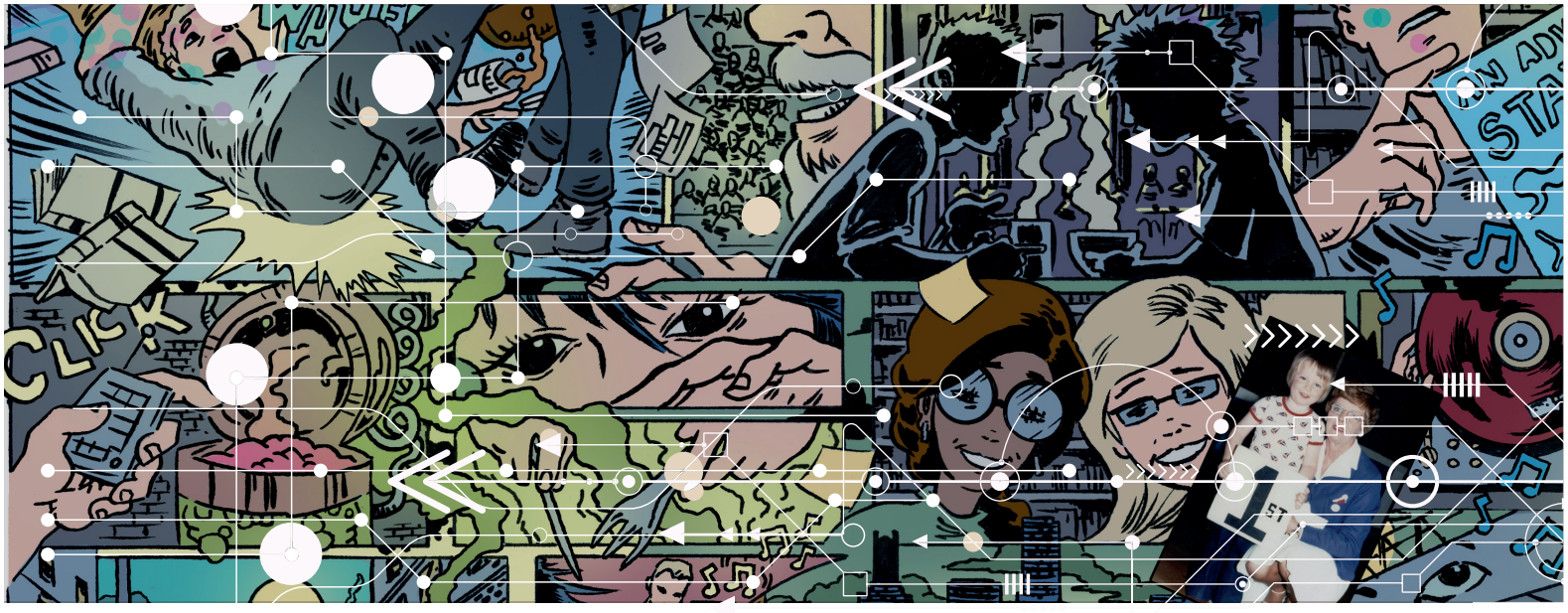
DISCOVERING STATISTICS USING IBM SPSS STATISTICS

CATISFIED CUSTOMERS

5TH EDITION

DISCOVERING STATISTICS USING IBM SPSS STATISTICS



ANDY FIELD

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

First edition published 2000
Second edition published 2005
Third editon published 2009. Reprinted 2009, 1010, 2011 (twice), 2012
Fourth edition published 2013. Reprinted 2014 (twice), 2015, 2016, 2017

Library of Congress Control Number: 2017954636

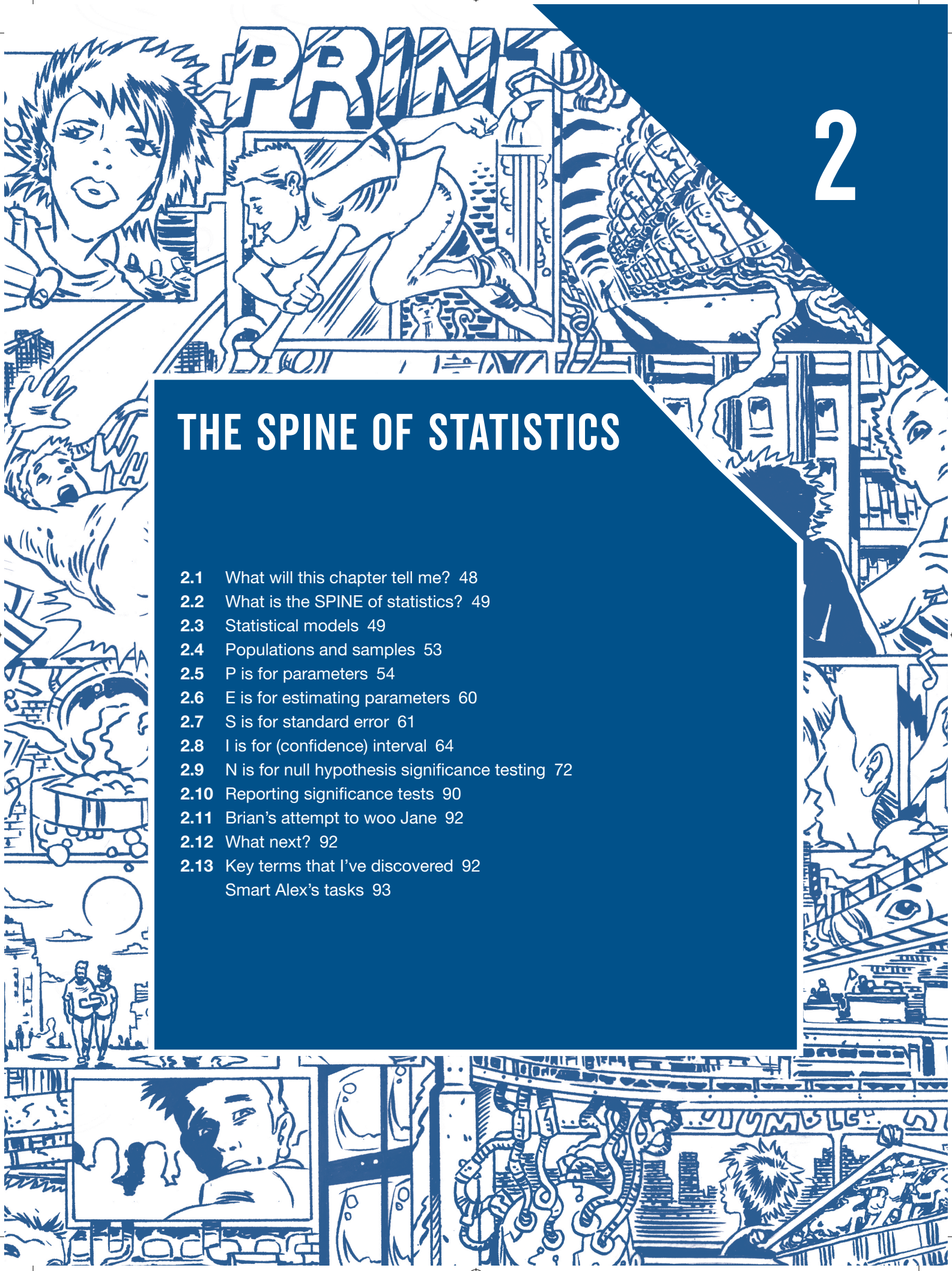
British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

Editor: Jai Seaman
Development editors: Sarah Turpie & Nina Smith
Assistant editors, digital: Chloe Statham
Production editor: Ian Antcliff
Copyeditor: Richard Leigh
Indexer: David Rudeforth
Marketing manager: Ben Griffin-Sherwood
Cover design: Wendy Scott
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in Germany by: Mohn Media Mohndruck GmbH

Illustrated by: James Iles

ISBN 978-1-5264-1951-4
ISBN 978-1-5264-1952-1 (pbk)



2

THE SPINE OF STATISTICS

2.1 What will this chapter tell me?

Although I had learnt a lot about golf clubs randomly appearing out of nowhere and hitting me around the face, I still felt that there was much about the world that I didn't understand. For one thing, could I learn to predict the presence of these golf clubs that seemed inexplicably drawn towards my apparently magnetic head? A child's survival depends upon being able to predict reliably what will happen in certain situations; consequently they develop a model of the world based on the data they have (previous experience) and they then test this model by collecting new data/experiences. Based on how well the new experiences fit with their original model, a child might revise their model of the world.

According to my parents (conveniently I have no memory of these events), while at nursery school one model of the world that I was enthusiastic to try out was 'If I get my penis out, it will be really funny'. To my considerable disappointment, this model turned out to be a poor predictor of positive outcomes. Thankfully for all concerned, I soon revised this model of the world to be 'If I get my penis out at nursery school the teachers and mummy and daddy will be quite annoyed'. This revised model may not have been as much fun but was certainly a better 'fit' of the observed data. Fitting models that accurately reflect the observed data is important to establish whether a hypothesis (and the theory from which it derives) is true.

Figure 2.1 The face of innocence ... but what are the hands doing?

You'll be relieved to know that this chapter is not about my penis but is about fitting statistical models. We edge sneakily away from the frying pan of research methods and trip accidentally into the fires of statistics hell. We will start to see how we can use the properties of data to go beyond our observations and to draw inferences about the world at large. This chapter and the next lay the foundation for the rest of the book.

2.2 What is the SPINE of statistics?

To many students, statistics is a bewildering mass of different tests, each with their own set of equations. The focus is often on 'difference'. It feels like you need to learn a lot of different stuff. What I hope to do in this chapter is to focus your mind on some core concepts that many statistical models have in common. In doing so, I want to set the tone for you focusing on the *similarities* between statistical models rather than the differences. If your goal is to use statistics as a tool, rather than to bury yourself in the theory, then I think this approach makes your job a lot easier. In this chapter, I will first argue that most statistical models are variations on the very simple idea of predicting an outcome variable from one or more predictor variables. The mathematical form of the model changes, but it usually boils down to a representation of the relations between an outcome and one or more predictors. If you understand that, then there are five key concepts to get your head around. If you understand these, you've gone a long way towards understanding any statistical model that you might want to fit. They are the SPINE of statistics, which is a clever acronym for:

- **S**tandard error
- **P**arameters
- **I**nterval estimates (confidence intervals)
- **N**ull hypothesis significance testing
- **E**stimation

I cover each of these topics, but not in this order because PESIN doesn't work nearly so well as an acronym.[1]

2.3 Statistical models

We saw in the previous chapter that scientists are interested in discovering something about a phenomenon that we assume exists (a 'real-world' phenomenon). These real-world phenomena can be anything from the behaviour of interest rates in the economy to the behaviour of undergraduates at the end-of-exam party. Whatever the phenomenon, we collect data from the real world to test predictions from our hypotheses about that phenomenon. Testing these hypotheses involves building statistical models of the phenomenon of interest.

Let's begin with an analogy. Imagine an engineer wishes to build a bridge across a river. That engineer would be pretty daft if she built any old bridge, because it might fall down. Instead, she collects data from the real world: she looks at existing bridges and sees from what materials they are made, their structure, size and so on (she might even collect data about whether these bridges are still standing). She uses this information to construct an idea of what her new bridge will be (this is a 'model'). It's expensive and impractical for her to build a full-sized version of her bridge, so she builds a scaled-down version. The model may differ from reality in several ways – it will be smaller, for a start – but the engineer will try to build a model that best fits the situation of interest based on the data available. Once the model has been built, it can be used to predict things about the real world: for example, the

1 There is another, more entertaining, acronym that fits well with the anecdote at the start of the chapter, but I decided not to use it because in a séance with Freud he advised me that it could lead to pesin envy.

