance showed that each result was a reliable observation.

Remember that adding an extra explanatory factor may turn the results on the head, and this may happen often in real life. In our examples, there were many more examples from the *males* and one explanation could have

to do with the fact that more data could also capture more of the range of male speech. The males might simply have talked more than the females.

In the example, the effect size was tiny or small, and this may be relevant. A small effect size would be easier to explain away than a large effect size.

The advice is to always include effect size in your formal results, as this makes it easier to evaluate the importance of your findings. The advice for the above small investigation might be to find more talkative females and retest, and the take-home message for the reader is never to trust any small study to give the full truth, as there could always be hidden variables that could explain the data in an alternative way. Are the results valid for all kind of speech? Monologues, elicited speech, talk between males, talk between females, and talk between the sexes? Here is an idea for project, but first find out what is published the area and see how you can improve on that.

The advice for using the *odds ratio* is to be extremely cautious. The interested reader may find out that the odds ratio is misused in a large proportion of published works, so if it is used in a publication your first task is to check that the ratio was correctly interpreted. The odds ratio is often wrongly interpreted as risk reduction.

It is also hard to know if an odds ratio is large for the kind of study at hand. However, one benefit of the odds ratio is that it is invariant on the size of the data (since magnifying the data by multiplying with a constant cancels out in the calculation of the ratio). It may be a good number to compare across several versions of the same experiment. Since the odds ratio can be extremely sensitive, and give an impression of a large effect where the real effect is quite moderate, it is recommended to use logarithms to scale the odds ratio.

This has the added effect that the sign shows the direction of the effect. The R function log accesses the natural logarithm. The log(odds ratio) of the three experiments are: 0.48 -0.63 -0.25; with negative numbers indicating females use more pronouns. Odds ratios close to 1 are close to equal odds, log(odds ratio) close to 0 is close to equal odds (i.e. no detected preference).

3.2.1 Yates correction

In the results from **chisq.test** a Yates' correction is sometimes reported. This correction is a very conservative correction for possible rounding errors, and it demands that any difference between observed and expected frequencies must be larger than 0.5 to fully count. The correction assumes that all differences, also a 0 difference, has a rounding error that is at most 0.5. Yates' correction ensures that we are not looking at effects that stem only from rounding errors.

If R reports that Yates' correction was used, you should also mention it when you report your formal results. This correction is crucial to use when there are cells in the table with few observations.

3.2.2 Larger tables

For larger cross tables, you need to construct a larger matrix. For a 3 by 3 table

```
> x <- matrix(c(1,2,3,4,5,6,7,8,9), ncol = 3,
        dimnames = list(c("row 1", "row 2", "row 3"),
                        c("column 1", "column 2","column 3")))
```

	$column_1$	$column_2$	$column_3$
row_1	1	4	7
row_2	2	5	8
row_3	3	6	9

It is just as easy to do the cross tables analysis using **chisq.test(x)**, but for larger tables you need to adjust the calculation of the effect size. It is however increasingly difficult to interpret larger tables, and to pinpoint where in the table the effect shows best. The chi-square test only let you know if there is an association between rows and columns, i.e., if there is a significant deviation from the expected frequencies of the cells. Cramér suggests the following correction for effect size calculation, where k is the smallest number of the number of rows or columns:

$$\Phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

In R, for a 3 by 3 table:

```
> Phi_c = sqrt(chisq.test(x)$statistic/(sum(x)*(3-1)))
```

3.3 McNemar's test for repeated measures

It is often a useful idea to use paired tests, where you measure on the same subjects more than once. This violates the independence assumption of the χ^2 test, so we have to use another test. McNemar's test is the test of choice to see if there are any changes between two tests (repeated measures) of the same subjects. The test actually only considers the changes. If the test is

significant then there is a net change in some direction. You will easily find the direction of the change by looking at the data.

The McNemar test is very similar to a χ^2 test. The following table illustrates the cross table. At Test 1 subjects either answered yes or no. At test time 2 the subjects could either give the same answer or change their mind. The interesting hypothesis is if anything has happened: do the subject answer more positively (or negatively) at the retest? If approximately the same number of subjects changed from yes to no, as from no to yes there is no change.

	Test2	
Test1	yes	no
yes	X	a
no	b	Y

In the table, X and Y do not have any consequence for the outcome of the test. However, the numbers could be given to show how large the samples were. The only interest is how many have changed their answer between $Test_1$ and $Test_2$.

$$\chi^2_{McNemar} = \frac{(a-b)^2}{a+b}$$

Let us consider a constructed example. The research question is: Does repetition have an effect for spelling, in the absence of feedback? That is, are words spelled more (or less) correct the second time they are written? In order to test this a list of hundred different words were dictated, in random order twice to 20 subjects. There was a pause of two seconds between each word, which would allow for correction of typing errors. The task was to type the words as correctly as possible.

	2nd Try	
1st Try	correct	error
correct	2750	445
error	555	250

The data could be entered in R as:

```
> data <- matrix(c(2750, 555, 445, 250), nrow = 2,
            dimnames = list("1st Try" = c("correct", "error"),
            "2nd Try" = c("correct", "error")))
```

View the data, and make sure it was put in correctly.

```
> mcnemar.test(data)
```

```
McNemar's Chi-squared test with continuity correction
data:  data
McNemar's chi-squared = 11.881, df = 1, p-value = 0.0005671
```

This could be reported as: McNemar's test shows a significant difference between the first try and the second try. $\chi^2_{(1)} = 11.88$, $p < 0.001$ ***. More errors were corrected than errors created in the second try.

One criticism of the above test is that the effects for subjects have been inflated by giving them many words to type. This can be remedied by calculating the average count per subject (dividing by 20).

	2nd Try	
1st Try	correct	error
correct	137.50	22.25
error	27.75	12.50

On average, 22.25 errors were created and 27.75 were corrected per subject in the second try. This is in fact a too small effect to be significant at the subject level, and it is not certain that individual subjects will benefit from a second try. At the same time more errors will be corrected the second time if we consider the whole group.

```
> data2 = data/20
> mcnemar.test(data2)
McNemar's Chi-squared test with continuity correction
data:  data2
McNemar's chi-squared = 0.405, df = 1, p-value = 0.5245
```

However, the first test showed that it does matter for the whole group. Note that the data quality has not become better by dividing, even though we now have two decimals, we still have frequency counts of an underlying nominal variable. The test does not estimate individual variance, so we do not get information about individual differences. In the typical application of the McNemar test, there are only one binomial choice for each individual (e.g. opinions before and after an event).

One other way to handle the situation would be to treat the error rate of each individual as measurement on an interval scale (one argument might be that we have a large number of measurements, enough to make relative error rates lay approximately on a continuous scale); the data may then qualify for using a paired t-test, if we can justify that the difference in error rates is approximately normally distributed. If the assumption of normality does not hold, we may back down to using a non-parametric test such as a paired Wilcoxon test.

McNemar's test is often applied to a simple question with two exclusive alternatives (yes/no, approve/not approve) that is asked on two different occasions to the same subjects. This can be useful for measuring attitude change. It may also be attractive that it is very easy to calculate the statistic, using little more than paper and pen.

3.4 Summary

For nominal data we may compare proportions. The test may be either a simple test between frequencies of two categories, or a structured test of a cross tabulation. There is also a paired version, if it is possible to obtain frequencies before and after a meaningful event, from the same research subjects. The paired version can help detect significant changes in for example opinions after an event. There might be alternatives that are more convenient to use, if the number of data points is large and the collected data can be shown to approximate some demands on for example distribution. It is important to notice that significance can be achieved by reliably detecting small differences if we make a large enough number of observations. The effect size is therefore very useful information. The effect size tells us something about how noticeable the observed differences are, and this in turn may be related to how important the observed differences are. The size of a relevant effect size depends on the field of study and the goals of the investigation. Any findings will need an interpretation and an argumentation from the researchers to put results in a proper perspective.

Chapter 4

Ordinal Data

4.1 Introduction

Ordinal data makes it possible to sort data. A typical situation is that we cannot use the mean as a reliable measure of central tendency. We might be concerned about the shape of the distribution in the data we have, there could be many outliers, the distribution does not look symmetric, a formal test shows that the normal distribution is an unlikely model, or the data level is not at least interval data. We are then in a situation where we cannot calculate the parameters (the mean and standard variation) that we need for parametrized tests.

There are tests that avoid such assumptions on the distribution of data, or the quality of data. Such tests avoid the need to estimate the parameters of the population. This class of tests is called non-parametric tests, and we will specifically look at the Wilcoxon (Mann-Whitney) tests that are available in R as **wilcox.test**.

The non-parametric tests we consider are based on sorting data, and noting how the groups sort in relation to each other. The key operation is sorting, and significance is the chance that the groups would sort so far apart as we have observed from the samples. The median (middle value) is a good extra measure. Since the median marks the midpoint where half the values are greater and half are smaller, a difference in the median values are related to how far apart the two populations are. However, the tests consider how all values sort relative each other. The null-hypothesis is stated as the *allocation shift* being 0 between the groups, and the alternative hypothesis is that we can exclude, with some confidence, that the observed allocation shift is due to random chance.

In the following, we make *virtual* experiments. The *sample* function in

R, gives us a sample from a population that we may specify with how many different categories (items) and the probability of each category. The reason for using these virtual experiments are twofold: first you should become aware of the idea of sampling from a (potentially infinite) population, and second we can construct data sets for exercises automatically.

4.1.1 An example: Swedish "congruency"

Let us investigate an example. It is very common in language research to ask subjects to rate alternatives for acceptability. The scale can be some small interval, but we often want a middle point so an odd number of alternatives is preferred, usually 5 or 7. Such a scale is called a Likert scale, and it is widely used when information is sampled from questionnaires and the like. Let us use 5 alternatives in the example. Let 1 stand for unacceptable, 3 for ok, and 5 for unproblematic. The investigator has to think of which terms to use, and how subjects would understand how the terms correspond to judgments (this is often not trivial).

In Swedish there are some rare occasions when people may have alternatives for gender congruency. This is noted by which word in a phrase will give the form of the determiner. Alternative a) would be "early congruency" with the word for 'a kind', for example *En tax är ett slags hund*. Alternative b) would be "late congruency" with the head noun, as in *En tax är en slags hund*. Both example sentences translate to "A dachshund is a kind of dog". Are these alternatives equally acceptable?

First we might ask 20 subjects about alternative a (embedded in a larger test batch, where each alternative has to be handled within a second), and then another 20 subjects are asked about alternative b in the same experimental setup. This is an independent values design. In a paired design we might ask the same subjects about both alternatives, and make sure to pair the responses. This would be a paired values design.

4.2 How to construct a virtual experiment

We can simulate this experiment in R, using the *sample* function. In order to simulate this we may model the outcomes on a frequency distribution. Say that we have done the experiment before, and we got the following frequency distribution:

```
> x <- c(1,1,1,2,2,2,2,2,2,2,3,3,3,4,4,4,4,4,5,5)
> f_early <- c(3,7,3,5,2)
```

Three people thought the example was unacceptable, and two thought it was unproblematic. We have counted up the sorted answers (in x) to get the distribution in f_{early}.

For late congruency, 1 subject found it unacceptable and 8 found it unproblematic:

```
> y <- c(1,2,2,3,3,3,4,4,4,4,4,4,5,5,5,5,5,5,5,5)
> f_late = c(1,2,3,6,8)
```

You may test the original data using a non-parametric test:
e.g. $wilcox.test(x, y)$. Make sure to look at the distributions using the histogram plot: **hist(x)** and **hist(y)**. Are the histograms symmetric around a central value?

In order to compare the two histograms in the graph, we will use a trick and divide the graphing space up in two parts using the **par** function. The command below creates two rows and one column using **mfrow=c(2,1)**. Later try switching the numbers and see the effect.

```
> par(mfrow=c(2,1))
```

Next we call **hist** to make histograms, with the **prob** parameter we call for a graph that uses proportions (probabilities) instead of counts, and with **ylim** set to the interval between 0 and 0.8 to create a common y-axis. We then add a **density** model and plot it in blue. The **lwd** sets the width of the density line. The **lines** command adds a line to the previous graph.

```
> hist(x, prob=TRUE, col="gray",ylim=c(0,0.8))
> lines(density(x), col="blue", lwd=2)

> hist(y, prob=TRUE, col="gray",ylim=c(0,0.8))
> lines(density(y), col="blue", lwd=2)
```

When you are done, you may want to reset the layout of the graphing window using the command **layout(1)**, otherwise you will continue to plot graphs in pairs.

For the sake of simulating experiments we can convert these data to likelihood estimates (probabilities of each answer 1...5), simply by dividing by the sum of frequencies. Note that the numbers are occurrences of each grading, from 1 to 5. Also note that we do not have any idea about the value of one step on this kind of scale. From 1 to 3 is a change from unacceptable to ok, but is this the same distance as from ok to unproblematic?

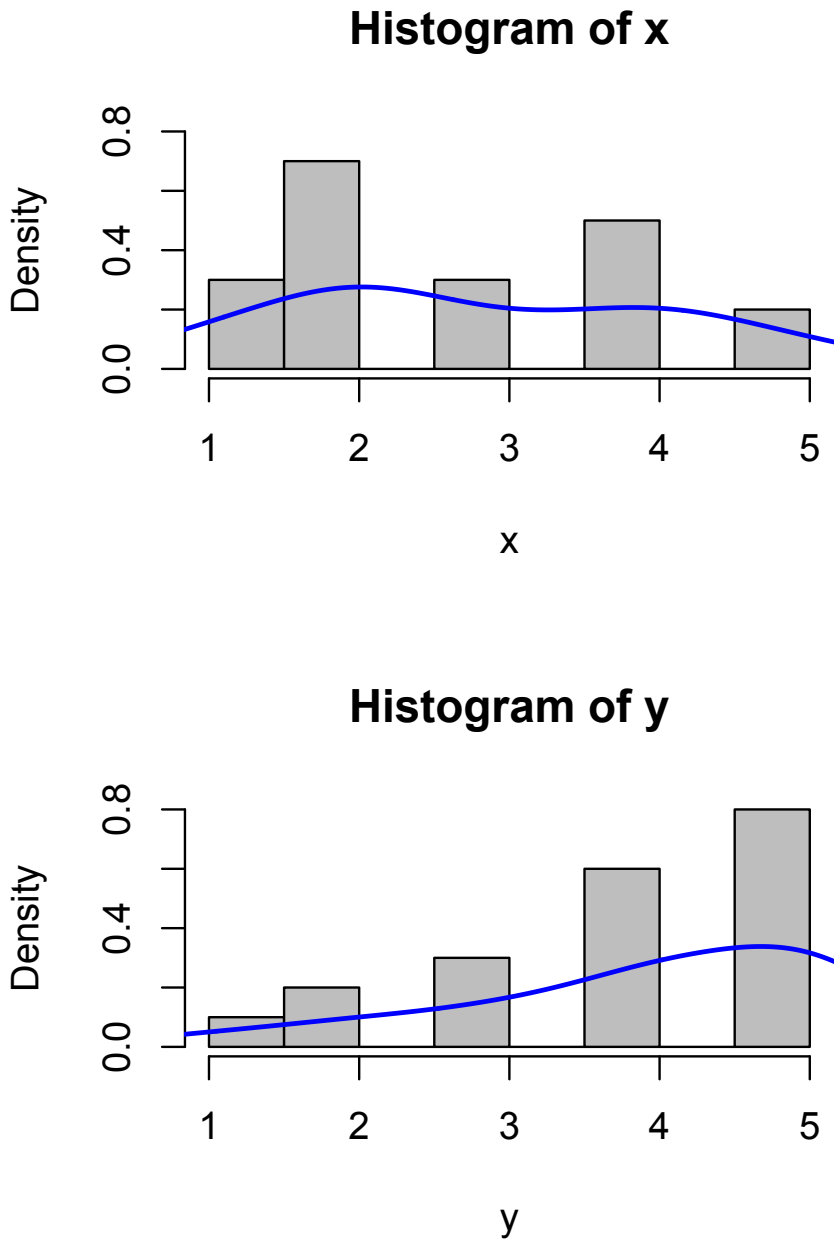


Figure 4.1: Visually compare two distributions

```
> p_early = f_early / sum(f_early)
> p_late = f_late / sum(f_late)
```

You may check that the sum of the probabilities are 1:

```
>   sum(p_early)
>   sum(p_late)
```

You may look at the probability vectors simply by entering their names.

```
> p_early
[1] 0.15 0.35 0.15 0.25 0.10
> p_late
[1] 0.05 0.10 0.15 0.30 0.40
```

Now we want a sample from this probability distribution.

```
> s_early=sample(1:5,20,prob=p_early,replace=T)
```

This says that we want integer numbers in the range from 1 to 5, according to the probability distribution in p_{early}, which has to have the same length as the interval as this defines the probability of each ordered category. The last item: $replace = T$ makes each draw independent of each other. The probability distribution is not affected by the sampling. This is sometimes called an urn model with replacement. If replace is false, then we will take out each number as we sample them, and if we draw 5 numbers without replacement we will get five different numbers, approximately ordered according to their probability of occurrence. The default is not to replace. This model of sampling is sometimes called an *urn model without* replacement. You may think of the alternatives as colored, or numbered, balls in an urn.