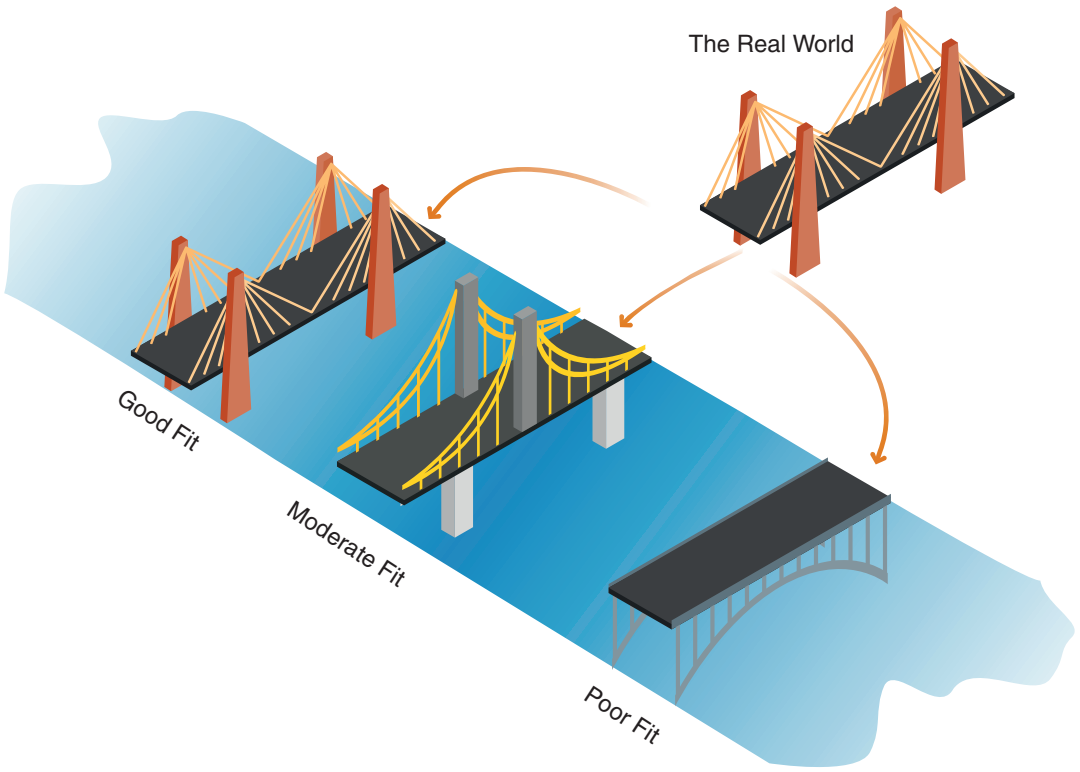
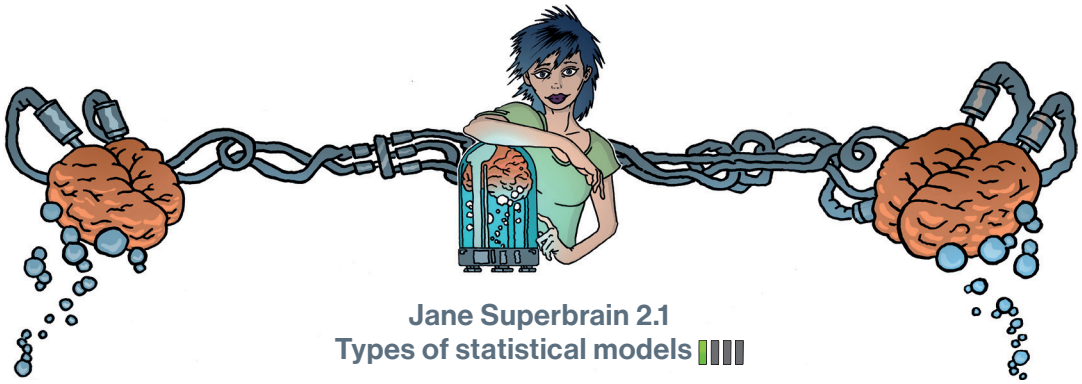


Figure 2.2 Fitting models to real-world data (see text for details)

engineer might test whether the bridge can withstand strong winds by placing her model in a wind tunnel. It is important that the model accurately represents the real world, otherwise any conclusions she extrapolates to the real-world bridge will be meaningless.

Scientists do much the same: they build (statistical) models of real-world processes to predict how these processes operate under certain conditions (see Jane Superbrain Box 2.1). Unlike engineers, we don't have access to the real-world situation and so we can only *infer* things about psychological, societal, biological or economic processes based upon the models we build. However, like the engineer, our models need to be as accurate as possible so that the predictions we make about the real world are accurate too; the statistical model should represent the data collected (the *observed data*) as closely as possible. The degree to which a statistical model represents the data collected is known as the **fit** of the model.

Figure 2.2 shows three models that our engineer has built to represent her real-world bridge. The first is an excellent representation of the real-world situation and is said to be a *good fit*. If the engineer uses this model to make predictions about the real world then, because it so closely resembles reality, she can be confident that her predictions will be accurate. If the model collapses in a strong wind, then there is a good chance that the real bridge would collapse also. The second model has some similarities to the real world: the model includes some of the basic structural features, but there are some big differences too (e.g., the absence of one of the supporting towers). We might consider this model to have a *moderate fit* (i.e., there are some similarities to reality but also some important differences). If our engineer uses this model to make predictions about the real world then her predictions could be inaccurate or even catastrophic. For example, perhaps the model predicts that the bridge will collapse in a strong wind, so after the real bridge is built it gets closed every time a strong wind occurs, creating 100-mile tailbacks with everyone stranded in the snow, feasting on the crumbs of old sandwiches that



Jane Superbrain 2.1
Types of statistical models

Scientists (especially behavioural and social ones) tend to use **linear models**, which are models based on a straight line. As you read scientific research papers, you'll see that they are riddled with 'analysis of variance (ANOVA)' and 'regression', which are identical statistical systems based on the linear model (Cohen, 1968). In fact, most of the chapters in this book explain this 'general linear model'.

Imagine we were interested in how people evaluated dishonest acts.[2] Participants evaluate the dishonesty of acts based on watching videos of people confessing to those acts. Imagine we took 100 people and showed them a random dishonest act described by the perpetrator. They then evaluated the honesty of the act (from 0 = appalling behaviour to 10 = it's OK really) and how much they liked the person (0 = not at all, 10 = a lot).

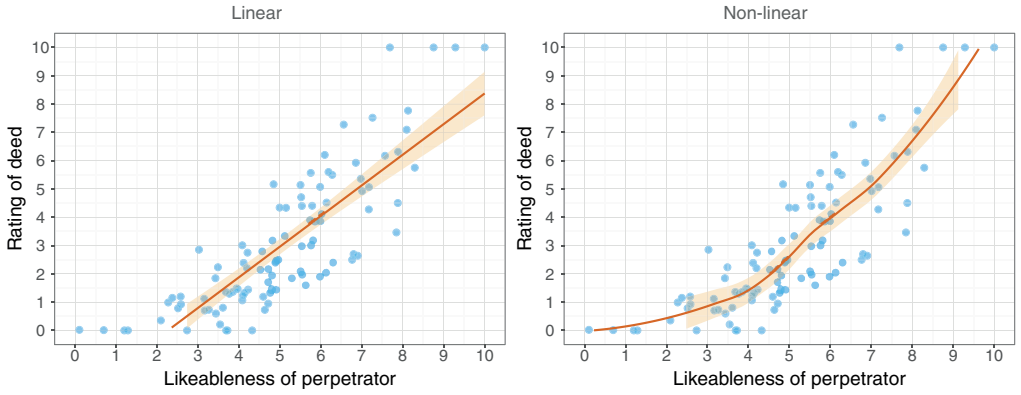
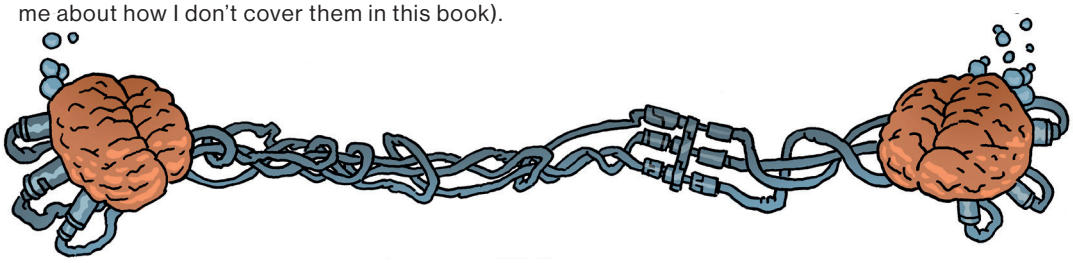


Figure 2.3 A scatterplot of the same data with a linear model fitted (left), and with a non-linear model fitted (right)

We can represent these hypothetical data on a scatterplot in which each dot represents an individual's rating on both variables (see Section 5.8). Figure 2.3 shows two versions of the same data. We can fit different models to the same data: on the left we have a linear (straight) model and on the right a non-linear (curved) one. Both models show that the more likeable a perpetrator is, the more positively people view their dishonest act. However, the curved line shows a subtler pattern: the trend to be more forgiving of likeable people kicks in when the likeableness rating rises above 4. Below 4 (where the perpetrator is not very likeable), all deeds are rated fairly low (the red line is quite flat), but as the perpetrator becomes likeable (above about 4) the slope of the line becomes steeper, suggesting that as likeableness rises above this value, people become increasingly forgiving of dishonest acts. Neither of the two models is necessarily correct, but one model will fit the data better than the other; this is why it is important for us to assess how well a statistical model fits the data.

2 This example came from a media story about the Honesty Lab set up by Stefan Fafinski and Emily Finch. However, this project no longer seems not to exist and I can't find the results published anywhere. I like the example though, so I kept it.

Linear models tend to get fitted to data because they are less complex and because non-linear models are often not taught (despite 900 pages of statistics hell, I don't get into non-linear models in this book). This could have had interesting consequences for science: (1) many published statistical models might not be the ones that fit best (because the authors didn't try out non-linear models); and (2) findings might have been missed because a linear model was a poor fit and the scientists gave up rather than fitting non-linear models (which perhaps would have been a 'good enough' fit). It is useful to plot your data first: if your plot seems to suggest a non-linear model then don't apply a linear model because that's all you know, fit a non-linear model (after complaining to me about how I don't cover them in this book).



they find under the seats of their cars. All of which turns out to be unnecessary because the real bridge was safe – the prediction from the model was wrong because it was a bad representation of reality. We can have a little confidence in predictions from this model, but not complete confidence. The final model is completely different than the real-world situation; it bears no structural similarities to the real bridge and is a *poor fit*. Any predictions based on this model are likely to be completely inaccurate. Extending this analogy to science, if our model is a poor fit to the observed data, then the predictions we make from it will be equally poor.

Although it's easy to visualize a model of a bridge, you might be struggling to understand what I mean by a 'statistical model'. Even a brief glance at some scientific articles will transport you into a terrifying jungle of different types of 'statistical model': you'll encounter tedious-looking names like *t*-test, ANOVA, regression, multilevel models, and structural equation modelling. It might make you yearn for a career in journalism, where the distinction between opinion and evidence need not trouble you. Fear not, though; I have a story that may help.