On my first run I got:

```
> s_early
 [1] 2 4 2 1 1 2 4 4 5 4 2 5 2 5 1 5 4 1 2 3
```

The next sample is from the model of the late congruency.

```
> s_late=sample(1:5,20,prob=p_late,replace=T)
> s_late
[1] 4 5 2 1 5 4 1 3 4 5 5 5 5 2 3 5 2 4 4 2
```

You will get different numbers, since the computer will simulate a random sequence for the sample function. If you repeat the full experiment a few times, you will get a feel for how often you will get significance from randomly sampling the given probability distributions.

You may look at the histograms for these samples, and see the shape of the sample distribution.

```
> par(mfrow=c(2,1))

> hist(s_early, prob=T, ylim=c(0,0.7))
> lines(density(s_early), col="blue", lwd=2)

> hist(s_late, prob=T, ylim=c(0,0.7))
> lines(density(s_late), col="blue", lwd=2)
```

We may also test if the distribution is significantly different from the normal distribution, by means of the Kolmogorov-Smirnoff test, **ks.test**.

```
> ks.test(s_early, pnorm, mean=mean(s_early), sd=sd(s_early))
...
> ks.test(s_late, pnorm, mean=mean(s_late), sd=sd(s_late))
```

I got $p < 0.14$ and $p < 0.33$ respectively; so normality cannot be excluded; but we know that the data is ordinal and no higher, and therefore a test should be very cautious about testing differences in the means. Because we have *ordinal* data we *should* use the **wilcox.test**.

4.3 How to test ordinal data

We will have to decide if we want a *paired* test. In this case, we have only one measurement from each subject.

```
paired = F
```

We may also think about a *two-tailed* or *one-tailed* test. In this case, we have no motivation why we would look only in one direction, so the default (*two-tailed test*) should be used. We may also ask R to calculate a confidence interval for us.

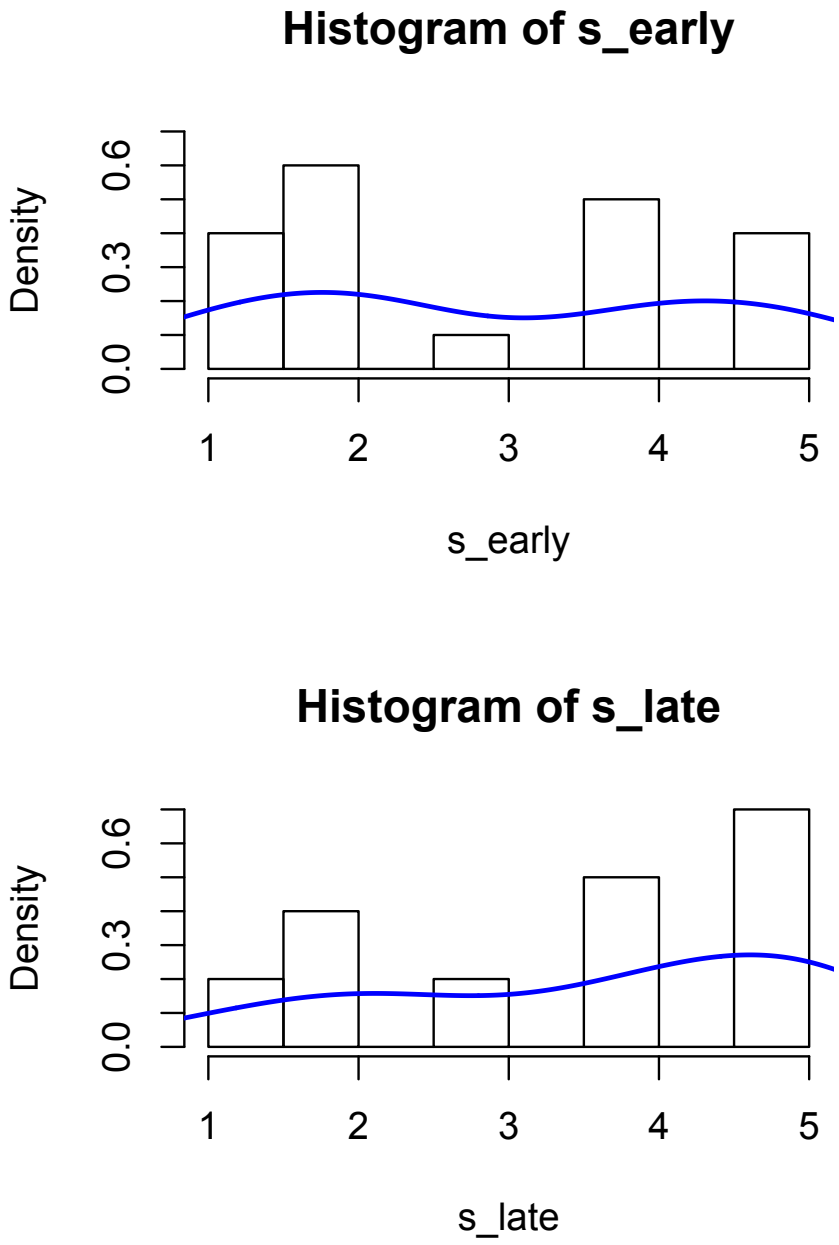


Figure 4.2: Visually compare two sample distributions

```
> wilcox.test(s_early,s_late, conf.int=T)

Wilcoxon rank sum test with continuity correction

data:  s_early and s_late
W = 153.5, p-value = 0.2005
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -1.999982e+00  3.128666e-05
sample estimates:
difference in location
            -0.9999565

Warning messages:
1: In wilcox.test.default(s_early, s_late, conf.int = T) :
  cannot compute exact p-value with ties
2: In wilcox.test.default(s_early, s_late, conf.int = T) :
  cannot compute exact confidence intervals with ties
```

This could be reported as: A Wilcoxon test ($W = 153.5$) showed no significant difference in the location shift, $p > 0.05$. You may report the median ratings for each sample.

```
> median(s_early)
[1] 2.5
> median(s_late)
[1] 4
```

The median rating of *late* is 4, and *early* is 2.5. We interpret this as possibly explained by the variance in the responses , and we cannot exclude that there in fact is no difference. As an exercise: Repeat the experiment with different samples and see how often you get signficance.

The reported confidence interval gives that the (pseudo-)median shift in location is between -2.0 and -0.0 steps, if we round to one decimal. The e-notation indicates the number of steps the decimal is shifted. Note $e + 00 = 10^0 = 1$ is no shift of the decimal, and $e - 05 = 10^{-5} = 0.00001$ is five steps to the left. However, the confidence interval we have established is *not* for the *difference in the median* values, but an estimated confidence interval for *the median of the differences* between the two groups. We cannot exclude that there is no location shift between the two populations.

There are warning messages from the test. These are warnings that the values have been approximated, but otherwise it is not an error message. When there are only five values possible, a tie is very likely to occur, so approximation of the test values are necessary. You may beware of using too many of the decimals. The p-value should generally be reported at $p < 0.05$, $p < 0.01$, or $p < 0.001$ if significant; and confidence intervals are limited by the precision of the data, and how many measurements were sampled.

If we used a t-test instead of the non-parametric wilcox.test, we would likely have similar results. For the data in the example above, the formal result is almost identical.An investigator should use the test that gives the best power to detect a significant difference, but only use tests where the data fits the assumptions of the test. Therefore we should avoid a t.test for these data sets, because the data is ordinal and the assumption of a normal distribution can be questioned, and the samples are small.

In the presence of outliers, a non-parametric test is more robust. We do not wish to report an effect that is based on the outliers pulling in the right direction for significance. The t-test is attractive to use, even for ordinal data, if outliers are under control, the sample size is large and the sample data pass some test of normal distribution. Neither is the case with the data above.

Non-parametric tests often have good power to detect significant differences, and in cases where outliers pull against significance a non-parametric test could show significance where a parametric test would not. The advice is to use a non-parametric test when it is the correct test to use. It might also be the more *cautious* test to use, as there are *less assumptions* on the shape of the distribution or the quality of data. For small samples, it is difficult to know if assumptions of normality holds, as there is too little data for reliably estimating the shape of the distribution, and therefor a non-parametric test may be preferred.

Some researchers may hesitate to use non-parametric tests because they give less information (no parameters are estimated). The Wilcoxon test with the calculated confidence interval, complemented by the median of the compared groups, may alleviate such concerns. Even though such complementary characteristics can be calculated, the test itself does not use them to arrive at its conclusion, and therefor the test is still considered to be a non-parametric test.

4.4 How to do a Paired test

Let us go through an example of a paired test. In the example above, we showed how sampling one value from each individual allowed us to detect a difference in acceptability between two different constructions. However, what if some people have preferences in one or the other direction? There might be a correlation between accepting one construction, and rejecting the other.

Imagine that we sampled the same subject twice. The same data as before will now look like below. Presented as rows instead of columns to save space.

```
s_subject 1 2 3 4 5 6 7 8 9  10 11 12 13 14 15 16 17 18 19 20
s_early   2 4 2 1 1 2 4 4 5   4  2  5  2  5  1  5  4  1  2  3
s_late    4 5 2 1 5 4 1 3 4   5  5  5  5  2  3  5  2  4  4  2
```

In order to balance the experiment we imagine that we have presented the early condition first in half of cases, and in the other half we presented the late condition. There would otherwise, in a real experiment, be an alternative explanation for any possible effects, namely the order of presentation.

Below is the result of a paired Wilcoxon test, for the example data above:

```
> wilcox.test(s_early,s_late, paired=T,conf.int=T)

Wilcoxon signed rank test with continuity correction

data:  s_early and s_late
V = 43, p-value = 0.2006
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -2.0000138  0.5000189
sample estimates:
(pseudo)median
     -0.724228
```

There could be warning messages. For example, there are no exact p-values, nor exact confidence intervals, if we have ties in the rank order. You should report that there were ties, and that the p-value and confidence intervals are reported with this caveat.

There does not seem to be any significant difference between the first and second sample. Note that the W of the non-paired test above is replaced by V, indicating a paired test. The formal result can be summed up as: A paired Wilcoxon test ($V = 43$) showed no significant difference in ranking

between examples of early and late congruency, $p > 0.05$.

The confidence interval for the difference is from -2.0 to $+0.5$, which again shows that there is little evidence for an actual difference. But a lack of evidence is not evidence for no difference. We simply have no evidence to claim a difference.

There are often much higher correlations between the choices of individuals than the simulated data has captured. The samples where indeed constructed so that they were statistically independent of each other. Therefore we should not expect to gain much from pairing the data. It is left as an exercise how you would construct data sets with individual bias for either choice.

4.5 Friedman test for multiple levels

Let us finish with a last thought experiment. Imagine that we have four measurements for each subject, and ten subjects. For example, we decided to include the example sentence in a short discourse. We can show the example sentence in a short discourse with early congruence, or with late congruence, and then ask the subjects to rate the sentences. This gives four pairings. a) Show early congruency, test early congruence. b) Show early congruence, test late congruence. c) Show late congruence, test early congruence. d) Show late congruence, test late congruence. To get better, more robust, measures, we can show many different sentences and take the *average* of the rating in each condition.

Let us test how to evaluate the idea in R:

```
> a = c(2.2,2.3,1.5,2.2,2.7,3.1,2.4,3.6,1.7,2.1)
> b = c(2.4,2.1,4.1,3.5,2.3,4.2,4.7,5.1,4.1,3.2)
> c = c(2.1,2.1,5.3,5.4,3.2,4.2,4.4,4.3,2.3,3.8)
> d = c(2.6,3.7,2.7,5.2,4.2,3.2,5.6,4.3,3.5,2.9)
```

Since the data is based on ordinal data, we do not feel comfortable with the assumption of normal distribution so we want to have a non-parametric test. We would also like to test all conditions at the same time, in the fashion of Analysis of Variance for repeated measures, a method that will be introduced in a later chapter. The test to use under these conditions is called a **Friedman** test.

We need to construct a matrix for our data that the test can analyze. In the code below, note how the columns have been structured using concatenation of the column vectors in the c() command, and how this structure is formatted into a matrix, where we have specified that we have entered data by column (not by row).

```
> data = matrix(c(a,b,c,d), nrow=10, byrow=FALSE,
       dimnames = list(1:10,
       c("earlyEarly", "earlyLate", "lateEarly", "lateLate")))
```

```
> data
   earlyEarly earlyLate lateEarly lateLate
1         2.2       2.4       2.1      2.6
2         2.3       2.1       2.1      3.7
3         1.5       4.1       5.3      2.7
4         2.2       3.5       5.4      5.2
5         2.7       2.3       3.2      4.2
6         3.1       4.2       4.2      3.2
7         2.4       4.7       4.4      5.6
8         3.6       5.1       4.3      4.3
9         1.7       4.1       2.3      3.5
10        2.1       3.2       3.8      2.9
```

We can make a boxplot of this data.
This shows how easy it is to handle structured data in R.

```
> boxplot(data)
```

The summary function gives an overview of the data.

```
> summary(data)
```

Look at the span from the smallest to the largest value, and the median in each condition.

The data matrix can be analyzed by the Friedman test.

```
> friedman.test(data)
Friedman rank sum test
data:  data
Friedman chi-squared = 10.299, df = 3, p-value = 0.01619
```

A Friedman test shows significant differences between conditions, $\chi^2 = 10.3$, $p < 0.05$ *.

The test shows that there are differences between the columns, but to investigate where the differences originate we need to apply a pairwise Wilcoxon test. The bad news is that we need to reformat our data, so that the *pairwise.wilcox.test* can handle it correctly.

```
> data2 = data.frame(score=c(a,b,c,d),
 condition = factor(c(rep("earlyEarly",10),rep("earlyLate",10),
 rep("lateEarly",10),rep("lateLate",10))),
 subject = factor(rep(1:10,4)))
```

You should always perform a summary of the data you have entered. This allows you to make a sanity check on the data frame, and ensure that everything has been entered correctly. We have created a variable that allows us to keep track of the repeated measures from the same subject.

```
> summary(data2)
     score            condition     subject
 Min.   :1.500   earlyEarly:10   1      : 4
 1st Qu.:2.300   earlyLate :10   2      : 4
 Median :3.200   lateEarly :10   3      : 4
 Mean   :3.362   lateLate  :10   4      : 4
 3rd Qu.:4.200                   5      : 4
 Max.   :5.600                   6      : 4
                                 (Other):16
```

The summary shows that the measurement data has been entered under the label score, and that scores range from 1.5 to 5.6, but most importantly the score is correctly entered as numeric values. All four conditions are entered, with 10 values in each condition. There are four values for each subject, and a total of $(4 * 6 = 24) + 16 = 40$ values; i.e., 10 subjects that have given four data points each.

Look at the data:

```
> data2
```

You should see the values for subjects 1 to 10 repeated for the four conditions.

You may now attach the data.frame (*data*2) so that you can refer to only the headings (i.e., the name of the dependent variable: score, and the factor *condition* which contains 4 levels, and subject tells which subject gave the line of data). We are ready to perform the pairwise Wilcoxon test, in its paired (repeated measures) version. The pairwise function constructs all unique pairs of factors. The levels are available in the grouping variable that we have called *condition*, and the dependent measure is in the variable called *score*. The pairwise function then performs the specified test (this time it is the *wilcox.test* on all the pairs, and adjust the significance for the fact that we perform *multiple comparisons*. The default adjustment method is *Holm*, but you may find out how to select other methods (use the ? command with the function name as an argument).

```
> attach(data2)

> pairwise.wilcox.test(score,condition,paired=T)
Pairwise comparisons using Wilcoxon signed rank test
data:  score and condition
          earlyEarly earlyLate lateEarly
earlyLate 0.087      -         -
lateEarly 0.049      1.000     -
lateLate  0.012      1.000     1.000

P value adjustment method: holm
```

The pairwise test revealed that the earlyEarly condition, is significantly different (lower ranked) from lateEarly and lateLate (both $p < 0.05$, Holm method was used to compensate for multiple comparisons).

4.6 Summary

Whenever there is doubt if the normal distribution is a good model for your data you may use the **wilcox.test** if you can sort your data in a meaningful way. When you have decided to use this non-parametric test you should also specify if you have performed repeated measures on your subjects. If you have you should make sure that you have paired your data points from the same individuals.

There is also the choice of using a one-tailed or a two-tailed test. There need to be a very good motivation to use a one-tailed test. If differences in your data indeed are accounted for by random chance then using a one-tailed

test simply doubles your risk of making a bad decision and falsely reject your null hypothesis on no real evidence.

Non-parametric tests are sometimes avoided because it is very tempting to use a model that increases your chances of finding significance when the assumptions of the model hold and there is evidence in your data. However, non-parametric tests are more robust, and have fewer assumptions on the underlying model for your data.

Chapter 5

Parametrized models, and interval data

5.1 Introduction

A model for interval data is often based on the normal distribution, possibly adjusted for sample size as in the **t-test**. The tests assume normal distribution, which entails a symmetric distribution around a central value. The same tests may be used on data that we can show satisfy an assumption of approximate normal distribution. This can be tested formally using for example Kolmogorov-Smirnoffs test (R: **ks.test**).

We assume that the mean is a good estimate for effects, when using these tests. Changes in the mean corresponds to a change in the central tendency of the whole population, shifting the whole population in one direction. However, variation might also be affected by the experimental conditions, which is considered when the tests determine how large a shift is needed to get a significant difference. There are situations where the change in variance is important in itself, for example if we want faster type writing and more accurate type writing. Changes in writing speed may be indicators of writing problems, such as finding the right word or wondering about the spelling of a word.

5.2 Normal distribution

The normal distribution has two parameters, the mean and the standard deviation. These parameters are properties of a population of measurements. There is only one mean and one standard deviation in the population of all possible measures we could get. They are assumed to be constant for the

whole population, but when we sample from the population we will likely get different estimates for different samples. The mean is the central tendency and thus the peak of the distribution.

The value of the mean in a sample is easy to calculate, and it is our best guess at the true population mean.

The value of the standard deviation is likewise easy to calculate in a sample, and it gives a best guess at the true deviation of the population.

Tests that are built on the idea of estimation of population parameters are commonly referred to as parametric tests. Our problem is that it is not practical, or in most cases even possible, to sample the whole population. So we have to settle for a sample, and this sample is often very small compared to the population of interest. We must then make sure that our estimates are not unreasonable. We will often assume that the population does not change over time. If it does we will need to estimate the effect of time on the measures.

A guideline is that we want each estimated parameter to be close to the true population values, if we resampled the population many times, and averaged our estimates for the number of times we resampled. That is, we want the estimates to be true to expectations, which means that we expect them, on the average, to converging towards their true values.

One assumption is that the population parameters are not changing. In an experimental situation this can be a faulty assumption. We might have seasonal variations, or variations that are inherently due to learning effects, or maturation effects, or changes in the population. If there is considerable time between measurements, or time is inherently an important factor, then we deal with a different class of measurements that would take a time series analysis and estimations of trends or patterns in time. This is often handled by a baseline condition, where no experimental manipulation is performed, that would estimate the effect of time alone.

The discussion will be limited to cases where we can think of variation as a property of the population we are measuring. *If* time is involved it could be thought of as discrete time, for example before and after a certain manipulation. Variation could be due to individual variation, variation of the measurement equipment, or other factors. However, if these variations are due to random error (i.e., a non-intentional, non-systematic, difference from the mean) they can be thought of as normally distributed with a mean of 0, i.e., errors are expected to cancel out on the average if there are only enough measurements. If errors are systematic, this would not be true. For example, variation that is due to fatigue effects would likely get more serious with an increased number of measurements on the same increasingly tired subject.

5.3 Using t-tests: independent or paired design

Data that is approximately normally distributed is one requirement for t-tests. The t-test is based on a model called a t-distribution, which has two (population) parameters that we need to estimate from samples, namely the mean and the standard deviation. The t-distribution is similar to the normal distribution, but it allows for a higher proportion of values in the tails of the distribution, depending on how small the samples are. In a small sample, outliers will be able to affect the estimation of the population parameters much more than in a large sample. This problems is handled, to some extent, using a t-distribution that depends on the size of the sample(s), as measured by the degrees of freedom of the test.

We will concentrate on comparing means, although there are occasions when it would make more sense to compare variance or standard deviation; for example imagine a new heart drug that stabilize the rhythm of the heart. It would be nice to be able to reduce the highs and the lows in heart frequency without necessarily affect the average that much. Likewise, in language studies there might be periods where linguistic variation is increasing without seriously affecting any measure of central tendency.

The Mean

The mean is easy to estimate, we sum up measurements and divide by the number of measurements. However, the intent of calculating the mean is to have an idea of the central tendency in the population and to get an idea which value is the most expected. We must have data that it makes sense to calculate means for. Is the mean always a possible value for what we measure? For example, the mean number of children per woman may be around 1.8 these days, but no woman has 1.8 children. If we participate in two races, and the outcome is 1 gold and 1 bronze, we still did not get a silver even if the mean place is the second place. Imagine a country where a few are very rich and most are very poor, on the average the expectation may be a middle point that does not exist or is unrepresentative of the population.

The mean makes sense with interval data, and a distribution that is symmetric around the mean, where values around the mean are the most expected values. A bell shape curve, which is another common name for a normal distribution.

How many values do we need in our sample to estimate the mean? Actually, one is enough to estimate the mean, but we will most certainly miss the actual population mean.

If a Martian suddenly beamed into the room, so that we could measure her height it would be the best estimate we have of the expected height of Martians. Depending on the which hypothesis we have of why the Martian beamed into our room, we might have different ideas of the population mean though. A physicist may claim that beaming from Mars to Earth would take energy in proportion to size, and therefore it would be easier to send smaller individuals. A political scientist might say it makes sense to send the most impressive representative of their species. The point is that when we claim the mean of our observations as a measure of central tendency it is in absence of any better hypothesis.

Standard deviation, and degrees of freedom

We have no idea yet of the variation of height in Martians, so the standard deviation of the Martian height is undefined until we get one more observation of another observation. But we would only have to wait for one more Martian, and then we would have a possibility to measure *two* differences from the mean, if the individuals are not identical in height. This make it possible to get a number on variation, and standard deviation. Note that we would only have *one* difference between the two individuals. When we calculate the expected distance to the mean, we created an extra measurement (the mean) that is not a new observation as it contains information from all the measurements. We account for this in the calculations. Therefore we will divide by one less than the number of measurements, to be true to expectations. We would not expect no variance with one measurement, rather we would expect infinite variance or complete ignorance of the variance if we only have one measure.

The entity that is named variance in statistics can be estimated from the sum of the squared distances to the mean divided by one less than the number of distances to the mean. The squaring is done to avoid negative numbers. Summing up just the signed distances to the mean would sum to zero. If we square each distance, before summing them up, we get a series of 'areas' that then sum to a larger area. Dividing it up gets an 'average' area, and the side of that area is the *standard deviation*. The standard deviation is the expected distance of any point to the central point i.e., the mean. Remember that we only have sample estimates from the real population mean and standard deviation.

In math typeset this would be expressed more compactly as below:

mean: $\bar{x} = \frac{1}{N}\sum_i x_i$
variance: $v = \frac{1}{(N-1)}\sum_i (x_i - \bar{x})^2$
standard deviation: $s = \sqrt{v}$

The magic of working with a model, such as the normal distribution, now makes it possible for us to state the uncertainty of the data using a confidence interval around the mean. Given that the model is correct and our estimates of the true mean and standard deviation are correct, a 95% confidence interval is two standard deviations to either side of the mean. This defines a data interval where about 19 out of 20 data points ought to be if they were randomly drawn from the theoretical normal distribution with that mean and that standard deviation.

The testing procedure for a t-test is based on comparing the means of two groups, or one group to a fixed value. How sure are we of the estimated mean value? We need to estimate the variance of the distribution of possible means. Remember that there is only one true mean in the population, and if our sample was the entire population there would be no variance of the mean. The variance of the mean will get smaller and smaller with increasingly large samples. The error of the mean is, like previously, the side of the *variance area*; a number which is reached by taking the square root of that area. The error of the mean makes it possible to model the mean itself by a normal distribution with the deviation equal to the error of the mean. One error out from the mean, on either side, will cover about 68% of the possible mean values, and two standard errors out will cover about 95%.

The standard error of the mean

The standard deviation of the normal distribution is a constant property of the population that is modeled, a descriptive statistic, just as the mean. The standard deviation of the data does not change consistently with the number of data points in our sample: We may miss the correct value, but increasing the sample cannot make the standard deviation smaller. Just as the population mean it is a fixed property that we try to estimate from a limited sample. The standard error of the mean is different, because we do get more certainty with larger samples.

If we could sample the whole population there would be no uncertainty in the mean. One way to think of this is to imagine that you had a large number of measurements of the true mean, each one with an error due to random sampling. If you summed all of these measures, and averaged over the number of measurements, then you would get the mean + a sum of errors divided by the number of measures: the mean error, or the error of the mean.

This error gets smaller by the number of measures, since the sum of errors is limited by errors in both directions.

$\frac{1}{N}\sum_i(\bar{x}+\epsilon_i) = \frac{1}{N}\sum_i(\bar{x}) + \frac{1}{N}\sum_i(\epsilon_i) = \bar{x} + \bar{epsilon}$

The question is how fast the expected error shrinks with increased samples. The total sum squared error (SSE) in a sample is variance: $\sum_i(x_i - \bar{x})$, and the variance of the sample was to divide the SSE by the number of data points. The smallest sample we could make (for estimating variance) is a pair of data points, and without order information we could make $N * N$ pairs (for example, 1,1,1,2,2,1,2,2), and excluding the N data points that pair with themselves gives $N * N - N = N(N-1)$ pairs. If we divide the SSE by this number of pairs we get the definition formula for the variance of the sample mean, and taking the square root delivers the definition formula for the error of the sample mean. Note that this is not a general proof, but it might make it easier to remember the definition formula for the error of the sample mean.

Variance of sample mean: $v = \frac{1}{N(N-1)}\sum_i(x_i - \bar{x})^2$
Error of sample mean: $sqrt\frac{1}{N(N-1)}\sum_i(x_i - \bar{x})^2 =$
$=\frac{1}{\sqrt{N}}\sqrt{\frac{1}{(N-1)}\sum_i(x_i - \bar{x})^2}=$
$=\frac{sd}{\sqrt{N}}$

According to the definition formula, it is possible to reduce the standard error of the mean by increasing the sample. The standard deviation of the data is a parameter of the distribution model, and we can only estimate its value. However, the error of the mean varies with the square root of the number of data points. That is, if we want to half the error of the mean, i.e. divide by 2, then we would need to quadruple the number of data points, as $\sqrt{4N} = \sqrt{4}\sqrt{N} = 2\sqrt{N}$.

The real proofs involve showing that the estimates of the parameters (mean, standard deviation, standard error of the mean) are true to expectations, and never consistently over- or under-estimates the true population parameters given enough repetitions.

We can never guard against occasional misses of the population parameters, but given enough new samples we should be sure to hone in on the true parameters. This is what the definition formulas are doing. Exactly why this is so is beyond the scope of this book, but it should help to know the intentions behind the estimates. For now we have to trust that the researchers that have established the formulas and practices have done their work.

5.4 Comparing two groups

When we have interval data, and suspect that data is approximately normally distributed, we may compare the means of two groups, since the mean is a *fair estimate of central tendency* under these circumstances. Normal distribution may be visually verified by making a histogram.

For example, 50 random number from a normal distribution with default values mean=0, and sd=1.

```
>   x=rnorm(50)
>   hist(x)
```

Check with mean(x) and sd(x) that the mean is close to the expected mean (0) and that the standard deviation is close to the expected deviation (1). Try again with increasingly larger samples: 100, 1000, 10000. What happens? Can you characterize what happens in one sentence?

With increasingly larger samples, the estimation of the population mean and standard deviation gets increasingly accurate. Note that nor the mean nor the standard deviation are expected to get smaller or larger, they are simply approaching the values of the population. Of course, in this case we know the population values from the known theoretical population specified by the *rnorm* function. In real life it is very rare that we know the real values of these parameters, as we cannot sample the whole population.