Many centuries ago there existed a cult of elite mathematicians. They spent 200 years trying to solve an equation that they believed would make them immortal. However, one of them forgot that when you multiply two minus numbers you get a plus, and instead of achieving eternal life they unleashed Cthulhu from his underwater city. It's amazing how small computational mistakes in maths can have these sorts of consequences. Anyway, the only way they could agree to get Cthulhu to return to his entrapment was by promising to infect the minds of humanity with confusion. They set about this task with gusto. They took the simple and elegant idea of a statistical model and reinvented it in hundreds of seemingly different ways (Figures 2.4 and 2.5). They described each model as though it were completely different from the rest. 'Ha!' they thought, 'that'll confuse students.' And confusion did indeed infect the minds of students. The statisticians kept their secret that all statistical models could be described in one simple, easy-to-understand equation locked away in a wooden box with Cthulhu's head burned into the lid. 'No one will open a box with a big squid head burnt into it', they reasoned. They were correct, until a Greek fisherman stumbled upon the box and, thinking it contained some vintage calamari, opened it. Disappointed with the contents, he sold the script inside the box on eBay. I bought it for €3 plus postage. This was money well spent, because it means that I can now give you the key that will unlock the mystery of statistics for ever. Everything in this book (and statistics generally) boils down to equation (2.1).

$$\text{outcome}_i = (\text{model}) + \text{error}_i \tag{2.1}$$

Figure 2.4 Thanks to the Confusion machine, a simple equation is made to seem like lots of unrelated tests

This equation means that the data we observe can be predicted from the model we choose to fit plus some amount of error.[3] The 'model' in the equation will vary depending on the design of your study, the type of data you have and what it is you're trying to achieve with your model. Consequently, the model can also vary in its complexity. No matter how long the equation that describes your model might be, you can just close your eyes, reimagine it as the word 'model' (much less scary) and think of the equation above: we predict an outcome variable from some model (that may or may not be hideously complex) but we won't do so perfectly so there will be some error in there too. Next time you encounter some sleep-inducing phrase like 'hierarchical growth model', just remember that in most cases it's just a fancy way of saying 'predicting an outcome from some variables'.

2.4 Populations and samples

Before we get stuck into a specific form of statistical model, it's worth remembering that scientists are usually interested in finding results that apply to an entire **population** of entities. For example, psychologists want to discover processes that occur in all humans, biologists might be interested in processes that occur in all cells, economists want to build models that apply to all salaries, and so on. A population can be very general (all human beings) or very narrow (all male ginger cats called Bob). Usually, scientists strive to infer things about general populations rather than narrow ones. For example, it's not very interesting to conclude that psychology students with brown hair who own a pet hamster named George recover more quickly from sports injuries if the injury is massaged (unless you happen to be a psychology student with brown hair who has a pet hamster named George, like René Koning).[4] It will have a much wider impact if we can conclude that *everyone's* (or most people's) sports injuries are aided by massage.

3 The little i (e.g., outcome$_i$) refers to the ith score. Imagine, we had three scores collected from Andy, Zach and Zoë. We could replace the i with their name, so if we wanted to predict Zoë's score we could change the equation to: outcome$_{Zoë}$ = model + error$_{Zoë}$. The i reflects the fact that the value of the outcome and the error will be different for each person.

4 A brown-haired psychology student with a hamster called Sjors (Dutch for George, apparently), who emailed me to weaken my foolish belief that I'd generated an obscure combination of possibilities.

Remember that our bridge-building engineer could not make a full-sized model of the bridge she wanted to build and instead built a small-scale model and tested it under various conditions. From the results obtained from the small-scale model she inferred things about how the full-sized bridge would respond. The small-scale model may respond differently than a full-sized version of the bridge, but the larger the model, the more likely it is to behave in the same way as the full-sized bridge. This metaphor can be extended to scientists: we rarely, if ever, have access to every member of a population (the real-sized bridge). Psychologists cannot collect data from every human being, and ecologists cannot observe every male ginger cat called Bob. Therefore, we collect data from a smaller subset of the population known as a **sample** (the scaled-down bridge) and use these data to infer things about the population as a whole. The bigger the sample, the more likely it is to reflect the whole population. If we take several random samples from the population, each of these samples will give us slightly different results but, on average, the results from large samples should be similar.

2.5 P is for parameters

Remember that parameters are the 'P' in the SPINE of statistics. Statistical models are made up of variables and **parameters**. As we have seen, variables are measured constructs that vary across entities in the sample. In contrast, parameters are not measured and are (usually) constants believed to represent some fundamental truth about the relations between variables in the model. Some examples of parameters with which you might be familiar are: the mean and median (which estimate the centre of the distribution) and the correlation and regression coefficients (which estimate the relationship between two variables).

Statististicans try to confuse you by giving estimates of different parameters different symbols and letters ($\overline{X}$ for the mean, r for the correlation, b for regression coefficients) but it's much less confusing if we just use the letter b. If we're interested only in summarizing the outcome, as we are when we compute a mean, then we won't have any variables in the model, only a parameter, so we could write our equation as:

$$\text{outcome}_i = (b_0) + \text{error}_i \tag{2.2}$$

However, often we want to predict an outcome from a variable, and if we do this we expand the model to include this variable (predictor variables are usually denoted with the letter X). Our model becomes:

$$\text{outcome}_i = (b_0 + b_1 X_i) + \text{error}_i \tag{2.3}$$

Now we're predicting the value of the outcome for a particular entity (i) not just from the value of the outcome when there are no predictors (b_0) but from the entity's score on the predictor variable (X_i). The predictor variable has a parameter (b_1) attached to it, which tells us something about the relationship between the predictor (X_i) and outcome.

If we want to predict an outcome from two predictors then we can add another predictor to the model too:

$$\text{outcome}_i = (b_0 + b_1 X_{1i} + b_2 X_{2i}) + \text{error}_i \tag{2.4}$$

In this model, we're predicting the value of the outcome for a particular entity (i) from the value of the outcome when there are no predictors (b_0) and the entity's score on two predictor variables (X_{1i} and X_{2i}). Each predictor variable has a parameter (b_1, b_2) attached to it, which tells us something about the relationship between that predictor and the outcome. We could carry on expanding the model with

more variables, but that will make our brains hurt, so let's not. In each of these equations I have kept brackets around the model, which aren't necessary, but I think it helps you to see which part of the equation is the model in each case.

Hopefully what you can take from this section is that this book boils down to a very simple idea: we can predict values of an outcome variable based on a model. The form of the model changes, but there will always be some error in prediction, and there will always be parameters that tell us about the shape or form of the model.

To work out what the model looks like, we estimate the parameters (i.e., the value(s) of b). You'll hear the phrase 'estimate the parameter' or 'parameter estimates' a lot in statistics, and you might wonder why we use the word 'estimate'. Surely statistics has evolved enough that we can compute exact values of things and not merely estimate them? As I mentioned before, we're interested in drawing conclusions about a population (to which we don't have access). In other words, we want to know what our model might look like in the whole population. Given that our model is defined by parameters, this amounts to saying that we're not interested in the parameter values in our sample, but we care about the parameter values in the population. The problem is that we don't know what the parameter values are in the population because we didn't measure the population, we measured only a sample. However, we can use the sample data to *estimate* what the population parameter values are likely to be. That's why we use the word 'estimate', because when we calculate parameters based on sample data they are only estimates of what the true parameter value is in the population. Let's make these ideas a bit more concrete with a very simple model indeed: the mean.

2.5.1 The mean as a statistical model

We encountered the mean in Section 1.8.4, where I briefly mentioned that it was a statistical model because it is a hypothetical value and not necessarily one that is observed in the data. For example, if we took five statistics lecturers and measured the number of friends that they had, we might find the following data: 1, 2, 3, 3 and 4. If we want to know the mean number of friends, this can be calculated by adding the values we obtained, and dividing by the number of values measured: $(1 + 2 + 3 + 3 + 4)/5 = 2.6$. It is impossible to have 2.6 friends (unless you chop someone up with a chainsaw and befriend their arm, which is probably not beyond your average statistics lecturer) so the mean value is a *hypothetical* value: it is a model created to summarize the data and there will be error in prediction. As in equation (2.2), the model is:

$$\text{outcome}_i = (b_0) + \text{error}_i$$

in which the parameter, b_0, is the mean of the outcome. The important thing is that we can use the value of the mean (or any parameter) computed in our sample to estimate the value in the population (which is the value in which we're interested). We give estimates little hats to compensate them for the lack of self-esteem they feel at not being true values. Who doesn't love a hat?

$$\text{outcome}_i = (\hat{b}_0) + \text{error}_i \tag{2.5}$$

When you see equations where these little hats are used, try not to be confused, all the hats are doing is making explicit that the values underneath them are estimates. Imagine the parameter as wearing a little baseball cap with the word 'estimate' printed along the front. In the case of the mean, we estimate the population value by assuming that it is the same as the value in the sample (in this case 2.6).

Figure 2.5 Thanks to the Confusion machine, there are lots of terms that basically refer to error

2.5.2 Assessing the fit of a model: sums of squares and variance revisited

It's important to assess the fit of any statistical model (to return to our bridge analogy, we need to know how representative the model bridge is of the bridge that we want to build). With most statistical models we can determine whether the model represents the data well by looking at how different the scores we observed in the data are from the values that the model predicts. For example, let's look what happens when we use the model of the mean to predict how many friends the first lecturer has. The first lecture was called Andy; it's a small world. We observed that lecturer 1 had one friend and the model (i.e., the mean of all lecturers) predicted 2.6. By rearranging equation (2.1) we see that this is an error of –1.6:[5]

$$\begin{aligned} \text{outcome}_{\text{lecturer 1}} &= \hat{b}_0 + \varepsilon_{\text{lecturer 1}} \\ 1 &= 2.6 + \varepsilon_{\text{lecturer 1}} \\ \varepsilon_{\text{lecturer 1}} &= 1 - 2.6 \\ &= -1.6 \end{aligned} \tag{2.6}$$

You might notice that all we have done here is calculate the deviance, which we encountered in Section 1.8.5. The *deviance* is another word for *error* (Figure 2.5). A more general way to think of the deviance or error is by rearranging equation (2.1) into:

$$\text{deviance} = \text{outcome}_i - \text{model}_i \tag{2.7}$$

In other words, the error or deviance for a particular entity is the score predicted by the model for that entity subtracted from the corresponding observed score. Figure 2.6 shows the number of friends that each statistics lecturer had, and the mean number that we calculated earlier on. The line representing the mean can be thought of as our model, and the dots are the observed data. The diagram also has a series of vertical lines that connect each observed value to the mean value. These lines represent the

5 Remember that I'm using the symbol $\hat{b}_0$ to represent the mean. If this upsets you then replace it (in your mind) with the more traditionally used symbol, $\bar{X}$.

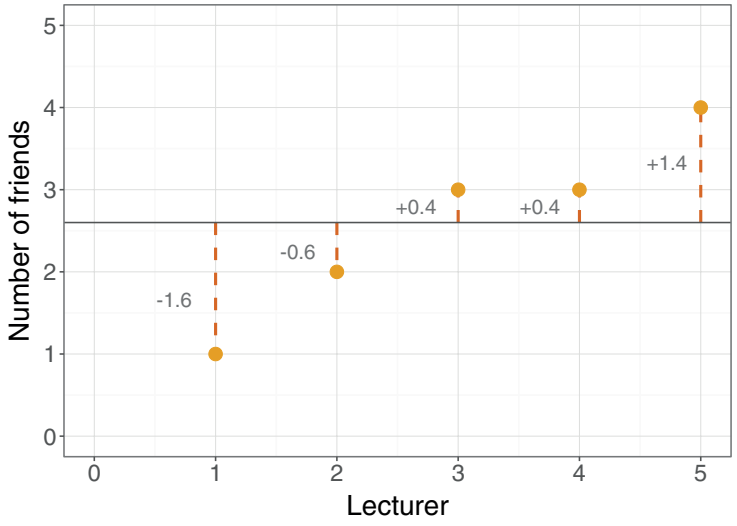


Figure 2.6 Graph showing the difference between the observed number of friends that each statistics lecturer had, and the mean number of friends

error or deviance of the model for each lecturer. The first lecturer, Andy, had only 1 friend (a glove puppet of a pink hippo called Professor Hippo) and we have already seen that the error for this lecturer is –1.6; the fact that it is a negative number shows that our model *overestimates* Andy's popularity: it predicts that he will have 2.6 friends but, in reality, he has only 1 (bless!).

We know the accuracy or 'fit' of the model for a particular lecturer, Andy, but we want to know the fit of the model *overall*. We saw in Section 1.8.5 that we can't add deviances because some errors are positive and others negative and so we'd get a total of zero:

$$\begin{aligned}\text{total error} &= \text{sum of errors}\\ &= \sum_{i=1}^{n}(\text{outcome}_i - \text{model}_i)\\ &= (-1.6)+(-0.6)+(0.4)+(0.4)+(1.4)=0\end{aligned} \tag{2.8}$$

We also saw in Section 1.8.5 that one way around this problem is to square the errors. This would give us a value of 5.2:

$$\begin{aligned}\text{sum of squared errors (SS)} &= \sum_{i=1}^{n}(\text{outcome}_i - \text{model}_i)^2\\ &= (-1.6)^2+(-0.6)^2+(0.4)^2+(0.4)^2+(1.4)^2\\ &= 2.56+0.36+0.16+0.16+1.96\\ &= 5.20\end{aligned} \tag{2.9}$$

Does this equation look familiar? It ought to, because it's the same as equation (1.6) for the sum of squares in Section 1.8.5 – the only difference is that equation (1.6) was specific to when our model is the mean, so the 'model' was replaced with the symbol for the mean ($\bar{x}$), and the outcome was replaced by the letter x, which is commonly used to represent a score on a variable (Eq. 2.10).

$$\sum_{i=1}^{n}(\text{outcome}_i - \text{model}_i)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.10}$$

However, when we're thinking about models more generally, this illustrates that we can think of the total error in terms of this general equation:

$$\text{total error} = \sum_{i=1}^{n} (\text{observed}_i - \text{model}_i)^2 \tag{2.11}$$

This equation shows how something we have used before (the sum of squares) can be used to assess the total error in any model (not just the mean).

We saw in Section 1.8.5 that although the sum of squared errors (SS) is a good measure of the accuracy of our model, it depends upon the quantity of data that has been collected – the more data points, the higher the SS. We also saw that we can overcome this problem by using the average error, rather than the total. To compute the average error we divide the sum of squares (i.e., the total error) by the number of values (N) that we used to compute that total. We again come back to the problem that we're usually interested in the error in the model in the population (not the sample). To estimate the mean error in the population we need to divide not by the number of scores contributing to the total, but by the **degrees of freedom** (df), which is the number of scores used to compute the total adjusted for the fact that we're trying to estimate the population value (Jane Superbrain Box 2.2):

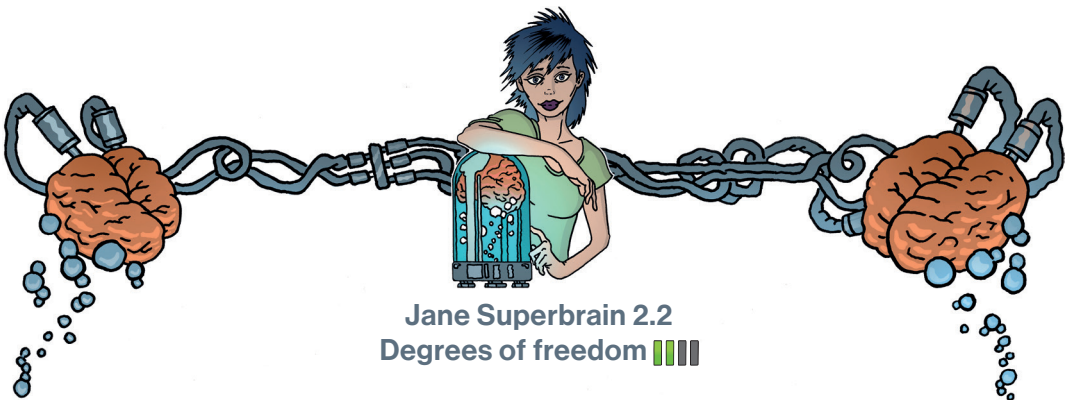
$$\text{mean squared error} = \frac{\text{SS}}{df} = \frac{\sum_{i=1}^{n} (\text{outcome}_i - \text{model}_i)^2}{N-1} \tag{2.12}$$

Does this equation look familiar? Again, it ought to, because it's a more general form of the equation for the variance (Eq. 1.7). Our model is the mean, so let's replace the 'model' with the mean ($\bar{x}$), and the 'outcome' with the letter x (to represent a score on the outcome). Lo and behold, the equation transforms into that of the variance:

$$\text{mean squared error} = \frac{\text{SS}}{df} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N-1} = \frac{5.20}{4} = 1.30 \tag{2.13}$$

To sum up, we can use the sum of squared errors and the mean squared error to assess the fit of a model. The mean squared error is also known as the variance. As such, the variance is a special case of a more general principle that we can apply to more complex models, which is that the fit of the model can be assessed with either the sum of squared errors or the mean squared error. Both measures give us an idea of how well a model fits the data: large values relative to the model indicate a lack of fit. Think back to Figure 1.10, which showed students' ratings of five lectures given by two lecturers. These lecturers differed in their mean squared error:[6] lecturer 1 had a smaller mean squared error than lecturer 2. Compare their graphs: the ratings for lecturer 1 were consistently close to the mean rating, indicating that the mean is a good representation of the observed data – it is a good fit. The ratings for lecturer 2, however, were more spread out from the mean: for some lectures, she received very high ratings, and for others her ratings were terrible. Therefore, the mean is not such a good representation of the observed scores – it is a poor fit.

6 I reported the standard deviation, but this value is the square root of the variance (a.k.a. the mean square error).



Jane Superbrain 2.2
Degrees of freedom

The concept of degrees of freedom (*df*) is very difficult to explain. I'll begin with an analogy. Imagine you're the manager of a sports team (I'll try to keep it general so you can think of whatever sport you follow, but in my mind I'm thinking about soccer). On the morning of the game you have a team sheet with (in the case of soccer) 11 empty slots relating to the positions on the playing field. Different players have different positions on the field that determine their role (defence, attack, etc.) and to some extent their physical location (left, right, forward, back). When the first player arrives, you have the choice of 11 positions in which to place this player. You place their name in one of the slots and allocate them to a position (e.g., striker) and, therefore, one position on the pitch is now occupied. When the next player arrives, you have 10 positions that are still free: you have 'a degree of freedom' to choose a position for this player (they could be put in defence, midfield, etc.). As more players arrive, your choices become increasingly limited: perhaps you have enough defenders so you need to start allocating some people to attack, where you have positions unfilled. At some point, you will have filled 10 positions and the final player arrives. With this player you have no 'degree of freedom' to choose where he or she plays – there is only one position left. In this scenario, there are 10 degrees of freedom: for 10 players you have a degree of choice over where they play, but for 1 player you have no choice. The degrees of freedom are one less than the number of players.

In statistical terms, the degrees of freedom relate to the number of observations that are free to vary. If we take a sample of four observations from a population, then these four scores are free to vary in any way (they can be any value). We use these four sampled values to estimate the mean of the population. Let's say that the mean of the sample was 10 and, therefore, we estimate that the population mean is also 10. The value of this parameter is now fixed: we have held one parameter constant. Imagine we now want to use this sample of four observations to estimate the mean squared error in the population. To do this, we need to use the value of the population mean, which we estimated to be the fixed value of 10. With the mean fixed, are all four scores in our sample free to be sampled? The answer is no, because to ensure that the population mean is 10 only three values are free to vary. For example, if the values in the sample we collected were 8, 9, 11, 12 (mean = 10) then the first value sampled could be any value from the population, say 9. The second value can also be any value from the population, say 12. Like our football team, the third value sampled can also be any value from the population, say 8. We now have values of 8, 9 and 12 in the sample. The final value we sample, the final player to turn up to the soccer game, cannot be any value in the population, it *has to be* 11 because this is the value that makes the mean of the sample equal to 10 (the population parameter that we have held constant). Therefore, if we hold one parameter constant, then the degrees of freedom must be one fewer than the number of scores used to calculate that parameter. This fact explains why when we use a sample to estimate the mean squared error (or indeed the standard deviation) of a population, we divide the sums of squares by $N - 1$ rather than N alone. There is a lengthier explanation in one of my other books (Field, 2016).

2.6 E is for estimating parameters

We have seen that models are defined by parameters, and these parameters need to be estimated from the data that we collect. Estimation is the 'E' in the SPINE of statistics. We used an example of the mean because it was familiar, but it will also illustrate a general principle about how parameters are estimated. Let's imagine that one day we walked down the road and fell into a hole. Not just any old hole, though, but a hole created by a rupture in the space-time continuum. We slid down the hole, which turned out to be a sort of U-shaped tunnel under the road, and we emerged out of the other end to find that not only were we on the other side of the road, but we'd gone back in time a few hundred years. Consequently, statistics had not been invented and neither had the equation to compute the mean. Happier times, you might think. A slightly odorous and beardy vagrant accosts you, demanding to know the average number of friends that a lecturer has. If we didn't know the equation for computing the mean, how might we do it? We could guess and see how well our guess fits the data. Remember, we want the value of the parameter $\hat{b}_0$ in this equation:

$$\text{outcome}_i = \hat{b}_0 + \text{error}_i$$

We know already that we can rearrange this equation to give us the error for each person:

$$\text{error}_i = \text{outcome}_i - \hat{b}_0$$

If we add the error for each person, then we'll get the sum of squared errors, which we can use as a measure of 'fit'. Imagine we begin by guessing that the mean number of friends that a lecturer has is 2. We can compute the error for each lecturer by subtracting this value from the number of friends they actually had. We then square this value to get rid of any minus signs, and we add up these squared errors. Table 2.1 shows this process, and we find that by guessing a value of 2, we end up with a total squared error of 7. Now let's take another guess; this time we'll guess the value is 4. Again, we compute the sum of squared errors as a measure of 'fit'. This model (i.e., guess) is worse than the last because the total squared error is larger than before: it is 15. We could carry on guessing and calculating the error for each guess. We *could* – if we were nerds with nothing better to do – but you're probably rad hipsters too busy doing whatever it is that rad hipsters do. I, however, am a badge-wearing nerd, so I have plotted the results in Figure 2.7, which shows the sum of squared errors that you would get for various values of the parameter $\hat{b}_0$. Note that, as we just calculated, when b is 2 we get an error of 7, and when it is 4 we get an error of 15. The shape of the line is interesting, though, because it curves to a minimum value – a value that produces the lowest sum of squared errors. The value of b at the lowest point of the curve is 2.6, and it produces an error of 5.2. Do these values seem familiar? They should, because they are the values of the mean and sum of squared errors that we calculated earlier. This example illustrates that the equation for the mean is designed to estimate that parameter to minimize the error. In other words, it is the value that has the least error. This doesn't necessarily mean that the value is a *good* fit to the data, but it is a better fit than any other value you might have chosen.

Throughout this book, we will fit lots of different models, with parameters other than the mean that need to be estimated. Although the equations for estimating these parameters will differ from that of the mean, they are based on this principle of minimizing error: they will give you the parameter that has the least error given the data you have. Again, it's worth reiterating that this is not the same thing as the parameter being accurate, unbiased or representative of the population: it could just be the best of a bad bunch. This section has focused on the principle of minimizing the sum of squared errors, and this is known as the **method of least squares** or **ordinary least squares OLS**. However, we'll also encounter other estimation methods later in the book.