5.4.1 Checking for normality

How does the distribution look like? The size of each bar is automatically calculated, from the sample distribution. The first run I made gave a fairly symmetric curve, with a center around 0, which fairly much looked like a normal distribution. There are formal test for how close a sample distribution is to a given distribution, for example the *Kolmogorov-Smirnov* test (R: **ks.test**, using **pnorm** to select a test for normal distribution).

```
>     ks.test(x,pnorm,mean=0,sd=1)
```

This would show significance (e.g. $p < 0.05$) if the sample distribution is different than the distribution tested against. If you test

```
> ks.test(x,pnorm,mean=mean(x),sd=sd(x))
```

you would test the shape of the curve against a normal distribution with *the same mean and standard deviation* as those derived from the sample. The Kolmogorov-Smirnov test can be used to test if two samples come from the same distribution, or if a sample distribution fits a given distribution. There are many options to explore. For simple examples, it is often enough to look at the histogram of the data and see if it is reasonably symmetric, but formal verification will need a formal test, such as the ks.test.

The preferred choice for testing differences between two groups, given interval data or better, is the t-test. The t-test assumes normally distributed data, but calculates with a higher risk that extreme values affect the results, therefore it works with the related t-distribution that has a higher proportion in the tails of the distribution. The more degrees of freedom (i.e., the larger the samples) the closer the t-distribution is to a normal distribution.

We will simulate some experiments to show the logic of the tests. We will begin with two statistically independent samples. We will show how plots of statistically independent data will look like if we plot the data. Look out for cloud shapes, rather than lines, curves and other patterns. A roundish, cloud shape, plot is an indication of statistical independence between the samples. Statistical independence means that you will have any advantage in predicting a the value in the second sample from knowing the corresponding value in the first sample.

5.4.2 The t-test for independent samples

Imagine a reaction time experiment. Two groups, different individuals, are selected. The first group is called the control and they get to see words (and non-words to make it a task) and the task is to recognize the words (and reject the non-words) as fast as possible.

A set of target words have been constructed according to some principles (e.g., using familiar but low frequency words with two or three syllables denoting concrete nouns). The same set of target words is used for both groups, but the task group get to see a very brief presentation (typically less than 100 ms) of a related word, called a prime, directly before the target word. It is checked that there is no initial overlap between the prime words and the target words. For example, the target word *apple* may be preceded by *banana*. Each individual will be presented with many different target words. Reaction times are recorded as averages for the correctly recognized words only. Only individuals that manage 90% accuracy will be included in the study; individuals are sampled until the required number of subjects have passed the requirements. This has implications for how representative the study is, and the drop-out rate should be presented and discussed in an

actual research report.

From previous studies, the experimenters expect that this control task can be performed with reaction times around 550ms (i.e., a little slower than half a second) with a standard deviation of 25ms. The task, given a related prime word, is expected to be reacted to faster at around 500ms with a standard deviation of 25ms.

The first task is to calculate the group size, needed to have a fair chance of getting results if there actually is a difference. The expected difference is 50ms, and standard deviation is 25ms. We want a power of 90%, and significance is at the conventional 1/20 (5%) level.

```
> power.t.test(sig.level=0.05, power=0.90, delta=50, sd=25)
```

The power test shows that an effect that large will need 7 individuals in each group (always round up). Thus, a fairly low number, practical for a quick demonstration.

Let us use R to simulate such an experiment (remember, you cannot publish such a study as you have not looked at any actual task).

```
> control=rnorm(7,mean=550,sd=25)
> task=rnorm(7,mean=500,sd=25)
> hist(control)
> hist(task)
```

We may also visually verify that the control and task are statistically independent. Look for patterns when we plot task by control. With so little data it is easy to imagine some pattern, but in the first run I did not detect anything unexpected.

```
> plot(control,task).
```

In one run, the task had $\sigma = 19.4$ and the control was 21.5. The mean of control was 565.8, and the mean of the task was 504.6; so this looks promising and stronger than expected from the power calculation.

```
> t.test(control, task)
```

A first run gave $t_{(df=11.87)} = 5.597$; $p < 0.001$ (***-significance). Note that the degrees of freedom is not reported as $(7-1)+(7-1) = 12$; which is the usual text book example that we lose 1 degree of freedom for each mean we calculate for the test as that last value is not free to vary given that the mean is known (i.e., a population parameter)). The value (11.87 in this example) is rather close to 12, but since there seem to be some small amount of correlation in the data. It looks like we have some small gain in guessing, if we know measurements for one group and based our guess on that information. The compensated degree of freedom accounts for this observation, and make it a tiny bit harder to find significance than a naive model (without considering possible correlation).

The t-test also gives a 95% confidence interval of the difference: the confidence interval is from 37 to 85 ms., and estimates of the mean reaction time of control (566ms) and task (505ms).

Cohen's d

We may want to give an estimate of the effect size, as seen between the samples. Cohen's d is the difference in means between control and task divided by the pooled standard deviation. The pooled standard deviation can be calculated in R as:

```
> pool = sqrt( ((7-1)*sd(control)*sd(control)+
  (7-1)*sd(task)*sd(task))/(7+7) )
```

where 7 is the group size. A first run gave Cohen's d = 3.3; which is a very large effect.

```
> d= (mean(control)-mean(task))/pool
```

Glass's Δ

Glass's Δ would be

```
 > delta = (mean(control)-mean(task))/sd(control)
```

the first run gave 2.8, which is still a large effect. Note that your numbers will be different, depending on how the samples were drawn from the normal distribution. You may wish to go through the calculations in R, and select different samples. Note down how significance (and power and effect size) differ between different draws from the exact same normal distribution.

5.4.3 Caution

Note that in a real experimental situation we will not be so sure that the model of normal distribution fits the experimental setting. We may also occasionally get bad estimates of the parameters. Some amount of error may be introduced from the measure instruments, or from individuals learning the task, or getting tired of the task, or other fluctuations between measures. These extra sources of variation are often thought to be of a lower magnitude, and crucially not consistently related to our experimental groups; thus over a number of trials these effects will tend to be small on average. However, it is a reason to select larger groups than the minimum required, as calculated from the analysis of statistical power.

5.5 The t-test for repeated measures (paired data)

Imagine exactly the same task as above, but this time we have decided to use each individual as its own control. In order to avoid repetition effects, each individual is only allowed to see each word once, so there need to be separate lists of primed and non-primed target words, and no overlap between primed and non-primed words. The first half of the non-primed list and the second half of the primed list can be presented to the first half of subjects. The second half of the non-primed list and the first half of the primed list is presented to the last half of subjects. These details are mentioned just to make you aware of the lengths you may go to avoid spurious factors in the experimental design, and in this case to isolate a priming effect from a repetition effect. As before, for each subject the average reaction time for correctly recognized target words were recorded, in the primed and non-primed condition; two paired values for each individual.

It is usually a good idea to design a study such that each subject is used as its own control. Some people are slower, and some are faster, but in this kind of study we are only interested in the difference between the primed and non-primed condition. If the *speed* of lexical recognition is roughly constant within each individual than we may expect to get more precise measurements this way, as individual differences for lexical recognition will cancel out between primed and non-primed, and focus will be on how large the effect of priming a target word will be. Some people are faster than others.

A nice effect is that we can usually make do with fewer subjects, although we measure a pair of data from each subject. This is especially important when there is a high cost for using each subject.

Let us first do the power analysis for paired samples, we expect the same difference and the same variance for the prime effect:

```
> power.t.test(sig.level=0.05, power=0.90, delta=50, sd=25,
  type="paired")
```

The run give that we may get the effect using only five subjects, which is better than the independent group analysis above. Although this is getting close to the minimum requirements for the tests, so paired tests would be more useful for smaller effects that need more subjects.

Let us use R to simulate such an experiment. The control is expected at 550ms, and sd=(25/550)*550. The task is expected to reduce the reaction time by an average of 50ms, with sd=(25/550)*50; i.e., 2.3 if we keep the same relative σ (about 4.5%); or sd=25 if we keep the same absolute σ for the priming effect. Let us simulate the case with sd=50, also for the effect.

```
> control=rnorm(5,mean=550,sd=25)
> task=control-rnorm(5,mean=50,sd=25)
```

Histograms may not be helpful, due to the extremely small sample size. In one run, significance was achieved with $t(4) = 7.7$; $p < 0.01$ (**). The effect size measurement would be Pearsons r, computed by:

```
> r = cor(control,task)
```

The value of one run was 0.84, indicating a very strong effect with 84% explained variance. As previously noted, you may get very different numbers. Caution would be advised as the sample size is so extremely small. In a real life experiment that small, the fear of uncontrolled factors, or unknown factors, would be very real. Uncontrolled factors could known, but not controlled for in the experiment, or unknown until their effect show up.

We may interpret the formal results as showing that a related word in close approximation to a target word typically makes lexical recognition faster. One possible way this could happen is that the target word is partly activated by the related word, and this pre-activation makes recognition of the target word faster. Note that this interpretation may follow from the motivation for the experiment, but it is not a direct consequence of the detected significance. The observed effect is consistent with predictions from a model that assumes spreading activations, but is it the only model that could account for the observations?

5.5.1 Suggested exercise

Rerun the simulations a few times, and note if the significance is as stable as indicated by the power analysis. Rerun the whole procedure for an estimated difference of 10ms, instead of 50ms. (Keep sd= 25ms). How large samples are needed for a power of 90%? Give the group size for both paired and independent samples. Look at the histograms, do they look more or less normal?

5.6 Summary

Measurement data for the t-test requires that the mean is a good estimate of the central tendency, so we need approximate normal distribution and data on at least interval level. Before conducting an experiment it is important to know how many measurements need to be sampled (i.e., how many subjects to include). This can be done using power analysis for the t-test, if the difference in mean can be predicted either from previous experiments, similar experiments, or from knowing the lowest difference we would care about. The standard deviations could be similarly approximated or stipulated.

It is important to recognize if the design of the experiment is with independent groups (i.e., different individuals) or with repeated measures on the same subjects. The paired design can also be used for estimating effects in highly correlated subjects (e.g., siblings, twins or difference).

The calculation of the statistic using software such as R is straight forward. You specify the two groups to be compared, and if the test is paired or not; independent measurements are assumed as a default. The test gives a t-value, with the degrees of freedom, and a p-value for significance.

The p-value is used to *take a decision* on keeping or rejecting the null hypothesis. If the null hypothesis is rejected it is also necessary to interpret the formal result, and state the cause of the difference. How is the manipulation (e.g., showing a prime word) related to the observed difference?

5.7 Further examples

5.7.1 Language Change and Voice Onset Time

Voice Onset Time (VOT) is a measurement of the time it takes from the release of a full closure in a stop consonant until voice is detected in the following vowel. Voiced plosives (such as /b/,/d/, and /g/) have voice onset before the release (thus a negative VOT), voiceless unaspirated plosives (e.g., /p/,/t/,/k/) have voice onset near the release of the closure, and aspirated voiceless plosives take some time before voice onset. The measurement can be extracted from speech spoken at a fairly constant speech rate, usually controlled by letting speakers read out lists of words, in the context, words where the target stop consonants are followed directly by a vowel. The sounds can be visualized using a spectrogram technique where both closure and release of the stop is detectable, and the time until voice onset can be measured accurately, nowadays most often using computer software (such as Praat (REF)). In order to get more accurate measures from each speaker, there are often multiple measurements and the average measurement is reported. The reason for reporting only the average is that there is ideally only one measure for each subject, allowing more risk critique for pseudo-replication, which may boost the significance of small differences.

One question in language change is how pronunciation is affected by language contact. Speech data was collected from five speakers of an indigenous language, in 1970. Twenty years later another five subjects were sampled for the same test words. The investigation focussed on two stops (/t/ and /k/). Aspiration at the release of these stops will be visible as a longer VOT. The language contact was mainly with English, which has a low level of aspiration on these stops, and thus fairly short VOTs. It was predicted that language contact would shorten VOT.

The data was as follows:

```
stop year vot
t early 20
t early 50
t early 58
t early 46
t early 69
t late 34
t late 55
t late 24
t late 47
```

```
t late 62
k early 65
k early 85
k early 62
k early 72
k early 88
k late 35
k late 42
k late 67
k late 63
k late 37
```

You may input these data into three columns in a spreadsheet of your choice (e.g., Excel, or OpenOffice) and save as comma separated text, specifying the column separator as a tabulator character. This will allow you to access the data through R using:

```
> data <- read.delim(file.choose())

> summary(data)

> attach(data)
```

The **file.choose()** function allows you to select the file, you need to know the name of the file and where you saved it. The summary function is used to check that the data was correctly read from the file. If all is correct you may use **attach**, to allow R to access data from the names of the columns.

You summary should look something like this:

```
 stop      year         vot
 k:10   early:10   Min.   :20.00
 t:10   late :10   1st Qu.:40.75
                   Median :56.50
                   Mean   :54.05
                   3rd Qu.:65.50
                   Max.   :88.00
```

Is there an effect of year?

```
> t.test(vot~year)
```

The ~ sign may be read out as *'grouped by year'*. The p-value < 0.1, does not reach significance. However, we have not used the information that we predict that the effect would go only in one direction (shortening of VOT, or loss of aspiration).

```
> t.test(vot~year, alternative="greater")
```

That is, we will test that the value for VOT is greater for the early years compared to the late years. If the test is done this way, we have reached significance, in a *one-tailed* test of significance. We have tested that there is a difference in VOT for stop consonants (/t/ and /k/) in the predicted direction.

Is it true that there is a difference for both /t/ and /k/? Note that we *have not tested this*! How to test this is left as an exercise, but a hint is that we need to test for an interaction effect, which is best done using a *factorial analysis of variance*. Save the data set, and come back to this question after reading the chapter on analysis of variance.

Graphing the data is always a good idea. Let us try a boxplot, and see if we get any more insights.

```
> boxplot(vot~year*stop)
```

You may read this as : plot VOT grouped by year and stop. Is the difference between early and late /k/ the same as the difference between early and late /t/?

5.8 Summary of example

When you get a data set you need to determine how the measurements are grouped. The choice of an independent test is based on the information that the subjects were different in the two samples. If the subjects had been the same, maybe the effect was due to the aging of the subjects? Further questions are related to how well matched the two samples were. Are they matched for age and gender? Is education a factor?

The choice to do a one-tailed test is motivated by the prediction that aspiration would most likely go in the direction of the contact language, i.e., aspiration would be lost to some degree if the phonetic change is due to language contact. This would need to be argued in detail, as it is important for the conclusion of the test. Furthermore, the number of subjects is only five in each group. This might be too few to draw any conclusions about the effects. We will also need to calculate the effect size.

The overall difference between early and late years is $61.5 - 46.6 = 14.9$ To calculate the effect size we need the pooled standard deviation:

```
> s1=sd(vot[year='early'])
> s2=sd(vot[year='late'])
> pool=sqrt((9*s1+9*s2)/(10+10))
```

Cohen's d = 14.9/*pool*; which is around 0.9 so the estimated effect is large, and should be possible to replicate, if we find more recorded material. Let us try a power analysis.

```
> power.t.test(sig.level=0.05, power=0.80, delta=14, sd=20,
  alternative="one.sided")
```

A conservative estimate of the effect at 14ms, and standard deviation at 20ms, would require about 26 subjects in a larger experiment, to have a fair chance (here: 80%) of replicating the study.