Table 2.1 Guessing the mean

Number of Friends (x_i)	b_1	Squared error $(x_i - b_1)^2$	b_2	Squared error $(x_i - b_2)^2$
1	2	1	4	9
2	2	0	4	4
3	2	1	4	1
3	2	1	4	1
4	2	4	4	0
		$\sum_{i=1}^{n}(x_i - b_1)^2 = 7$		$\sum_{i=1}^{n}(x_i - b_2)^2 = 15$

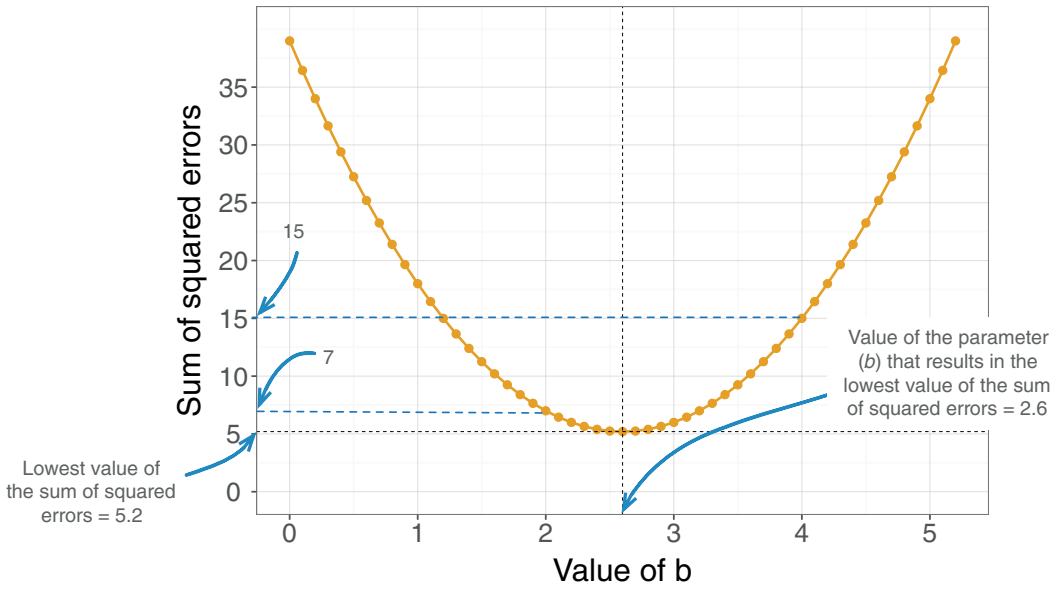


Figure 2.7 Graph showing the sum of squared errors for different 'guesses' of the mean

2.7 S is for standard error

We have looked at how we can fit a statistical model to a set of observations to summarize those data. It's one thing to summarize the data that you have actually collected, but in Chapter 1 we saw that good theories should say something about the wider world. It is one thing to be able to say that a sample of high-street stores in Brighton improved profits by placing cats in their store windows, but it's more useful to be able to say, based on our sample, that all high-street stores can increase profits by placing cats in their window displays. To do this, we need to go beyond the data, and to go beyond the data we need to begin to look at how representative our samples are of the population of interest. This idea brings us to the 'S' in the SPINE of statistics: the *standard error*.

In Chapter 1 we saw that the standard deviation tells us about how well the mean represents the sample data. However, if we're using the sample mean to estimate this parameter in the population, then we need to know how well it represents the value in the population, especially because samples

from a population differ. Imagine that we were interested in the student ratings of all lecturers (so lecturers in general are the population). We could take a sample from this population, and when we do we are taking one of many possible samples. If we were to take several samples from the same population, then each sample would have its own mean, and some of these sample means will be different. Figure 2.8 illustrates the process of taking samples from a population. Imagine for a fleeting second that we eat some magic beans that transport us to an astral plane where we can see for a few short, but beautiful, seconds the ratings of all lectures in the world. We're in this astral plane just long enough to compute the mean of these ratings (which, given the size of the population, implies we're there for quite some time). Thanks to our astral adventure we know, as an absolute fact, that the mean of all ratings is 3 (this is the *population mean*, μ, the parameter that we're trying to estimate).

Back in the real world, where we don't have magic beans, we also don't have access to the population, so we use a sample. In this sample we calculate the average rating, known as the *sample mean*, and discover it is 3; that is, lecturers were rated, on average, as 3. 'That was fun,' we think to ourselves, 'Let's do it again.' We take a second sample and find that lecturers were rated, on average, as only 2. In other words, the sample mean is different in the second sample than in the first. This difference illustrates **sampling variation**: that is, samples vary because they contain different members of the population; a sample that, by chance, includes some very good lecturers will have a higher average than a sample that, by chance, includes some awful lecturers.

Imagine that we're so excited by this sampling malarkey that we take another seven samples, so that we have nine in total (as in Figure 2.8). If we plotted the resulting sample means as a frequency distribution, or histogram,[7] we would see that three samples had a mean of 3, means of 2 and 4 occurred in two samples each, and means of 1 and 5 occurred in only one sample each. The end result is a nice symmetrical distribution known as a **sampling distribution**. A sampling distribution is the frequency distribution of sample means (or whatever parameter you're trying to estimate) from the same population. You need to imagine that we're taking hundreds or thousands of samples to construct a sampling distribution – I'm using nine to keep the diagram simple. The sampling distribution is a bit like a unicorn: we can imagine what one looks like, we can appreciate its beauty, and we can wonder at its magical feats, but the sad truth is that you'll never see a real one. They both exist as ideas rather than physical things. You would never go out and actually collect thousands of samples and draw a frequency distribution of their means, instead very clever statisticians have worked out what these distributions look like and how they behave. Likewise, you'd be ill-advised to search for unicorns.

The sampling distribution of the mean tells us about the behaviour of samples from the population, and you'll notice that it is centred at the same value as the mean of the population (i.e., 3). Therefore, if we took the average of all sample means we'd get the value of the population mean. We can use the sampling distribution to tell us how representative a sample is of the population. Think back to the standard deviation. We used the standard deviation as a measure of how representative the mean was of the observed data. A small standard deviation represented a scenario in which most data points were close to the mean, whereas a large standard deviation represented a situation in which data points were widely spread from the mean. If our 'observed data' are *sample* means then the standard deviation of these sample means would similarly tell us how widely spread (i.e., how representative) sample means are around their average. Bearing in mind that the average of the sample means is the same as the population mean, the standard deviation of the sample means would therefore tell us how widely sample means are spread around the population mean: put another way, it tells us whether sample means are typically representative of the population mean.

7 This is a graph of possible values of the sample mean plotted against the number of samples that have a mean of that value – see Section 1.8.1 for more details.

The standard deviation of sample means is known as the **standard error of the mean** (**SE**) or **standard error** for short. In the land where unicorns exist, the standard error could be calculated by taking the difference between each sample mean and the overall mean, squaring these differences, adding them up, and then dividing by the number of samples. Finally, the square root of this value would need to be taken to get the standard deviation of sample means: the standard error. In the real world, it would be crazy to collect hundreds of samples, and so we compute the standard error from a mathematical approximation. Some exceptionally clever statisticians have demonstrated something called the **central limit theorem**, which tells us that as samples get large (usually defined as greater

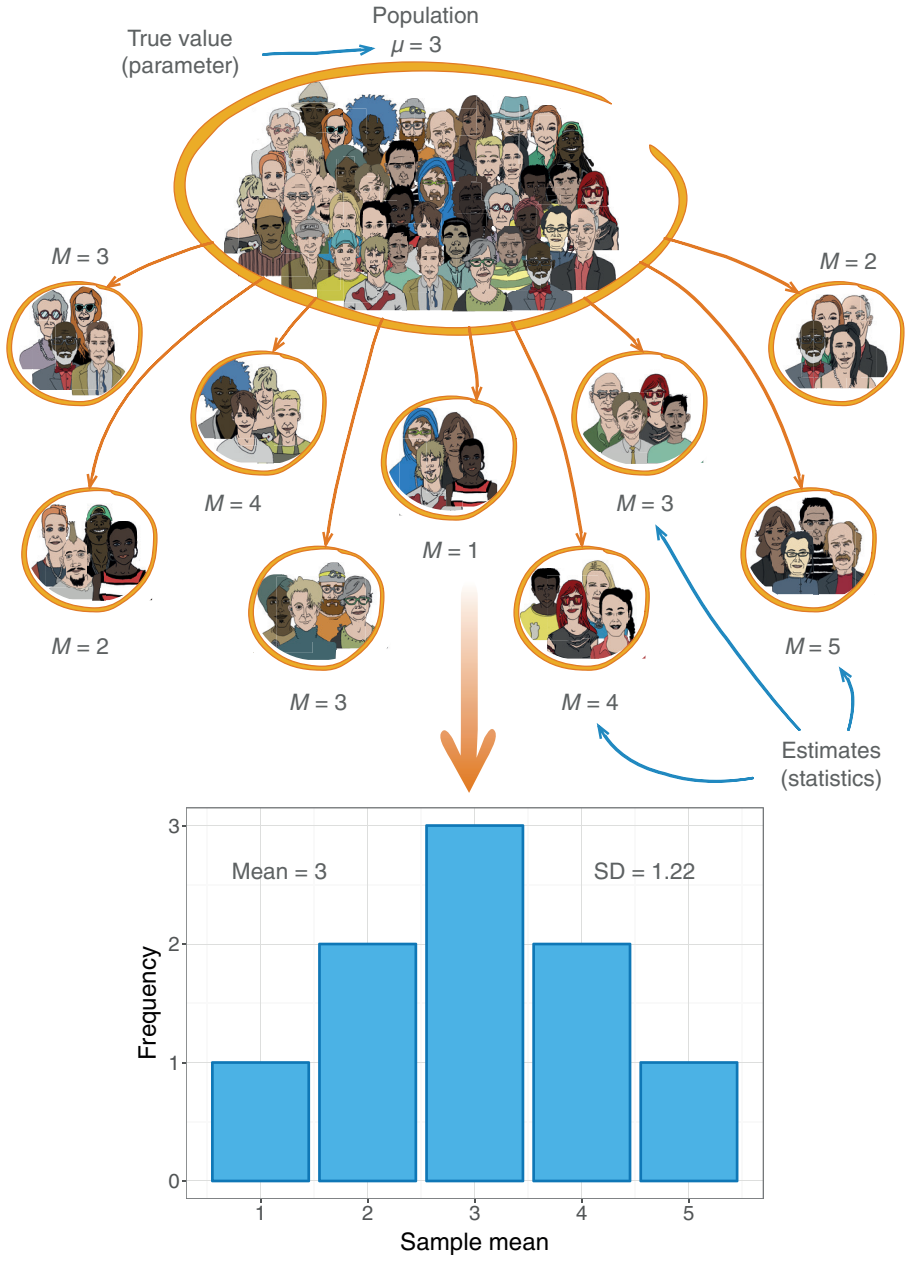


Figure 2.8 Illustration of the standard error (see text for details)

than 30), the sampling distribution has a normal distribution with a mean equal to the population mean, and a standard deviation shown in equation (2.14):

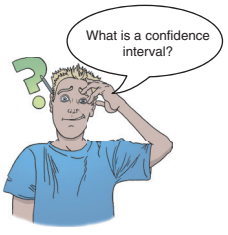
$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}} \tag{2.14}$$

We will return to the central limit theorem in more detail in Chapter 6, but I've mentioned it here because it tells us that if our sample is large we can use equation (2.14) to approximate the standard error (because it is the standard deviation of the sampling distribution).[8] When the sample is relatively small (fewer than 30) the sampling distribution is not normal: it has a different shape, known as a *t*-distribution, which we'll come back to later. A final point is that our discussion here has been about the mean, but everything we have learnt about sampling distributions applies to other parameters too: any parameter that can be estimated in a sample has a hypothetical sampling distribution and standard error.

Cramming Sam's Tips
The standard error

- The standard error of the mean is the standard deviation of sample means. As such, it is a measure of how representative of the population a sample mean is likely to be. A large standard error (relative to the sample mean) means that there is a lot of variability between the means of different samples and so the sample mean we have might not be representative of the population mean. A small standard error indicates that most sample means are similar to the population mean (i.e., our sample mean is likely to accurately reflect the population mean).

2.8 I is for (confidence) interval



The 'I' in the SPINE of statistics is for 'interval'; confidence interval, to be precise. As a brief recap, we usually use a sample value as an estimate of a parameter (e.g., the mean) in the population. We've just seen that the estimate of a parameter (e.g., the mean) will differ across samples, and we can use the standard error to get some idea of the extent to which these estimates differ

8 In fact, it should be the *population* standard deviation (σ) that is divided by the square root of the sample size; however, it's rare that we know the standard deviation of the population and, for large samples, this equation is a reasonable approximation.

across samples. We can also use this information to calculate boundaries within which we believe the population value will fall. Such boundaries are called **confidence intervals**. Although what I'm about to describe applies to any parameter, we'll stick with the mean to keep things consistent with what you have already learnt.