It should be noted that we have allowed two measurements from each subject, and though they are measures of different stop consonants there is a concern that the results have been inflated. Can you see from the boxplot which consonant we should pick if we want a study with only one measurement per subject (analyzed by a t-test)? Does including both /t/ and /k/ help the results, or make it more difficult to detect the effect? If the effect is similar for both sounds, it would help to have twice as many measurements. If the effect of one consonant is much larger than the other, the effect would be diminished by including both (doubling the measurements for the variable 'stop').

In this case, we might have considered doing the analysis on only the consonant with the best effect. However, we could only find out the best consonant by looking at the data before deciding the test, which is a risky strategy. You might also have considered doing both consonants, and then

you would have performed multiple comparisons, which need to be compensated if the risk of falsely rejecting H_0 is to be correctly estimated. More on this theme when we introduce pairwise tests.

There are obviously problems with using a t-test when we have multiple factors to consider (in this case, when the sampling was done, and which consonant was sampled) and if we have multiple levels on one variable (imagine having sampled more than two consonants). Analysis of variance offers some remedy to these problems, and it is a very useful procedure. This is the theme of the next chapter.

Chapter 6

Analysis of Variance

6.1 Introduction

There are many cases where we need to go beyond analyzing two groups against each other. One disciplined way is within the Analysis of Variance (aov, anova) paradigm. The logic behind $ANOVA$ is to compare variance within groups with variance between groups, or factors, i.e. to see how much of the variance can be explained by where the data came from. This chapter will only be an introduction to the logic and a show case of some very common situations. If you need to do anything more complicated, you will likely need to read up on the literature. There are a lot of information about analysis of variance on the net and in text books, for example (cf.

`http://en.wikipedia.org/wiki/Analysis_of_variance`

For our purposes it is sufficient, for the moment, to notice that anova is a very useful and elegant method. We will look closely at assumptions of the model. As for the t-test, normality is assumed and variance should be similar in the groups, with a similar number of measurements in each group. If the design uses repeated measures from the same subject the design *must be* balanced so that it is possible to pair measurements. This will need special treatment of issues such as outliers and missing values, but these are more advanced topics and may be researched later. One other common assumption is that the variance in each group is approximately equal. In favor of the model is that it is fairly robust, and it will tolerate some deviance from the assumptions of normality and equal variance in groups.

The text will give examples of some common designs, and analyses. We will introduce models that assume fixed factors, of two varieties: Intrinsic factors, such as gender, that cannot be manipulated, and factors that can

be experimental manipulated and cover all the alternatives the experimenter want to consider. One example of such an experimentally controlled variable is to compare reaction times from a presentation with a prime word and a presentation without any priming stimulus.

Sometimes fixed factors are just a selection from many more possibilities. If we want to test the effect of caffeine on reaction time, we may choose to provide a cup of coffee in one condition and no coffee in another condition, but there could be many other levels to control such as if the subjects have had coffee or tea (is the effect of tea the same as for coffee?) or had any special diet or use any medications. If we want to generalize our results, we may have to consider that we have only looked at one possible way of adding caffein to the subjects, and it could be of interest to see how different subjects differ in their reaction. This will typically demand a larger design, and also a design that accounts for random effects (looking more closely at the individual variation).

Think of this chapter as a starting point. There will be situations where a fixed factors model will not be ideal, and a random effects or a mixed effects model will be more appropriate.

First a quick look at how to enter model designs using formulas in R, followed by examples of very simple models using either independent measures or repeated measures. The goal of the chapter is to prepare the reader for modeling data with the intent to find out effects of controlled factors. The reader should be able to perform analyses that are fairly close to the examples, by finding a good model of the data and testing that the assumptions are a reasonably met.

6.2 Formulas in R

R provides one convenient format for modeling data. The model is built up using a small set of operators that relate variables to each other. The ~ (tilde) sign could be read *is modeled by*. The $*$ sign can be read as *and* (including interaction between factors) and the $+$ sign is an *and* that excludes interaction between terms.

For example, when the dependent variable *score* is modeled by the factors *group* and *condition* (and their interaction), then the formula would read: *model score by group and condition and their interaction*, and the formula would be:

```
score ~ group * condition
```

If the analysis is for a *repeated measures* design then an explicit error term need to be constructed. A repeated measures is when you have several measures of the dependent variable (*score*) from the *same* subjects. You will have to specify an error term that tells the analysis how you have made the repeated measures. The error term contains a / sign that can be read *by*. For example, *model score by group and condition and their interaction with Error from repeated measures of subject by condition.* This states that the *levels* of condition (e.g. baseline and treatment) come from the same subject. It looks easier to follow in a formula:

```
score ~ group * condition + Error(subject/condition)
```

There are examples below that use both types of formulas. The names in the formula come from your data. In the examples, score has been used as a name for a measurement variable (the dependent variable) and group and condition are names of factors that are used to group the measurements in the model at hand.

6.3 One Way Anova

6.3.1 Generating Examples

The following function is a small program in R that generates data for a one-way anova. The data is intended for the user to generate different scenarios and then use the R function aov to see if the effects can be detected.

There are several parameters that go into the function, but they all have default values. M1 is a *baseline* value that gives the expected mean in group a. M1 has 500 as a default, it could be thought of as 500ms in a reaction time experiment. N is the number of measurements in each group. SD1 is the standard deviation in group a. The default is set at ten percent of M1. SD2 is the standard deviation in group b, and SD3 is the standard deviation in group c. Defaults are the same standard deviation as in group a. D1 is the effect in group b, and its default is half of the standard deviation in group b. D2 is the effect in group c, with a default set to half of the standard deviation of group c.

Two positive effects In order to generate data for a situation with an *additive* effect you may let D1 and D2 go in the same direction. For example,

```
data <- genaov.data(D1=50,D2=75)
```

One positive and one negative effect You may let the effect go in opposite directions (note the sign):

```
data <- genaov.data(D1=50,D2=-50)
```

In real life, people may not be so easy to model and within one group you may have individuals that have no effect, some have a positive effect, and some may show a negative effect. In the version of analysis of variance presented here, we will assume that there is only one measurement per subject, or one measurement per subject and "treatment"-group (in the case of repeated measures). If you have many measurements per subject you may have to average data so that you get data that conforms to the model of either independent measurements, or repeated measures. We also assume an equal number of measurements for all groups in a repeated design. We want to avoid *pseudo-replication*, which has an inflationary effect. It may make significance appear since the analysis assumes more independent measures than you really have, so your study looks larger than it is to the analysis. However, if you have a lot of measurements for each individual one solution is to model each subject to find out the limits for the true mean of each subject, but this is a more advanced technique described elsewhere (Baayen, 2009). One common solution is to get mean values for each individual. This solution keeps the model within the anova paradigm, which is familiar to more of the audience for your research, while controlling for pseudo-replication.

The function *genaov.data* below simulates a subject distribution assuming that measures in group a, b, and c are from the same individuals. The dependent variable (i.e. the measurement variable) is labelled *score* and the independent variable (i.e. the grouping variable) is called *group*. The repeated measures analysis of this data is available since there is a variable that marks *subjects*. However, it would be more correct to model each subject as having their own distribution; this is left as an excercise for the more ambitious readers. Subjects could be intrinsically slower or faster (i.e. they could have different expected means) and they could be more or less variable in their performance. The analysis is mostly concerned about differences between conditions, and the function will provide practice material as well as some insights: for example that we might sometimes fail to detect a real difference, and that we may sometimes claim a difference where there is none.

```
genaov.data <- function(M1=500,N=100,SD1=M1/10,SD2=SD1,SD3=SD1,
                                        D1=SD2/2,D2=SD3/2) {
    a = rnorm(N, mean=M1, sd=SD1)
    b = rnorm(N, mean=M1+D1, sd=SD2)
    c = rnorm(N, mean=M1+D2, sd=SD3)
```

```
    data2 = data.frame(
       subject = factor(c(rep(1:N,1), rep(1:N,1),rep(1:N,1)) ),
       group = factor( c( rep("a",N),rep("b",N),rep("c",N) ) ),
       score=c(a,b,c))

    cat("\n",
     paste(rep("-",21)),"\n",
     sprintf("   \tExpected     \tActual"),"\n",
     sprintf("a"),"\n",
     sprintf("  :\t%f\t\t%f", M1, mean(a)) ,"\n",
     sprintf("b"),"\n",
     sprintf("  :\t%f\t\t%f", M1+D1, mean(b)) ,"\n",
     sprintf("c"),"\n",
     sprintf("  :\t%f\t\t%f", M1+D2, mean(c)) ,"\n",
     paste(rep("-",21)),"\n"
    )
    return(data2)
}
```

The following examples will use data generated by genaov. There are seven measurements ($N = 7$) in each group. You should test with your own data. Because the data are drawn from a normal distribution of random numbers (with mean and standard deviation specified) you will get slightly different results between different applications of the simulation function.

```
data <- genaov.data(N=7,D1=50,D2=75)

 - - - - - - - - - - - - - - - - - - - - -
     Expected      Actual
 a
   : 500.000000 470.885128
 b
   : 550.000000 533.643622
 c
   : 575.000000 578.946426
 - - - - - - - - - - - - - - - - - - - - -

attach(data)
```

The command attach(data) makes the data available for analysis without having to specify which data set until you do *detach(data)*.

Do not forget to call for a summary of your data. This gives you a sanity control. You will expect 7 subjects in each group, i.e. three group measures for each subject, since we called the function with $N = 7$.

```
subject group     score
1:3     a:7   Min.   :436.5
2:3     b:7   1st Qu.:487.0
3:3     c:7   Median :526.6
4:3           Mean   :527.8
5:3           3rd Qu.:565.8
6:3           Max.   :689.8
7:3
```

6.3.2 Independent measures

You may look at the data graphically using the boxplot function.

```
boxplot(score~group}
```

This gives you an idea of the range of data in the groups (e.g. treatment conditions). You may think of group *a* as a *baseline* (i.e. no treatment). Imagine that group *b* are measures of when a semantically related word was shown before the word. Imagine that group *c* are measures for word that starts with the same initial syllable as the word presented before the one the subject is asked to react on. When we analyze the situation as independent measures we assume that all measures are independent, and most crucially that they come from different subjects that are independent of each other. This means that we assume 21 subjects instead of 7, in our example.

```
model <- aov(score~group)
```

We ask *R* to model the score by the group. The function *aov* builds up a linear model that we can get an analysis of variance table from, simply by typing *summary(model)*. The summary function is extremely useful, and it adapts to the data that it receives.

```
summary(model)
            Df Sum Sq Mean Sq F value   Pr(>F)
group        2  41226   20613   12.43 0.000407 ***
Residuals   18  29854    1659
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The signficance table tells us that there is a highly significant effect of group ($p < 0.001$, marked as a three star significance, $***$). It does not tell which group or groups that provides the effect, so we must do one further test, a *pairwise t-test*.

```
pairwise.t.test(score,group)

Pairwise comparisons using t tests with pooled SD

data:  score and group

  a      b
b 0.0198 -
c 0.0003 0.0520

P value adjustment method: holm
```

From the above pairwise test, we found that groups b and c are different from group a, with $p < 0.05$ and $p < 0.001$ respectively, but group b and c failed significance, since $p > 0.05$, although with a very narrow margin, which sometimes is marked as a trend towards significance.

The pairwise t-test used pooled standard deviation and the Holm-method to compensate for multiple comparisons. Because we have made three comparisons we have increased our chances of getting significance by a factor 3; the Holm method builds on that logic (sometimes called Bonferroni compensation for multiple comparisons). Holm's method is slightly less conservative for reporting the second and smaller significant effects. The interested reader will find much information on the internet on Bonferroni correction, and Holm's method. It is important that some compensation has been made if there were multiple comparisons involved, since each comparison may show significance by random chance.

The formal null hypothesis for the one-way anova is, as usual, that there is no difference between the levels (groups). The alternative hypothesis is that there is at least one group that is different from the other groups. We have to make specific tests to pinpoint exactly where the difference is, or where the differences are, if we have found out that the null hypothesis can be rejected. The formal report would be stated along the lines: $F(2,18)=12.43$; $p < 0.001$

(***), the number 2 comes from the degrees of freedom for *group*, and the 18 comes from the degrees of freedom for the residuals that are not accounted for by the groups ($18 = 21 - 3$). The F is the statistic that needs to be reported, and it quantifies the deviance from the expected together with the degrees of freedom for the group variable. The degree of freedom from the residuals is the degree of freedom not accounted for by the groups. There are 3 times 7 individual measurements and 3 groups, i.e. $21 - 3$ measures are free to vary when we have accounted for group means. We lost one degree of freedom from the original three groups, as knowing the mean for all groups and any two of the groups makes it possible to calculate the mean of the third group. We should state that pairwise t-tests show that group b and c were significantly different from group a, and you may state the $p - values$ for each significant comparison, for example, a is significantly different from both b ($p < 0.05$) and c ($p < 0.001$).

6.3.3 Repeated measures

If we have made repeated measures, then we have used the same subjects in all groups. The rational for this is that we may account for individual differences in a much more precise manner by comparing each "treatment" paired for the same subjects. This will allow the analysis to use that individuals carry their own characteristics for how fast they respond, if we assume that the individuals have not changed between measurements. It is typically very beneficial for the experimental design to use the individuals as their own controls, and the number of subjects can often be reduced using this design. We can use the data that you previously generated from the genaov.data function, which you *attached*, i.e. made available for quick reference with the $attach(data)$ command. When you are done with a data set you may remember to use $detach(data)$.

Important: If we have made repeated measures we are *obliged* to analyze the situation as a repeated measures (paired) design, since the measurements are not independent of each other.

You may look at the data graphically using the boxplot function.

```
boxplot(score~group}
```

This gives you an idea of the range of data in the groups (or treatment conditions). However, you may also construct a graph that plots the range of effects for individuals. If group a is the baseline, this value may be subtracted from each other group, such that we get a difference score for each individual. This is fairly easy to construct in a spreadsheet, and is suggested as an exercise.

When we analyze a repeated measures design we have to specify where the *error* (variance) stems from. In this case, we state that the error comes from the subjects. For a one-way anova it suffice to specify the subject in the Error term, but to be explicit we can also include the factor that we have repeated measures for (i.e., group, see formula below). Try the formula with "Error(subject)" as well, and see that it delivers the same results.

```
> model <- aov(score~group+Error(subject/group))
> summary(model)

Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  6   8072    1345

Error: Within
          Df Sum Sq Mean Sq F value  Pr(>F)
group      2  41226   20613   11.36 0.00171 **
Residuals 12  21783    1815
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The formal null hypothesis for the one-way anova with repeated measures is that there is no difference that cannot be explained by random variation. The alternative hypothesis is that there is at least one treatment (group) that is different from the other treatments. We have to perform more specific tests to find out where we can find differences. Do not forget that we have repeated measures, and thus the specific tests are going to be paired measures.

The formal report would be stated along the lines: F(2,12)=11.36; $p < 0.01$ (**). The degree of freedom from the residuals is the degree of freedom not accounted for by the groups. There are 3 times 7 measurements, 7 subjects and 3 groups, and we have stated that error stem from subjects, with 6 degrees of freedom: i.e. $12 = 21 - 3 - 6$ measures are free to vary when the source of error has been accounted for.