2.8.1 Calculating confidence intervals

Domjan, Blesbois, & Williams (1998) examined the learnt release of sperm in Japanese quail. The basic idea is that if a quail is allowed to copulate with a female quail in a certain context (an experimental chamber), then this context will serve as a cue to a mating opportunity and this in turn will affect semen release (although during the test phase the poor quail were tricked into copulating with a terry cloth with an embalmed female quail head stuck on top).[9] Anyway, if we look at the mean amount of sperm released in the experimental chamber, there is a true mean (the mean in the population); let's imagine it's 15 million sperm. Now, in our sample, we might find the mean amount of sperm released was 17 million. Because we don't know what the true value of the mean is (the population value), we don't know how good (or bad) our sample value of 17 million is as an estimate of it. So rather than fixating on a single value from the sample (the **point estimate**), we could use an **interval estimate** instead: we use our sample value as the midpoint, but set a lower and upper limit as well. So, we might say, we think the true value of the mean sperm release is somewhere between 12 million and 22 million sperm (note that 17 million falls exactly between these values). Of course, in this case, the true value (15 million) does falls within these limits. However, what if we'd set smaller limits – what if we'd said we think the true value falls between 16 and 18 million (again, note that 17 million is in the middle)? In this case the interval does not contain the population value of the mean.

Let's imagine that you were particularly fixated with Japanese quail sperm, and you repeated the experiment 100 times using different samples. Each time you did the experiment you constructed an interval around the sample, mean as I've just described. Figure 2.9 shows this scenario: the dots represent the mean for each sample, with the lines sticking out of them representing the intervals for these means. The true value of the mean (the mean in the population) is 15 million and is shown by a vertical line. The first thing to note is that the sample means are different from the true mean (this is because of sampling variation as described earlier). Second, although most of the intervals do contain the true mean (they cross the vertical line, meaning that the value of 15 million sperm falls somewhere between the lower and upper boundaries), a few do not.

The crucial thing is to construct the intervals in such a way that they tell us something useful. For example, perhaps we might want to know how often, in the long run, an interval contains the true value of the parameter we're trying to estimate (in this case, the mean). This is what a confidence interval does. Typically, we look at 95% confidence intervals, and sometimes 99% confidence intervals, but they all have a similar interpretation: they are limits constructed such that, for a certain percentage of samples (be that 95% or 99%), the true value of the population parameter falls within the limits. So, when you see a 95% confidence interval for a mean, think of it like this: if we'd collected 100 samples, and for each sample calculated the mean and a confidence interval for it (a bit like in Figure 2.9), then for 95 of these samples, the confidence interval contains the value of the mean in the population, and in 5 of the samples the confidence interval does not contain the population mean. The trouble is, you do not

9 This may seem a bit sick, but the male quails didn't appear to mind too much, which probably tells us all we need to know about males.

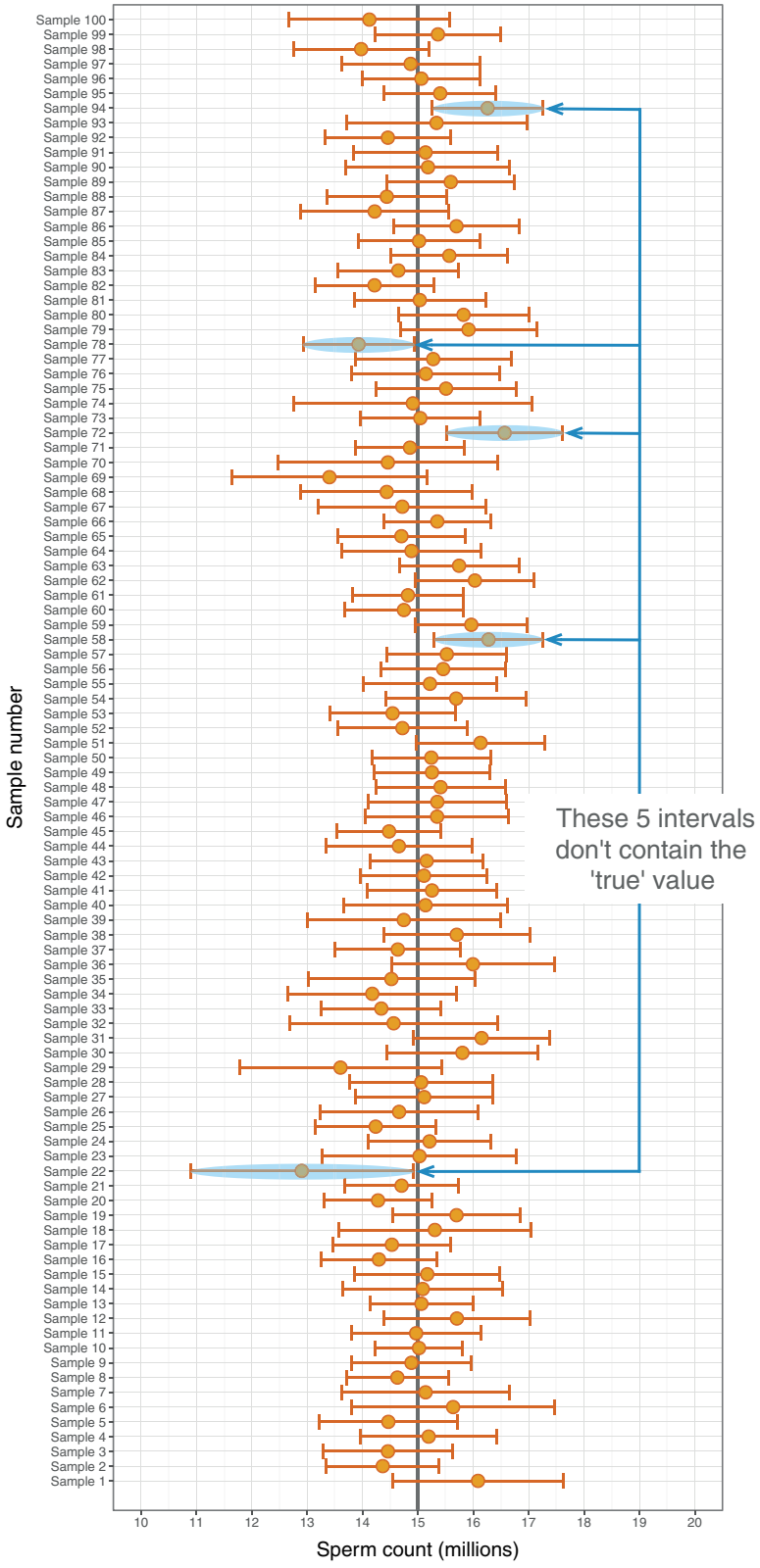


Figure 2.9 The confidence intervals of the sperm counts of Japanese quail (horizontal axis) for 100 different samples (vertical axis)

know whether the confidence interval from a particular sample is one of the 95% that contain the true value or one of the 5% that do not (Misconception Mutt 2.1).

To calculate the confidence interval, we need to know the limits within which 95% of sample means will fall. We know (in large samples) that the sampling distribution of means will be normal, and the normal distribution has been precisely defined such that it has a mean of 0 and a standard deviation of 1. We can use this information to compute the probability of a score occurring, or the limits between which a certain percentage of scores fall (see Section 1.8.6). It was no coincidence that when I explained all of this in Section 1.8.6 I used the example of how we would work out the limits between which 95% of scores fall; that is precisely what we need to know if we want to construct a 95% confidence interval. We discovered in Section 1.8.6 that 95% of z-scores fall between –1.96 and 1.96. This means that if our sample means were normally distributed with a mean of 0 and a standard error of 1, then the limits of our confidence interval would be –1.96 and +1.96. Luckily we know from the central limit theorem that in large samples (above about 30) the sampling distribution *will* be normally distributed (see Section 2.7). It's a pity then that our mean and standard deviation are unlikely to be 0 and 1 – except it's not, because we can convert scores so that they do have a mean of 0 and standard deviation of 1 (z-scores) using equation (1.9):

$$z = \frac{X - \overline{X}}{s}$$

If we know that our limits are –1.96 and 1.96 as z-scores, then to find out the corresponding scores in our raw data we can replace z in the equation (because there are two values, we get two equations):

$$1.96 = \frac{X - \overline{X}}{s} \qquad -1.96 = \frac{X - \overline{X}}{s}$$

We rearrange these equations to discover the value of X:

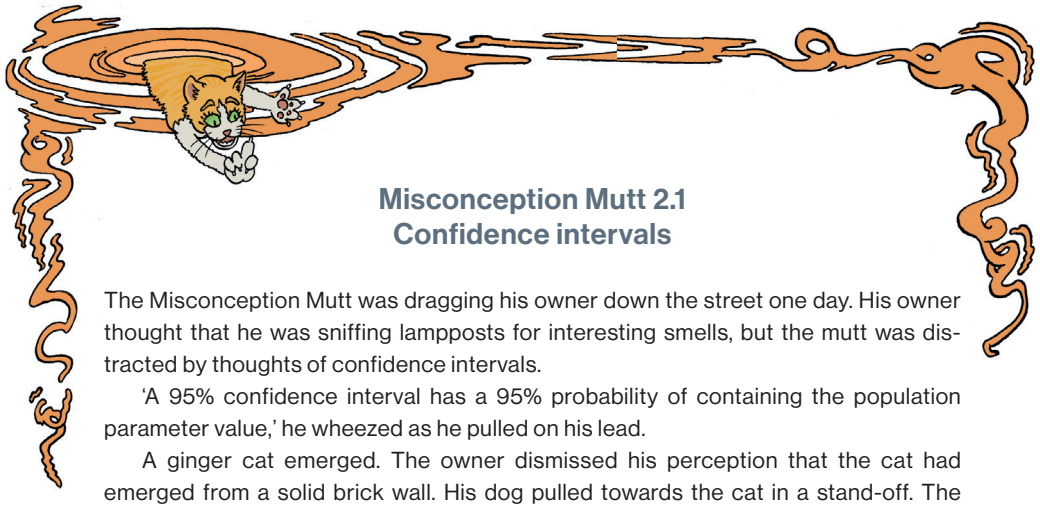
$$1.96 \times s = X - \overline{X} \qquad -1.96 \times s = X - \overline{X}$$

$$(1.96 \times s) + \overline{X} = X \qquad (-1.96 \times s) + \overline{X} = X$$

Therefore, the confidence interval can easily be calculated once the standard deviation (s in the equation) and mean ($\overline{X}$ in the equation) are known. However, we use the standard error and not the standard deviation because we're interested in the variability of *sample* means, not the variability in observations within the sample. The lower boundary of the confidence interval is, therefore, the mean minus 1.96 times the standard error, and the upper boundary is the mean plus 1.96 standard errors:

$$\begin{aligned} \text{lower boundary of confidence interval} &= \overline{X} - (1.96 \times SE) \\ \text{upper boundary of confidence interval} &= \overline{X} + (1.96 \times SE) \end{aligned} \tag{2.15}$$

As such, the mean is always in the centre of the confidence interval. We know that 95% of confidence intervals contain the population mean, so we can assume this confidence interval contains the true mean; therefore, if the interval is small, the sample mean must be very close to the true mean. Conversely, if the confidence interval is very wide then the sample mean could be very different from the true mean, indicating that it is a bad representation of the population. You'll find that confidence intervals will come up time and time again throughout this book.



Misconception Mutt 2.1
Confidence intervals

The Misconception Mutt was dragging his owner down the street one day. His owner thought that he was sniffing lampposts for interesting smells, but the mutt was distracted by thoughts of confidence intervals.

'A 95% confidence interval has a 95% probability of containing the population parameter value,' he wheezed as he pulled on his lead.

A ginger cat emerged. The owner dismissed his perception that the cat had emerged from a solid brick wall. His dog pulled towards the cat in a stand-off. The owner started to check his text messages.

'You again?' the mutt growled.

The cat considered the dog's reins and paced around, smugly displaying his freedom. 'I'm afraid you will see very much more of me if you continue to voice your statistical misconceptions,' he said. 'They call me the Correcting Cat for a reason'.