We then need to go forward with pairwise tests, but this time we need to specify that measurements are paired, i.e. each pairwise comparison stems from the same subjects ($paired = T$).

```
> pairwise.t.test(score,group,paired=T)
```

```
Pairwise comparisons using paired t tests

data:  score and group

  a     b
b 0.012 -
c 0.015 0.120

P value adjustment method: holm
```

We report that pairwise t-tests show that group b and c were significantly different from group a ($p < 0.05$ for both).

Note that the figures are slightly different from the analysis of independent measures. Assuming independent measures gives the analysis 21 independent measures from different subjects. Assuming repeated measures from 7 subjects in 3 different conditions gives a lower number for the degrees of freedom to vary.

In the first version of the experiment we would recruit 21 subjects, in the second version we would recruit only 7 subjects. Typically, we would benefit slightly from controlling for individual variation in a repeated measures design, but the main benefit is that the experiment could be carried out with fewer subjects.

Keep in mind that if a (true) repeated measures design is analyzed as if the measures are independent the significance is inflated by pseudo-replication (i.e. there were not as many independent sources for the measurements as the analysis would report). It is crucial to do the analysis correctly, and the needed information comes from the design of the experiment. Look for how many truly independent measures are available.

6.3.4 Summary One-way Anova

A one-way analysis is performed when we have a situation where there are many levels or groups that should be compared simultaneously. We need to determine if we have independent measures, or if we have a repeated measures design. If we have a baseline condition, it might be an idea to calculate difference scores, and in many cases this could reduce the analysis to a simpler case, if we are interested only in if two treatments are different. In the analyses provided above, using aov, we will also get information on which groups or treatments are different from the baseline condition. In addition, we need to think of how the baseline condition was identified.

It is often a very good idea to use a repeated measures design, as this limits the number of subjects, who are instead used as their own control. The repeated measures design specify which measurements are highly correlated. Do not forget that you need to specify paired tests when you look for specific effects using the pairwise function.

6.4 2x2 Factorial Anova

The first factorial design we will look at is a 2 by 2 design. This typically consists of one group variable, which states a feature that makes an individual belong to a group, it could for example be gender. Note that such intrinsic variable cannot be manipulated during the experiment. The second factor is typically a baseline compared to a treatment of some kind. For example, a baseline for a priming experiment could be to show nothing before the word that is to be decided, and the treatment could be to show a prime word that begins with the same syllable as the target word.

A factor is considered a *fixed effect factor* if all possibilities of interest are included, either an individual belongs to group 1 or 2, and there are no other alternatives.

A factor is considered a *random effect factor* if the values we have is just a random selection from a larger array of possible values, where other values are, or might be, of interest to our study.

For simplicity, we will assume that we have managed to find the factors that are relevant to our investigation. If we have only managed to find a sample of some possible factor values, we may need to look into other methods of analysis (for example, a linear model with random effects (lmer, cf. Baayen, 2009).

The 2x2 design is very attractive since we do not need to perform any searches for specific effects after we get the the analysis from the factorial analysis of variance. The main effects of the two individual factors, and the interaction effect between the two factors are readily available from the summary report of the analysis function (aov). More complicated designs may make it difficult to determine where the effect stems from.

6.4.1 Generating Examples

The following gives a simulation function that generates data for us to test analysis function. The simulation assumes two groups, and two conditions (baseline and treatment). We can specify an expected difference between baseline and condition for group 1 ($D1$) and group 2 ($D2$).

```
gen.data <- function(M1=500,M2=M1,N=100,SD1=M1/10,SD2=M2/10,
                                        D1=SD1/2,D2=SD2/2) {
    g1b = rnorm(N, mean=M1, sd=SD1)
    g1t = rnorm(N, mean=M1+D1, sd=SD1)
    g2b = rnorm(N, mean=M2, sd=SD2)
    g2t = rnorm(N, mean=M2+D2, sd=SD2)

    data2 = data.frame(
     subject=factor(c(rep(1:N,2), rep((N+1):(N+N),2))),
     group=factor(c( rep("group1",2*N),rep("group2",2*N))),
     condition=factor(c( rep("baseline",N),rep("treatment",N),
         rep("baseline",N),rep("treatment",N) ) ),
      score=c(g1b,g1t,g2b,g2t))

    cat("\n",
     paste(rep("-",21)),"\n",
     sprintf("          \tExpected     \tActual"),"\n",
     sprintf("group1"),"\n",
     sprintf("baseline:\t%f\t\t%f", M1, mean(g1b)) ,"\n",
     sprintf("treatment:\t%f\t\t%f", M1+D1, mean(g1t)) ,"\n",
     sprintf("group2"),"\n",
     sprintf("baseline:\t%f\t\t%f", M2, mean(g2b)) ,"\n",
     sprintf("treatment:\t%f\t\t%f", M2+D2, mean(g2t)) ,"\n",
     paste(rep("-",21)),"\n"
    )
    return(data2)
}
```

```
> data<-gen.data(N=7,D1=50,D2=-50)

 - - - - - - - - - - - - - - - - - - - - -
            Expected      Actual
 group1
 baseline: 500.000000 497.549106
 treatment: 550.000000 558.862296
```

```
 group2
 baseline: 500.000000 471.583668
 treatment: 450.000000 442.534698
 - - - - - - - - - - - - - - - - - - - - -
> attach(data)
```

Do not forget to get a summary of your data. We called for 7 subjects in 2 groups and 2 conditions (baseline and treatment), which makes a total of 28 measurements (*score*). Group1 has 7 subjects and 14 measurements, and likewise for group2. There are 14 measures in baseline and treatment, and thus we have 14 differences between baseline and treatment. If the design is a repeated design we made 2 measures, one baseline and one treatment, for each subject, and we do not have more independent measures than the number of subjects since we have only one difference for each subject.

However, if the design is an independent groups design each measurement comes from an individual that is independent from all other subjects, and we have 28 measurements.

If the design was repeated or independent measures design comes from the description of how the measurements were obtained, i.e., the design of the experiment. We will perform both types analysis to show how the analysis proceeds, but keep in mind that numbers cannot protest if an analysis breaks the conditions for the analysis.

```
> summary(data)
    subject      group        condition      score
 1      : 2   group1:14   baseline :14   Min.   :358.6
 2      : 2   group2:14   treatment:14   1st Qu.:470.3
 3      : 2                              Median :510.7
 4      : 2                              Mean   :510.2
 5      : 2                              3rd Qu.:539.1
 6      : 2                              Max.   :642.3
 (Other):16
```

Relevant graphs for this data are boxplot and interaction plots. The boxplot will help to find the range of data, and will often help to pinpoint anomalies in the data, such as outliers (marked out side of the whiskers) and differences in variation. An interaction plot will display more clearly if there is a crossover interaction between the two factors. If there is, the main effects may disappear as the effect is in different directions for the groups.

```
boxplot(score~group*condition)
interaction.plot(group,condition,score)
```

6.4.2 Independent measures

The first analysis assumes that we have no repeated measures, i.e. all measurements come from different individuals.

```
> model <- aov(score~group*condition)
> summary(model)
                Df Sum Sq Mean Sq F value  Pr(>F)
group            1    984     984   0.261 0.61396
condition        1      0       0   0.000 0.99836
group:condition  1  44112   44112  11.704 0.00224 **
Residuals       24  90453    3769
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As can be seen from the table, there are no main effects and a significant interaction effect: F(1,24)=11.7, $p < 0.01$ (**). This indicates that the effect of the treatment is different in the two groups, which should be clear in the interaction plot.

The degrees of freedom

We know that we have 28 independent measures in this analysis that assumes independent measures. The degree of freedom for *group* is the number of groups - 1, analogous to losing one degree of freedom for having an overall mean for groups. For condition, the reasoning is analogous and we have the number of conditions (here, baseline and treatment) - 1. The degrees of freedom for the interaction is calculated by multiplying the degrees of freedom for group and condition (i.e., 1 times 1 = 1).

The degree of freedom for residuals, sometimes referred to as the *within* degrees of freedom, is calculated from the total degrees of freedom minus the sum of degrees of freedom explained (df for group, condition, and interaction). The total degree of freedom is N - 1, since we need to keep all but one measures given that we know the sum of all measures. N-1 is here 27, and the sum of all the *explained* degrees of freedom is 3, thus residuals have 24 degrees of freedom.

Note that R will calculate these numbers of you, but you need to check that the numbers make sense. There should never be a number for degrees of freedom that is higher than the number of independent measures. When you report the F value you have two parameters for F, the degree of freedom "between" (i.e., the number for degrees of freedom for the effect; here, 1

for each of group, condition, and interaction) and the degrees of freedom "within" (*residuals*). The only significant effect in the table above is the interaction effect, reported as F(1,24)=11.7, $p < 0.01$ (**). The groups do not react the same to the treatment. This might indeed be expect from the simulated difference of 50 between baseline and treatment, that was asked to go in different directions for the two groups.

6.4.3 Repeated measures

The second analysis assumes that we have repeated measures; each individual provided one baseline measurement and one treatment measurement. Thus the pairing is for the *condition* factor, and this can be expressed in the formula using the added *Error* term to state repeated measures for subject by condition. We tell the analysis that we have one measurement for baseline and one measurement for treatment for each subject.

```
> model <- aov(score~group*condition+Error(subject/condition))
> summary(model)

Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
group      1    984     984   0.306   0.59
Residuals 12  38607    3217

Error: subject:condition
                Df Sum Sq Mean Sq F value Pr(>F)
condition        1      0       0    0.00 0.9985
group:condition  1  44112   44112   10.21 0.0077 **
Residuals       12  51847    4321
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results we see that there is no significant effect of either group or condition, but there is a significant interaction effect: F(1,12)=10.2, $p < 0.01$ (**). The group was not a significant factor, as F(1,12)=0.59, and $p = 0.59 > 0.05$.

Note that the degrees of freedom are different, as we now have a total of 14 subjects, and thus 14 independent pairings of baseline and treatment. In this design a subject belongs to one group only, and cannot be switched between groups (you may think of groups here as an intrinsic factor, for example gender). Notice that the table is split: the first gives the main effect of group

(which has not been repeated), and the second one gives the repeated factor *condition* and its interaction with *group*. The significant interaction effect is reported formally F(1,12)=10.2, $p < 0.01$ (**). The interpretation is that groups do not react in the same way to the treatment.

When you graph the results it is a good idea to calculate paired differences between baseline and condition.

```
> b=score[condition=="baseline"]
> t=score[condition=="treatment"]
> d=b-t
> data2plot <- data.frame(subject=rep(1:14),
   group=c(rep("group1",7),rep("group2",7)),
   diff=d)
> attach(data2plot)
> boxplot(diff~group)
```

The main difference between the two designs is that the repeated measures design should sample 14 individual subjects, where as the independent measures design should have 28 individual subjects. The number of measurements is the same, but the cost of measurement is largely affected by the number of subjects, and little else. If we had repeated measures we need to account for that in the analysis, otherwise we risk artificially inflating the significance of the results.

6.5 2x3 Factorial Anova

We will assume that we have a situation with two groups (say male and female) that have provided measurements in three different conditions (this could be baseline, semantically related, phonologically related).

6.5.1 Generating Examples

The two groups have means M1 and M2, default is the same mean, 500. The standard deviation is assumed to be the same, 50. All can be set to different values, but the interesting part is to set the effects of the treatments for each group: D11, D12, D21,D22.

```
gen2x3.data <-function(M1=500,M2=M1,N=100,
   SD1=M1/10,SD2=M2/10,
   SD11=SD1,SD12=SD1,
   SD21=SD2,SD22=SD2,
```

```
   D11=SD1/2,D12=SD1/2,
   D21=SD2/2,D22=SD2/2) {
    g1base = rnorm(N, mean=M1, sd=SD1)
    g1a = rnorm(N, mean=M1+D11, sd=SD11)
    g1b = rnorm(N, mean=M1+D12, sd=SD12)
    g2base = rnorm(N, mean=M2, sd=SD2)
    g2a = rnorm(N, mean=M2+D21, sd=SD21)
    g2b = rnorm(N, mean=M2+D22, sd=SD22)

    data = data.frame(subject = factor(c(rep(1:N,3),
          rep((N+1):(N+N),3))),
      group = factor(c( rep("group1",3*N),rep("group2",3*N))),
      condition = factor( c( rep("baseline",N),
         rep("treatment1",N),rep("treatment2",N),
         rep("baseline",N),
         rep("treatment1",N),rep("treatment2",N) ) ),
       score=c(g1base,g1a,g1b,g2base,g2a,g2b) )

    cat("\n",
     paste(rep("-",21)),"\n",
     sprintf("          \tExpected     \tActual"),"\n",
     sprintf("group1"),"\n",
     sprintf("baseline:\t%f\t\t%f", M1, mean(g1base)) ,"\n",
     sprintf("treatment:\t%f\t\t%f", M1+D11, mean(g1a)) ,"\n",
     sprintf("treatment:\t%f\t\t%f", M1+D12, mean(g1b)) ,"\n",
     sprintf("group2"),"\n",
     sprintf("baseline:\t%f\t\t%f", M2, mean(g2base)) ,"\n",
     sprintf("treatment1:\t%f\t\t%f", M2+D21, mean(g2a)) ,"\n",
     sprintf("treatment2:\t%f\t\t%f", M1+D22, mean(g2b)) ,"\n",
     paste(rep("-",21)),"\n"
    )
    return(data)
}

> data23 <- gen2x3.data(N=20,D11=50,D12=75,D21=75,D22=100)

 - - - - - - - - - - - - - - - - - - - - -
           Expected      Actual
 group1
 baseline: 500.000000 507.846795
 treatment: 550.000000 558.565645
```

```
 treatment: 575.000000 587.138334
 group2
 baseline: 500.000000 503.484271
 treatment1: 575.000000 566.189299
 treatment2: 600.000000 587.349751
 - - - - - - - - - - - - - - - - - - - - -

> summary(data23)
    subject       group         condition      score
 1      :  3   group1:60   baseline  :40   Min.   :411.4
 2      :  3   group2:60   treatment1:40   1st Qu.:513.8
 3      :  3               treatment2:40   Median :555.0
 4      :  3                               Mean   :551.8
 5      :  3                               3rd Qu.:591.6
 6      :  3                               Max.   :688.8
 (Other):102
```

6.5.2 Independent measures

First the analysis for the model using independent measures.

```
> model23 <- aov(score~group*condition)
> summary(model23)
                 Df Sum Sq Mean Sq F value   Pr(>F)
group             1     40      40   0.017    0.898
condition         2 139862   69931  28.742 7.81e-11 ***
group:condition   2    732     366   0.150    0.861
Residuals       114 277367    2433
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the table we can report a main effect of condition: F(2,114)=28.7, $p < 0.001$ (***), and no other significant effects. Our next problem is to locate the effect. Which conditions are different? A boxplot might give a lead, but we need to perform a pairwise t-test.

```
> pairwise.t.test(score,condition)

Pairwise comparisons using t tests with pooled SD

data:  score and condition

           baseline treatment1
treatment1 1.7e-06  -
treatment2 4.4e-11  0.024

P value adjustment method: holm
```

From the pairwise test we can report that both treatments are significantly different from the baseline ($p < 0.001$), and the treatments are significantly different from each other, i.e. treatment 2 gives higher values (when we look at the boxplot, or the individual means).

6.5.3 Repeated measures

Assuming repeated measures.

```
> model23<-aov(score~group*condition+Error(subject/condition))
> summary(model23)
```

```
Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
group      1     40    40.2   0.015  0.904
Residuals 38 103100  2713.2

Error: subject:condition
                Df Sum Sq Mean Sq F value   Pr(>F)
condition        2 139862   69931   30.50 1.89e-10 ***
group:condition  2    732     366    0.16    0.853
Residuals       76 174267    2293
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the table we can see that there is a significant effect for condition using repeated measures factorial anova: F(2,76)=30.5, $p < 0.001$ (***). We have to use a pairwise t-test to find out the differences.

```
> pairwise.t.test(score,condition, paired=T)

Pairwise comparisons using paired t tests

data:  score and condition

           baseline treatment1
treatment1 1.1e-05  -
treatment2 3.0e-08  0.014

P value adjustment method: holm
```

From the pairwise test we can report that both treatments are significantly different from the baseline ($p < 0.001$), and the treatments are significantly different from each other ($p < 0.05$), i.e. treatment 2 gives higher values (when we look at the boxplot, or the individual means). The paired analysis reveal a slightly stronger significance for a difference between the treatments. The main difference is, however, that instead of 120 subjects we now only need to sample 40.

6.6 An advanced repeated measures

It is possible to sample repeated measures for several factors. The problem is to specify the model for the repeated measure. Let us consider an example.

This is what to do if you have been clever enough to sample repeated measures. (data in *handouts*)

Below is the data for the first subject. Data stems from an experiment by Pitt & Shoaf, which is discussed in Keith Johnson's book *Quantitative Linguistics* (Chapter on repeated measures design pp. 126-134.) Note that there is no planned "between subjects"-factor (such as gender). As you can see the data is paired for position, with either overlap zero or three. It is different words, to avoid repetition/learning effects. It is always the case that reaction time for zero is the first in order, and for three is the second (this has to be checked, and is related to the ordering of differences). Paired data is thus represented by the order in the data file.

The experiment uses a Prime to Target design. The prime word either has a zero or a three speech sounds overlap with the target word. The overlap could be occur at either of three positions in the words: early, mid, or late (i.e. head-rhyme, mid, rhyme).

How can we use the fact that we have paired data?

subj	word	overlap	position	rt
AALA	plan	zero	early	1065
AALA	must	three	early	1149
AALA	close	zero	mid	701
AALA	blade	three	mid	845
AALA	sense	zero	late	748
AALA	stage	three	late	701

If we study the table above we find out that the design of the experiment is to collect reaction time measures from each subject, and that each subject will contribute six measures. The subject will contributed measures for zero (e.g. grit – plan) and three speech sound overlap (e.g. **mus**k – **mus**t) for three different positions in the word (early, mid, and late). We better use the planned paired design. The pairing factor is *subject*, and the pairs are defined by *overlap* and *position*. Note that there is *only one pair* for each combination of subject, and overlap & position! If you have used a design where you have measured *many* pairs from the same subject you may need to explicitly calculate averages per condition, to avoid pseudo-replication. Pseudo-replication

might not be a major worry, if you are careful to give effect sizes. However, if you only give significance then you may decrease the standard errors so that you can detect very minute differences with significance, thus boosting the data and possibly mislead the reader of the experiment.

The pairing procedure is to explicitly construct the *Error*-term (i.e. where we expect individual variation to stem from). We want to model reaction time by the effect of overlap **and** position and we know that the variation is attributed to an individual measured in the conditions of overlap **and** position. We know that we have one (remaining) measurement for each combination of overlap and condition. One measurement for zero overlap at the early position, one measurement for three overlap at the early position and so forth (cf. the table above).

This can be expressed as a model specifying formula:

```
rt ~ overlap*position + Error( subj/(overlap*position) )
```

This states that we model reaction times by overlap and position and their interaction + the variation that stems from the subjects over the factors overlap and position and their interaction. This formula describes the pairing. The following will discuss how to analyze data from 97 different subject, each with 6 different reaction time measures.

Let us first specify the analysis of variance (aov) model.

```
> model=aov(rt~overlap*position+Error(subj/(overlap*position)))
```

Now look at the ANOVA table.

```
> summary(model)
```

The main effect for overlap (paired for subject) is:

```
Error: subj:overlap
          Df  Sum Sq Mean Sq F value Pr(>F)
overlap    1   19758   19758  1.2692 0.2627
Residuals 96 1494378   15566
```

Remember that we have two levels of overlap (zero, three). We have 97 different subjects, $->$ 96 df. 96 = df overlap * df subj; check. We may write this result:
Main effect for overlap is not significant ($F(1, 96) = 1.27$; $p > 0.25$).

Now we turn to position.

```
Error: subj:position
           Df  Sum Sq Mean Sq F value Pr(>F)
position    2  111658   55829    2.15 0.1193
Residuals 192 4985721   25967
```

There are (df position) * (df subject) = 2 * 96 = 192 df for errors. We may write this result: Main effect for position is not significant ($F(1, 192) = 2.15$; $p > 0.1$).

So where is the beef? We now turn to the *interaction* effect.

```
Error: subj:overlap:position
                  Df  Sum Sq Mean Sq F value    Pr(>F)
overlap:position   2  284424  142212  10.121 6.623e-05 ***
Residuals        192 2697804   14051
```

There are (df for overlap*position)*(df subj) = 192 df for errors. We may write this result: The interaction effect for overlap and position is highly significant ($F(2, 192) = 10.12$; $p < 0.001$; ***).

We have now established that we have an interaction effect, but where is it? We have paired for overlap. The effect might be specific to early, mid, or late overlap. If we look closely at the table we see that for the early and mid positions we get a slow down when there is a three sound overlap, whereas there is a speed up for the late position. The table only illustrates one subject, so this is in no sense conclusive. It illustrates what an interaction effect is: the effect on reaction time for an overlap in speech sounds between a prime word and a target word depends on in which position the overlap is present.

It is an art to get the best plot to illustrate the effect. Generally a boxplot is useful, and the boxplot can be specified using the same formulas, although the box plot will not do the pairing for you.

```
> boxplot(rt ~ overlap*position)
```

One idea would be to calculate differences between zero and three overlap for each subject and then plot the data. The interaction effect will come out clearer in such a plot. This is left as an exercise.

6.7 Summary

We have looked through three varieties of anova-designs: the oneway anova, a 2 by 2 design and a 2 by 3 design.

Note that it could be necessary to perform a pairwise search for specific effects. This is particularly true if a factor has more than two levels. In the planning of experiments it is a good idea to use factors with only two levels if possible, as the results are easier to interpret.

The data generating functions can be used with different initial conditions to simulate different scenarios, including main effects, interaction effects, pure additive effects, different starting points and differences in standard deviation. Each time the function is run it will give a new data set to work with, and you are encouraged to test many different scenarios and try to find the effects you specified using analysis of variance.

Chapter 7

Statistical Hypothesis Testing

7.1 Introduction

You might be familiar with the Game Show Host problem. The setting is a Game Show where one of the games displays three boxes with only one containing a prize. The player is asked to choose a box, and then the host of the game opens another box showing that box to be empty. Next the player is asked to change the box or keep the previously selected box. Surprisingly few people choose to change boxes. It seems that the situation is thought of as a fifty-fifty choice, as there are two boxes and one of them contains a price. However, it is better to switch the boxes.

The player would double the chances to win by switching. How is that possible? The answer is surprisingly simple. When the first choice was done there was one chance in three to pick the right box (given no special talent to pick the right box, e.g. premonition). That means that there were two chances in three that the prize was in the other two boxes; the host opened one of them. If the first box was picked at random, opening the second box just eliminates one of the alternatives. Changing the box changes the chances of winning from the initial one in three to two in three – simply because the game show host's action does not make your previous action any smarter.

Some may object that this assumes that the game show host knows where the prize is. This is not necessary, the condition is only that the opened box is empty. If the box was not empty you would have lost the game, and not be given a chance to choose. The empty box is the new information. Knowing the history gives an advantage.

The example shows that choice, as well as knowledge of what has happened before, may affect the analysis of a situation. This is an important insight, and may be worth considering when we analyze sequences of events.

Most examples in this book will assume that measure data were collected as independent data points, unlike in the example above. The data points may depend on fixed or random factors, known or unknown factors, that can be investigated in a statistical model. The measures we collect in such a model are examples for the dependent variable, i.e. the variable that depends on the factors specified in the model.

Let us consider a more extreme example. This show has 10 boxes. You pick one box at random. In nine out of ten cases the prize would be in one of the other boxes. Now the kind host opens 8 boxes, and all of them are empty. If you change to the remaining box you would win in 9 out of 10 cases.

Let us consider one complication. You have watched this show twice before, and you saw that in both cases the prize was in box number 9. You quickly form a hypothesis that the prize is always in number 9, since finding the same number the second round only would occur in one out ten cases.

How would you test this hypothesis, if given the chance to participate in the game? Would you choose number 9 (i.e. your hypothesis)? I would not, because then I would not be able to change the box, which I know would give me an unbiased high chance of success. You may also think of this as trying to disprove your hypothesis.

So I would pick any other box. If the remaining box, after the rest are opened, is number 9 I would switch to it and strengthen my hypothesis if the switch is the correct thing to do. We might think of this as *jumping to conclusions* because the uncertainty of the hypothesis will remain, only now it is a bit more unlikely that three shows in a row had the prize in the same box. If the remaining box was any other box I would switch to it (and my hypothesis would already have been disproved), swallow my pride and accept the 90% chance of winning. Of course, it would psychologically feel bad to have switched from a winning box if that happened, but I would accept it because I did the rational choice.

Note that in neither case my hypothesis would have been proved. If the prize really were in the hypothesized box, I would do better trying to disprove the hypothesis because if the hypothesis was correct I could switch to it, and if the hypothesis was wrong I would still have probabilities on my side. The standard mode of scientific reasoning is to disprove, rather than prove, a hypothesis. When the impossible is dismissed only the possible remains. Doubt rather than belief.

Given that the support of my hypothesis was only that two same numbers in a row only happens once out of ten, what was the chance of the hypothesis

being correct? In cases like this the principle of maximum uncertainty would say that since there are two alternatives each alternative would have equal opportunity: Either the hypothesis is correct or not. Note that the hypothesis had not been tested previously. The evidence for it was based on just two observations; and indeed any number on the second show would have had one chance in ten.

7.2 Significance and Decision making

The most common case of hypothesis testing is to test if an observed difference between two groups is due to chance or not. This is often done through assuming a model. For example, a model based on an underlying distribution around a central tendency (e.g., the mean) with some estimated variance. The model is then used to make a decision whether to accept or reject the null hypothesis that there is no difference between the groups. There are two types of wrong decisions that can obviously be made: rejecting the null hypothesis when it is true and accepting it when it is false. These errors are called a type 1 and type 2 errors.

7.3 Type I errors

The probability of falsely rejecting the null hypothesis (H_0), i.e., claiming a difference where there is no difference, can be calculated if we have representative samples. If the data level allows it we may calculate a correct estimate of the variance in the groups and a correct estimate of the central tendency (typically the mean). Other tests may rely on sorting data, and calculating the chance that the two groups would sort out differently; without calculating a central tendency or the variance needed in a so called parametrized test (where the parameters are the central tendency and the variance). The probability of falsely rejecting H_0 is labelled α. We would like this probability to be as low as possible. Some like to think of this as parallel to convicting an innocent person in a criminal court. It is better to keep the null hypothesis when evidence is lacking. You might be more familiar with the p-value, which is the obtained probability of observing an equally or more extreme difference given that H_0 is indeed true. The observed p-value is crucially dependent on representative samples and correct estimation of parameters.

One important point is that if significance is not reached it maybe due to a lack of evidence rather than proof of no difference.

Lack of randomization may work both for and against significance.

The effect might also be smaller than can be detected given the size of the sample.

However, in the popular literature the scenario is often oversimplified: it often looks like a lack of significance is a proof of no difference. What lack of significance means, is that we cannot claim that an observed difference is real. Significance is not a proof of a real difference, but an indication of whether we have made an observation of a difference or not.

Type I errors may occur because we have failed to get a representative sample from the population we want to study. When we have small samples, outliers in the samples may have a very large impact. We may also have failed to account for some unseen variable. Both outliers and the effect of hidden variables can be quite unexpected. One example is a recent study on a vaccine that in smaller study showed a significant effect, but in a larger study failed to show significance. One difference, apart from study size, was that nearly all subjects in the smaller study were recruited in spring, whereas the larger study took much longer and recruited a much smaller proportion in spring. The smaller first study might involve a type I error, or the second study may involve a type II error due to a hidden factor, such as seasonal variation in immune responses.

Results from explorative statistical hypothesis testing is rarely as clear cut as the popular story of scientific 'proof' from experimental studies reported in the media often claims. It could be noted that many scientific discoveries have been made as a result of failures to find expected statistical significance.

One example is the background noise, that is critical to the Big Bang Theory. This noise was detected in radio telescopes, and at first commonly attributed to artifacts such as bird droppings on the disc of the radio telescope (since it occurred wherever the telescope was directed). More insights into the source of the background noise led to a Nobel prize.

Another popular story stems from a popular story on the discovery of penicillin. Contamination of Petri dishes by molds is a fairly common problem. Usually this leads to the dish being thrown away, but one day a researcher took the time to look more closely at the dish, and found that no bacteria grew near the mold. An insight into why this was so (that the mold secreted some anti-bacterial substance) led to an important discovery that changed how many diseases were treated in the 20th century.

The point here is that there were observations that were made repeatedly, but attributed to 'noise' until somebody thought of a different explanation.

However, it is true that much standard science is about planning experiments that will have decisive power if a significant difference is observed.

When experiments fail to detect significance, it is usually associated with a delay in the scientific process, and often a delay in reaching academic qualifications. This is one reason academic research need to have some tolerance for failure, as sometimes these failures are starts of more original thoughts.

We never know if we have made an error. If the results are replicated in other studies, we can feel increasingly sure that there really is a difference. If other studies fail to replicate the results, there might be more hidden variables. Identifying the hidden variables is one worthy scientific task.

7.4 Type II errors

Let us first consider some examples of failing to reject H_0 (possible type II errors)

Consider a trial for a heart medication. There were two groups, one group was given the new medication and the other group was given standard treatment. The results could not find any significant difference in the outcome. However, when looking closer at the groups a difference was found, and the medication is approved for treatment. The two groups were chosen by self-selection (for ethical reasons). The alternative was heart surgery. It turned out that more severe patients had preferred medication to an operation that they had little confidence in, given their weakened state. This consideration made the data look much more favorable for an effect of the drug, after accounting for the condition of the patients.

Another example, related to research in writing, is the effect of writing styles. One difference that is sometimes discussed is the difference between linear writers (i.e., writers that write from start to end without many revisions) and more complex writers (i.e., writers that edit their texts while writing). One might think that the complex writers would also come up with more complicated texts, more complex sentence structures, less spelling errors. However, the difference is at best a small difference in favor of complex writers. Does this mean that the complex strategy has an advantage? Note that there is not a baseline for comparison. All the writers have chosen a strategy themselves. There is actually no indication which strategy is the best. Indeed it could be questioned if there is a strategy that is best for all. Perhaps the writers have chosen a strategy that fits their skills, or cognitive abilities or cognitive profile? One test might be to see what happens if the writers are encouraged to use the other strategy. However, this is hard to do as the habits of writing likely have developed over many years. Who knows, maybe particular groups may benefit from a more linear style of writing?