The dog raised his eyebrows, inviting the feline to elaborate.

'You can't make probability statements about confidence intervals,' the cat announced.

'Huh?' said the mutt.

'You said that a 95% confidence interval has a 95% probability of containing the population parameter. It is a common mistake, but this is not true. The 95% reflects *a long-run* probability.'

'Huh?'

The cat raised his eyes to the sky. 'It means that if you take repeated samples and construct confidence intervals, then 95% of them will contain the population value. That is not the same as a particular confidence interval for a specific sample having a 95% probability of containing the value. In fact, for a specific confidence interval, the probability that it contains the population value is either 0 (it does not contain it) or 1 (it does contain it). You have no way of knowing which it is.' The cat looked pleased with himself.

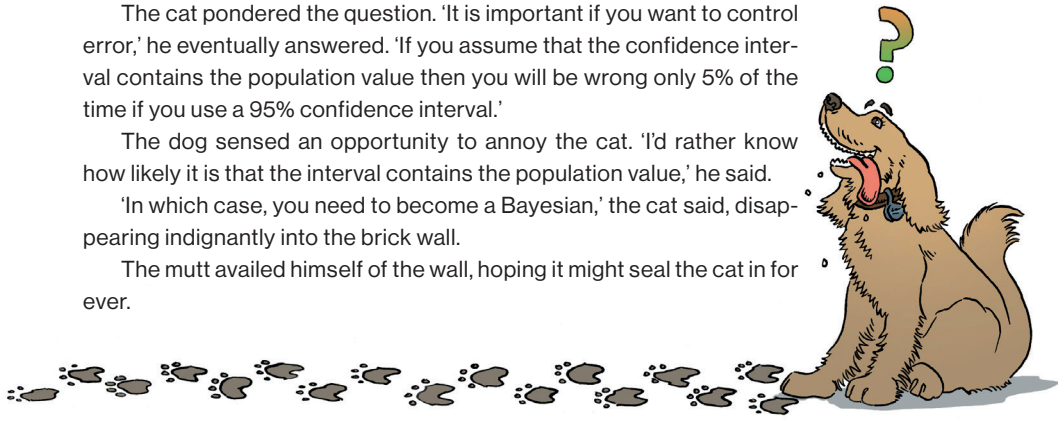
'What's the point of that?' the dog asked.

The cat pondered the question. 'It is important if you want to control error,' he eventually answered. 'If you assume that the confidence interval contains the population value then you will be wrong only 5% of the time if you use a 95% confidence interval.'

The dog sensed an opportunity to annoy the cat. 'I'd rather know how likely it is that the interval contains the population value,' he said.

'In which case, you need to become a Bayesian,' the cat said, disappearing indignantly into the brick wall.

The mutt availed himself of the wall, hoping it might seal the cat in for ever.



2.8.2 Calculating other confidence intervals

The example above shows how to compute a 95% confidence interval (the most common type). However, we sometimes want to calculate other types of confidence interval such as a 99% or 90% interval. The 1.96 and –1.96 in equation (2.15) are the limits within which 95% of z-scores occur. If we wanted to compute confidence intervals for a value other than 95% then we need to look up the value of z for the percentage that we want. For example, we saw in Section 1.8.6 that z-scores of –2.58 and 2.58 are the boundaries that cut off 99% of scores, so we could use these values to compute 99% confidence intervals. In general, we could say that confidence intervals are calculated as:

$$\begin{aligned}\text{lower boundary of confidence interval} &= \bar{X} - \left(z_{\frac{1-p}{2}} \times SE\right)\\ \text{upper boundary of confidence interval} &= \bar{X} + \left(z_{\frac{1-p}{2}} \times SE\right)\end{aligned} \tag{2.16}$$

in which p is the probability value for the confidence interval. So, if you want a 95% confidence interval, then you want the value of z for $(1 - 0.95)/2 = 0.025$. Look this up in the 'smaller portion' column of the table of the standard normal distribution (look back at Figure 1.14) and you'll find that z is 1.96. For a 99% confidence interval we want z for $(1 - 0.99)/2 = 0.005$, which from the table is 2.58 (Figure 1.14). For a 90% confidence interval we want z for $(1 - 0.90)/2 = 0.05$, which from the table is 1.64 (Figure 1.14). These values of z are multiplied by the standard error (as above) to calculate the confidence interval. Using these general principles, we could work out a confidence interval for any level of probability that takes our fancy.

2.8.3 Calculating confidence intervals in small samples

The procedure that I have just described is fine when samples are large, because the central limit theorem tells us that the sampling distribution will be normal. However, for small samples, the sampling distribution is not normal – it has a t-distribution. The t-distribution is a family of probability distributions that change shape as the sample size gets bigger (when the sample is very big, it has the shape of a normal distribution). To construct a confidence interval in a small sample we use the same principle as before, but instead of using the value for z we use the value for t:

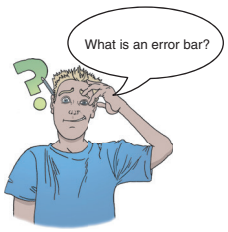
$$\begin{aligned}\text{lower boundary of confidence interval} &= \bar{X} - \left(t_{n-1} \times SE\right)\\ \text{upper boundary of confidence interval} &= \bar{X} + \left(t_{n-1} \times SE\right)\end{aligned} \tag{2.17}$$

The $n - 1$ in the equations is the degrees of freedom (see Jane Superbrain Box 2.2) and tells us which of the t-distributions to use. For a 95% confidence interval, we find the value of t for a two-tailed test with probability of 0.05, for the appropriate degrees of freedom.

In Section 1.8.3 we came across some data about the number of friends that 11 people had on Facebook. We calculated the mean for these data as 95 and standard deviation as 56.79.

- Calculate a 95% confidence interval for this mean.
- Recalculate the confidence interval assuming that the sample size was 56.

2.8.4 Showing confidence intervals visually



Confidence intervals provide us with information about a parameter, and, therefore, you often see them displayed on graphs. (We will discover more about how to create these graphs in Chapter 5.) The confidence interval is usually displayed using something called an error bar, which looks like the letter 'I'. An error bar can represent the standard deviation, or the standard error, but more often than not it shows the 95% confidence interval of the mean. So, often when you see a graph showing the mean, perhaps displayed as a bar or a symbol (Section 5.6), it is accompanied by this funny I-shaped bar.

We have seen that any two samples can have slightly different means (and the standard error tells us a little about how different we can expect sample means to be). We have seen that the 95% confidence interval is an interval constructed such that in 95% of samples the true value of the population mean will fall within its limits. Therefore, the confidence interval tells us the limits within which the population mean is likely to fall. By comparing the confidence intervals of different means (or other parameters) we can get some idea about whether the means came from the same or different populations. (We can't be entirely sure because we don't know whether our particular confidence intervals are ones that contain the population value or not.)

Taking our previous example of quail sperm, imagine we had a sample of quail and the mean sperm release had been 9 million sperm with a confidence interval of 2 to 16. Therefore, if this is one of the 95% of intervals that contains the population value, then the population mean is between 2 and 16 million sperm. What if we now took a second sample of quail and found the confidence interval ranged from 4 to 15? This interval overlaps a lot with our first sample (Figure 2.10). The fact that the confidence intervals overlap in this way tells us that these means could plausibly come from the same population: in both cases, if the intervals contain the true value of the mean (and they are constructed such that in 95% of studies they will), and both intervals overlap considerably, then they contain many similar values. It's very plausible that the population values reflected by these intervals are similar or the same.

What if the confidence interval for our second sample ranged from 18 to 28? If we compared this to our first sample we'd get Figure 2.11. These confidence intervals don't overlap at all, so one confidence interval, which is likely to contain the population mean, tells us that the population mean is somewhere between 2 and 16 million, whereas the other confidence interval, which is also likely to contain the population mean, tells us that the population mean is somewhere between 18 and 28 million. This contradiction suggests two possibilities: (1) our confidence intervals both contain the population mean, but they come from different populations (and, therefore, so do our samples); or (2) both samples come from the same population but one (or both) of the confidence intervals doesn't

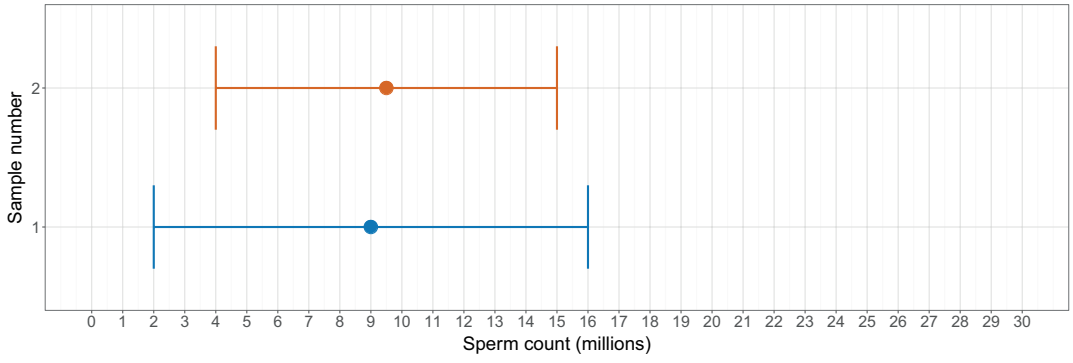


Figure 2.10 Two overlapping 95% confidence intervals

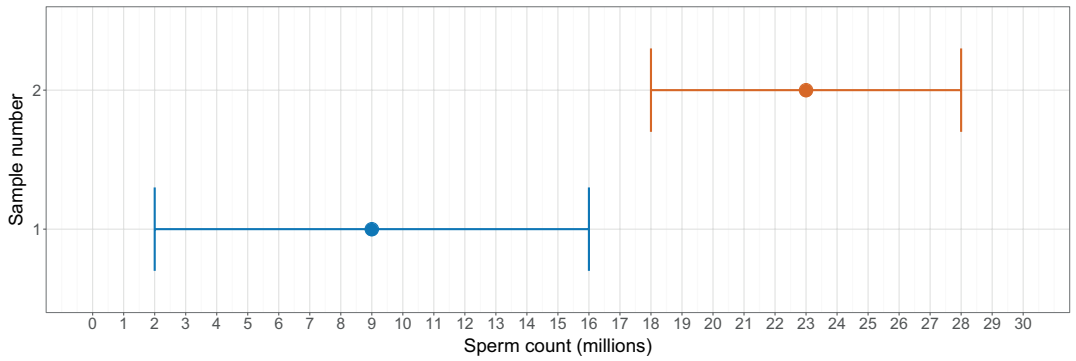


Figure 2.11 Two 95% confidence intervals that don't overlap

contain the population mean (because, as you know, in 5% of cases they don't). If we've used 95% confidence intervals, then we know that the second possibility is unlikely (this happens only 5 times in 100 or 5% of the time), so the first explanation is more plausible.

I can hear you all thinking, 'So what if the samples come from a different population?' Well, it has a very important implication in experimental research. When we do an experiment, we introduce some form of manipulation between two or more conditions (see Section 1.7.2). If we have taken two random samples of people, and we have tested them on some measure, then we expect these people to belong to the same population. We'd also, therefore, expect their confidence intervals to reflect the same population value for the mean. If their sample means and confidence intervals are so different as to suggest that they come from different populations, then this is likely to be because our experimental manipulation has induced a difference between the samples. Therefore, error bars showing 95% confidence intervals are useful, because if the bars of any two means do not overlap (or overlap by only a small amount) then we can infer that these means are from different populations – they are significantly different. We will return to this point in Section 2.9.9.

Cramming Sam's Tips
Confidence intervals

- A confidence interval for the mean is a range of scores constructed such that the population mean will fall within this range in 95% of samples.
- The confidence interval is *not* an interval within which we are 95% confident that the population mean will fall.

2.9 N is for null hypothesis significance testing

In Chapter 1 we saw that research was a six-stage process (Figure 1.2). This chapter has looked at the final stage:

- Analyse the data: fit a statistical model to the data – this model will test your original predictions. Assess this model to see whether it supports your initial predictions.