These examples highlight the need for a baseline, and the difficulties of comparing groups when self-selection is involved.

A type 2 error is the failure to reject H_0, when there really is a difference. This should be avoided, but it is very hard to avoid type 1 errors at the same time as avoiding type 2 errors. The probability of making a type 2 error is labeled β. In order to avoid falsely accepting H_0, we need to make sure that we have an experiment that is large enough to be able to detect a difference. This means we need to have an idea of the difference, either from previous experience or from knowing how large a difference needs to be to be worth consideration. We will also need to have an idea of the expected variance in the samples.

7.4.1 More on the source of errors

If we have detected a significant difference between groups, that means that the observed difference is likely not due to chance. The significance value indicate how often we would expect a difference of that magnitude, given that we have a representative sample, and we have succeeded in fulfilling the requirements of the tests. The differences should not be due to how the data was sampled, or how data was treated (e.g., rounded).

Researcher bias. Consider the case of the genius rats versus the intelligence challenged rats. One day a researcher comes in to the lab with a batch of rats that are claimed to be genetically engineered to be incompetent in finding their way in an experimental maze. The assistants happily go about testing the rats, and note down the results. The next week the researcher comes in with a new batch of rats, that are claimed to be genetically engineered genius rats, that will learn their way much faster than other rats. The assistants go about their work, and results are noted down. The results showed that there was a highly significant difference in favor of the genius rats. The researcher could finish his article on the behavior of his assistants. The rats were all genetically the same, but the assistants had somehow managed to get the results they expected. It is harder to know exactly what they did differently. Maybe they rounded number more in favor of the 'genius rats', or they might have given them a better treatment when starting the experiment. This would point to the need for blinded tests, where the researchers do not know the desired outcome for either the treated group or the control group.

Statistical significance means that the observed differences likely cannot

be explained by chance. This does not mean that the differences are due to the *explanation* implicit in our testing. The decision we take, when we get significance, is to reject the null hypothesis that there is no observed difference. The *interpretation* we want to make from this is that the difference is real, and not due to other possible sources for the differences, such as non-random sampling or consistent differences in handling the data. This is much harder to get a handle on.

Statistical significance does not necessarily mean that the observed difference is relevant to our research question; that this is so has to be argued and motivated by the researchers conducting the study.

Statistical significance does not necessarily mean that the difference is important. This also has to be argued and motivated by the researchers.

Statistical significance does not necessarily mean that we would have a much better chance than guessing, if we wanted to predict group from measurements. Very small differences may become significant, i.e. hard to explain by random errors, if the study is only large enough. This is related to the statistical power of the study, which can be calculated a priori for tests, given that we have either a pilot study or other reasons to estimate the effect size; for example from estimated population parameters such as variance and central tendency.

7.5 Candidates for type III errors

One obvious candidate for a third kind of error is being right for the wrong reason; this amounts to over-interpreting data. In the above game show host example, shifting to box number 9 because you believe the hypothesis that the prize is always in box number 9, is one example of doing the right choice for the wrong reason. It is indeed very rare that something is exactly what we think it is.

Note that keeping the H_0 is done by default, when significance is not achieved; so there is no need for an interpretation error that the two groups indeed are the same. When we fail to show significance it is simply that we do not have evidence for a difference. There is no other interpretation, and we do not know if the failure is because there is no difference or because we have a to small difference for detection.

7.6 The effect size

We may intuitively think of the effect size as the separation of the studied groups; how much does knowing a measure help predicting the group category? If two groups are completely separated by a measure then knowing the measure truly predicts the group.

In the case of a cross table test for nominal data, we might think of effect size as a measure of how much we have gained for predicting a factor by knowing the group, or vice versa. Remember that large studies may detect very minute details, that might not be so useful in the end.

There are excellent R-packages for calculating effect sizes, and these packages can help with effect sizes for other tests than the ones that will be demonstrated in this book. Once you have started to get past the beginner level, you should consider if there are better solutions to calculating the effect size. The suggestions below are mainly to demonstrate the logic behind effect sizes, and why we need effect size as a complementary measure to just statistical significance.

7.6.1 Effect size for a t-test

The first idea is based on the effect as measured by the difference in mean. You may do this with the absolute difference, and just note down which way the difference goes. The difference in means is then compared to the pooled standard deviation in the whole design. This can be estimated by simply combining all the data and calculate the standard deviation using sd.

7.6.2 Examples in R

Let us go through this in R. First we get two samples from a normal distribution (conducting a real experiment will do as well, but then we will not know the actual population parameters that are available in a theoretical distribution). Let us start with getting ten measures in two samples, x and y. The x sample has an expected mean of 0, and the mean of y is 0.5; thus the theoretical difference is 0.5 (y larger). The theoretical standard deviation is set at 1 in these example cases. However, we will have to calculate the actual mean and standard deviations as they are given by the extracted (random) samples. The variable pool is all the data taken together.

In R (sample size 10) :

```
x=rnorm(10,mean=0,sd=1)
y=rnorm(10,mean=0.5,sd=1)
```

Look at:

```
sd(x)
sd(y)
mean(x)
mean(y)
diff=mean(y)-mean(x)
diff
```

The sd function gives the standard deviation, i.e. the value for one step along the normal distribution. The calculation diff, gives the difference between the means (i.e. the peaks) of sample x and y. The samples are taken from populations were we expect the mean of sample x to be 0, and the mean of sample y to be 0.5. The standard deviation of both samples is expected to be 1 for each sample.

One run gave the following, within two decimals:

```
sd(x) = 0.94, sd(y)=0.74, sd(pool)=0.80
mean(x)= -0.27 mean(y)=0.45
diff=0.71 (kept all decimals)
```

You will get other values, because you will get different random samples.

Estimate Cohen's d

```
pool = sqrt(((10-1)*sd(y)*sd(y)+(10-1)*sd(x)*sd(x))/(10+10))
diff=mean(y)-mean(x)
d=diff/pool
```

With the above values, Cohen's d=0.88.You may get different values, depending on the values in your sample.

Cohen's d is sometimes interpreted in terms of a small, medium or large effect. A small effect would be a d around 0.2 to 0.3, a medium effect around 0.5, and a large effect may start at 0.8. Thus we would have a large effect size in the example. It should be noted that what constitutes small, medium and large effects depends on the kind of data and the field of research. Values above 1 are possible, in that case that the effect is much larger than the random fluctuations in the samples; i.e. the signal is then very good and would give a good chance to predict the group of an individual from knowing the value for the individual.

Let us repeat the simulation for larger samples.

```
x=rnorm(100,mean=0,sd=1)
y=rnorm(100,mean=0.5,sd=1)
```

One run gave the following, within two decimals:

```
sd(x) = 0.94, sd(y)=1.09, pool=1.01
mean(x)= -0.27 mean(y)=0.45
diff=0.41 (kept all decimals)
```

With the above values, Cohen's d=0.40.

The theoretical value, from the given population parameters would be d=0.5/1; so the larger sample did result in a better estimation of the effect size. The larger sample gives an estimated medium size, whereas the small sample indicated a large effect.

Estimate Glass's Δ

An alternative version of the above idea is called Glass's Δ. The main difference is that the standard deviations are not pooled, but rather the standard deviation of the control group is used. The main reason for using Δ, as an estimate of effect size, is when many groups are compared to a control. Glass's Δ makes it easier to compare to the same standard, as pooling pairwise, as in Cohen's d, would give a range of various pooled standard deviations.

7.6.3 Effect size for a paired t-test

If you use paired data in your study (e.g., you compare two measurements from the same individual) you may use Pearson r as an effect size measure. It is very attractive to use a paired test if possible, because much of the variance might be explained by individual variance that may be cancelled out between measurements from the same individual; for example, a slow or fast individual keeps that characteristic in both measurements.

Let us simulate this in R. First take a random sample, x, of 50 values (mean=0, sd=1). For the paired effect we are looking for, add a random sample of 50 values with a mean of 1 (sd=1) to create sample y. We now have samples that are related to each other. On the average, 1 has been added to each number in x (i.e. random numbers that average to one have been added to the first sample). Plot the values, to see if there is a pattern. The pattern should be less like a cloud, and more aligned along a line. The function cor gives the Pearson r that measures the correlation between x and y.

```
x = rnorm(50)
y=x+rnorm(50,mean=1)
plot(x,y)
```

```
r=cor(x,y)
r2=r*r
```

One run gave $r = 0.65$ (cor(x,y) gives Pearson r); you may get other values depending on the samples you drew from the normal distribution. It may also be interesting to give the r^2, sometimes called coefficient of determination, as this value tells how much of the variance is shared between the samples. The above value for r^2 would indicate that about 42% of the variance is shared. A qualitative classification, stemming from the social sciences, indicates a small effect when $r > 0.1$, a medium effect when $r > 0.3$, and a large effect when $r > 0.5$. A paired t-test of the example run (t.test(x,y,paired=T)) gave a very good p-value, so significance would have been achieved.

7.7 The power of test

The power of a test is related to the question of how likely it is that we would find statistical significance if we select a new randomized sample from the same population. We assume that we have correct estimates of standard deviation and central tendency.

Formally statistical power is the probability of rejecting the H_0 when the H_0 is false, i.e. the probability of not committing a type II error. It may be noted by $1-\beta$ where β is the probability of a type II error (i.e., failing to find a significant difference when there really is a difference); β is also referred to as the false negativity rate. Analysis of power is highly valuable to get the size of the experiment needed to find out if there is a difference. Without enough power, a rejection of the H_0 might very well be due to statistical accident.

A very useful online tool for Power Analysis can be found at

`http://www.divms.uiowa.edu/~rlenth/Power`

7.8 Exercise

Use the function power.t.test to find out the minimum number of subjects (measurements) in each group, if you have an expected difference of 0.5 in the population, and standard deviation of 1. Assume equal sized groups. You may try with different powers and significance levels. If more advanced power analysis is required, there are several packages in R that may be downloaded and installed. One is the package called pwr.

In R:

```
power.t.test(sig.level=0.05, power=0.8, delta=0.5, sd=1)
```

$Power = 0.8$ is a common choice if the significance level is 0.05. Recall β is the risk of keeping H_0 when it is false; i.e. not detecting a significant difference, $power = 1-\beta$. A four times higher risk of a type II error ($power = 1 - (4 * \alpha)$) is a common trade off between type I and II errors. However, if we are interested in reducing type II errors (maybe because we want to show that there likely is no difference), we should set β much lower, and thus power higher. The number in each group should be rounded up to the nearest integer; $n = 64$ in the example.

You can now go about testing how the prediction holds up, by simulating selecting groups. In R:

```
x=rnorm(64,mean=0,sd=1)
y=rnorm(64,mean=0.5,sd=1)
t.test(x,y)
```

Was significance ($p < 0.05$) achieved? The code above assigns 64 values from normal distributions with a mean of 0 and 0.5 respectively, and a standard deviation of 1 in both cases. You may check that the values are independent, visually, by plotting x and y.

```
plot(x,y)
```

The more round the cloud is the lower the correlation between x and y. You may also check standard deviation in both x and y.

```
sd(x)
sd(y)
```

Are the standard deviations close to the expected value (1)?

Repeat 20 times, and see how many times the test reach significance. We would expect failure to reach significance about 4 times out of 20.

```
x=rnorm(64,mean=0,sd=1)
y=rnorm(64,mean=0.5,sd=1)
t.test(x,y)
```

Try what happens with smaller groups (n=10) and an expected larger difference. Specify n=10, and try various values of delta, to calculate the power.

```
power.t.test(n=10, sig.level=0.05, delta=1.35, sd=1)
```

Redo the simulations.

```
x=rnorm(10,mean=0,sd=1)
y=rnorm(10,mean=1.35,sd=1)
t.test(x,y)
```

You may notice that the significance is not always near 0.05, but could be much better or worse, although the difference in the population is indeed the same in these cases (because we know the population, which is drawn directly from the normal distribution). The proportion of achieved significance ($p < 0.05$) is usually fairly close to the power in these simulations; but there might be deviations. Expecting not to achieve significance in 2 out of 10 trials is of course a fairly high risk, if this had been the final work before defending a thesis, or if the expense of each trial is high. For very important work, higher power would be reasonable, and that means larger samples. We would like to avoid the trap of repeating the experiment until significance is reached in one trial. For applications for funding it is often required that the power analysis has been done, and that it is properly motivated.

The lesson is that significance tells us if an observation is likely due to chance or not, but it does not , in itself, give a safe estimate of actual effect size. Larger samples, or repeating for several smaller samples, would help estimate the real effect size.

7.9 Summary

Null hypothesis testing tries to determine if there is a difference between two groups. The testing procedure keeps the null hypothesis by default, and try to estimate if there is enough evidence to claim a difference that is likely not due to the expected variance in the samples. Significance tells if an observation might be due to chance or if it is likely it is not. If a difference is significant, the test procedure is neutral on the cause of the difference. The difference could be for the reasons the investigators did the tests, but it could also be due to other causes: unexpected outliers, or hidden factors that were not part of the testing, or non-representative samples or the experimental procedure. If significance is achieved it is the task of the investigators to interpret what this means.

Significance should be complemented by effect size to determine if the difference is important. There are many ways that a significance result might be misleading. There is always a statistical risk that a difference really is due to random variance. This risk can be controlled but never completely eliminated. Any systematic difference, even unrelated to the task, may also

result in detecting a difference that is marked as significant (i.e., not explained by the estimated variance).

The risk (α) of making a type I error, claiming a difference when there actually is none, is affected by the sampling procedure and the size of the sample. Type II errors are related to failing to find a difference when there actually is one. The power of a test makes it possible to calculate the size of the groups, which is needed to plan the experiment. We may also start from a known group size and calculate either the minimum size of the effect or the maximum variance. Information about the expected difference may either be estimated from previous experiments, or from a stipulation of the minimum difference that we would like to consider. Likewise, the expected standard deviation from the mean may be estimated from previous experience with similar tasks.

Finally, there is always a chance that the effect we observe stems from other causes than the ones we have included in the experiments (i.e., to be 'right' for the wrong reasons, pointing to the wrong causation). An anecdotal example is early research indicating that aspirin might be effective for schizophrenia. However the results failed to replicate, and it was later discovered that one formulation contained lithium, and the other did not. Lithium was shown to have effect, and aspirin did not have an effect. Looking at the history of lithium, it shows that medical use of lithium may stem back to ancient times, and it was used to treat mania in the late 1800's. The anecdote may be just an anecdote, but it illustrates how unknown factors may suddenly pop up in experiments, and it is impossible to control all possible factors. It might be uncomfortable news for the budding scientist that experiments will never definitely prove any hypothesis, but never subjecting a hypothesis to a test is worse. Building knowledge on an untested hypothesis risks that everything built on the assumption of the hypothesis will fail. Therefore it is often not enough that one experiment shows a significant effect. If several experiments, all large enough to reliably detect an actual effect, really show an effect then we can proceed with one more piece of assumed knowledge. This is relatively rare, but fortunately we may also gain insights from failed experiments if we analyze them correctly.