I have shown that we can use a sample of data to estimate what's happening in a larger population to which we don't have access. We have also seen (using the mean as an example) that we can fit a statistical model to a sample of data and assess how well it fits. However, we have yet to see how fitting models like these can help us to test our research predictions. How do statistical models help us to test complex hypotheses such as 'Is there a relationship between the amount of gibberish that people speak and the amount of vodka jelly they've eaten?' or 'Does reading this chapter improve your knowledge of research methods?' This brings us to the 'N' in the SPINE of statistics: null hypothesis significance testing.

Null hypothesis significance testing (NHST) is a cumbersome name for an equally cumbersome process. NHST is the most commonly taught approach to testing research questions with statistical models. It arose out of two different approaches to the problem of how to use data to test theories: (1) Ronald Fisher's idea of computing probabilities to evaluate evidence, and (2) Jerzy Neyman and Egon Pearson's idea of competing hypotheses.

2.9.1 Fisher's *p*-value

Fisher (1925/1991) (Figure 2.12) described an experiment designed to test a claim by a woman that she could determine, by tasting a cup of tea, whether the milk or the tea was added first to the cup. Fisher thought that he should give the woman some cups of tea, some of which had the milk added first and some of which had the milk added last, and see whether she could correctly identify them. The woman would know that there are an equal number of cups in which milk was added first or last, but wouldn't know in which order the cups were placed. If we take the simplest situation in which there are only two cups, then the woman has a 50% chance of guessing correctly. If she did guess correctly, we wouldn't be that confident in concluding that she can tell the difference between the cups in which the milk was added first and those in which it was added last, because even by guessing she would be correct half of the time. But what if we complicated things by having six cups? There are 20 orders in which these cups can be arranged and the woman would guess the correct order only 1 time in 20 (or 5% of the time). If she got the order correct we would be much more confident that she could genuinely tell the difference (and bow down in awe of her finely tuned palate). If you'd like to know more about Fisher and his tea-tasting antics, see David Salsburg's excellent book *The Lady Tasting Tea* (Salsburg, 2002). For our purposes the take-home point is that only when there was a very small probability that the woman could complete the tea task by guessing alone would we conclude that she had a genuine skill in detecting whether milk was poured into a cup before or after the tea.

It's no coincidence that I chose the example of six cups above (where the tea-taster had a 5% chance of getting the task right by guessing), because scientists tend to use 5% as a threshold for confidence: only when there is a 5% chance (or 0.05 probability) of getting the result we have (or one more extreme) if no

effect exists are we confident enough to accept that the effect is genuine.[10] Fisher's basic point was that you should calculate the probability of an event and evaluate this probability within the research context. Although Fisher felt a $p = 0.01$ would be strong evidence to back up a hypothesis, and perhaps a $p = 0.20$ would be weak evidence, he never said $p = 0.05$ was in any way a magic number. Fast forward 100 years or so, and everyone treats 0.05 as though it *is* a magic number.

Figure 2.12 Sir Ronald A. Fisher, the cleverest person ever ($p < 0.0001$)

2.9.2 Types of hypothesis

In contrast to Fisher, Neyman and Pearson believed that scientific statements should be split into testable hypotheses. The hypothesis or prediction from your theory would normally be that an effect will be present. This hypothesis is called the **alternative hypothesis** and is denoted by H_1. (It is sometimes also called the **experimental hypothesis**, but because this term relates to a specific type of methodology it's probably best to use 'alternative hypothesis'.) There is another type of hypothesis called the **null hypothesis**, which is denoted by H_0. This hypothesis is the opposite of the alternative hypothesis and so usually states that an effect is absent.

Often when I write, my thoughts are drawn towards chocolate. I believe that I would eat less of it if I could stop thinking about it. However, according to Morewedge, Huh, & Vosgerau (2010), that's not true. In fact, they found that people ate less of a food if they had previously imagined eating it. Imagine we did a similar study. We might generate the following hypotheses:

- Alternative hypothesis: if you imagine eating chocolate you will eat less of it.
- Null hypothesis: if you imagine eating chocolate you will eat the same amount as normal.

The null hypothesis is useful because it gives us a baseline against which to evaluate how plausible our alternative hypothesis is. We can evaluate whether we think that the data we have collected are more likely, given the null or alternative hypothesis. A lot of books talk about accepting or rejecting these hypotheses, implying that you look at the data and either accept the null hypothesis (and therefore reject the alternative) or accept the alternative hypothesis (and reject the null). In fact, this isn't quite right because the way that scientists typically evaluate these hypotheses using p-values (which we'll come onto shortly) doesn't provide evidence for such black-and-white decisions. So, rather than talking about accepting or rejecting a hypothesis, we should talk about 'the chances of obtaining the result we have (or one more extreme), assuming that the null hypothesis is true'.

Imagine in our study that we took 100 people and measured how many pieces of chocolate they usually eat (day 1). On day 2, we got them to imagine eating chocolate and again measured how much chocolate they ate that day. Imagine that we found that 75% of people ate less chocolate on the second day than on the first. When we analyse our data, we are really asking, 'Assuming that imagining eating chocolate has no effect whatsoever, is it likely that 75% of people would eat less chocolate on the second day?' Intuitively, the answer is that the chances are very low: if the null hypothesis is true, then everyone should eat the same amount of chocolate on both days. Therefore, we are very unlikely to have got the data that we did if the null hypothesis were true.

10 Of course it might not be true – we're just prepared to believe that it is.

What if we found that only 1 person (1%) ate less chocolate on the second day? If the null hypothesis is true and imagining eating chocolate has no effect whatsoever on consumption, then no people should eat less on the second day. The chances of getting these data if the null hypothesis is true are quite high. The null hypothesis is quite plausible given what we have observed.

When we collect data to test theories we work in these terms: we cannot talk about the null hypothesis being true or the experimental hypothesis being true, we can talk only in terms of the probability of obtaining a particular result or statistic if, hypothetically speaking, the null hypothesis were true. It's also worth remembering that our alternative hypothesis is likely to be one of many possible models that we could fit to the data, so even if we believe it to be more likely than the null hypothesis, there may be other models of the data that we haven't considered that are a better fit, which again means that we cannot talk about the hypothesis as being definitively true or false, but we can talk about its plausibility relative to other hypotheses or models that we have considered.

Hypotheses can be directional or non-directional. A directional hypothesis states that an effect will occur, but it also states the direction of the effect. For example, 'If you imagine eating chocolate you will eat less of it' is a one-tailed hypothesis because it states the direction of the effect (people will eat less). A non-directional hypothesis states that an effect will occur, but it doesn't state the direction of the effect. For example, 'Imagining eating chocolate affects the amount of chocolate you eat' does not tell us whether people will eat more or less.

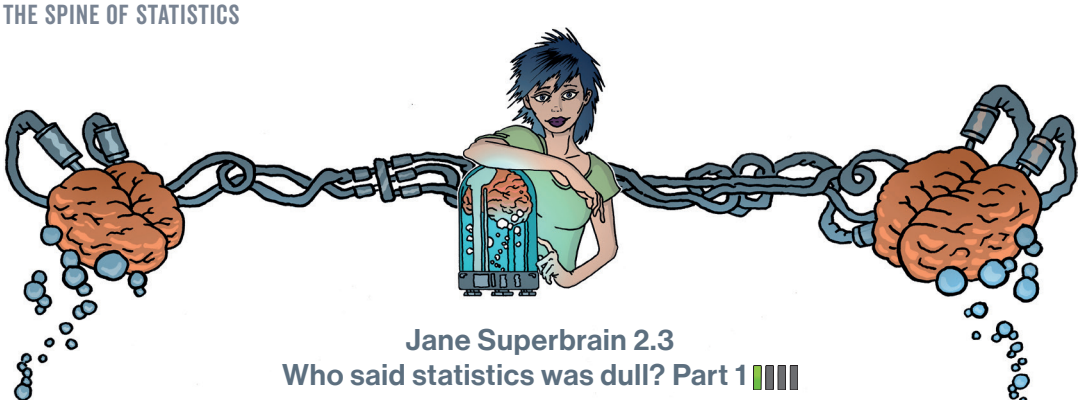
What are the null and alternative hypotheses for the following questions:

- 'Is there a relationship between the amount of gibberish that people speak and the amount of vodka jelly they've eaten?'
- 'Does reading this chapter improve your knowledge of research methods?'

2.9.3 The process of NHST

NHST is a blend of Fisher's idea of using the probability value p as an index of the weight of evidence against a null hypothesis, and Jerzy Neyman and Egon Pearson's idea of testing a null hypothesis *against* an alternative hypothesis (Neyman & Pearson, 1933). There was no love lost between these competing statisticians (Jane Superbrain Box 2.3). NHST is a system designed to tell us whether the alternative hypothesis is likely to be true – it helps us to decide whether to confirm or reject our predictions.

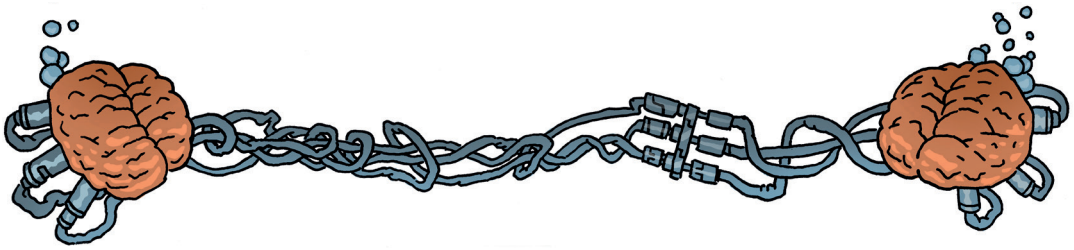
Figure 2.13 outlines the steps in NHST. As we have seen before, the process starts with a research hypothesis that generates a testable prediction. These predictions are decomposed into a null (there is no effect) and alternative hypothesis (there is an effect). At this point you decide upon the long-run error rate that you are prepared to accept, alpha (α). In other words, how often are you prepared to be wrong? This is the significance level, the probability of accepting an effect in our population as true, when no such effect exists (it is known as the Type I error rate, which we'll discuss in more detail in due course). It is important that we fix this error rate before we collect data, otherwise we are cheating (see Jane Superbrain Box 2.4). You should determine your error rate based on the nuances of your research area, and what it is you're trying to test. Put another way, it should be a meaningful decision. In reality, it is not: everyone uses 0.05 (a 5% error rate) with barely a thought for what it means or why they're using it. Go figure.



Jane Superbrain 2.3
Who said statistics was dull? Part 1

Students often think that statistics is dull, but back in the early 1900s it was anything but dull, with prominent figures entering into feuds on a regular basis. Ronald Fisher and Jerzy Neyman had a particularly impressive feud. On 28 March 1935 Neyman delivered a talk to the Royal Statistical Society, at which Fisher was present, in which he criticized some of Fisher's most important work. Fisher directly attacked Neyman in his discussion of the paper at the same meeting: he more or less said that Neyman didn't know what he was talking about and didn't understand the background material on which his work was based. He said, 'I put it to you, sir, that you are a fool, an imbecile, a man so incapacitated by stupidity that in a battle of wits with a single-cell amoeba, the amoeba would fancy its chances.' He didn't really say that, but he opened the discussion without proposing a vote of thanks, which would have been equally as rude in those days.

Relations soured so much that while they both worked at University College London, Neyman openly attacked many of Fisher's ideas in lectures to his students. The two feuding groups even took afternoon tea (a common practice in the British academic community of the time and one which, frankly, we should reinstate) in the same room but at different times. The truth behind who fuelled these feuds is, perhaps, lost in the mists of time, but Zabell (1992) makes a sterling effort to unearth it. Basically, the founders of modern statistical methods, despite being super-humanly intelligent,[11] acted like a bunch of squabbling children.



Having not given the thought you should have to your error rate, you choose a sampling distribution. This involves working out what statistical model to fit to the data that will test your hypothesis, looking at what parameters that model has, and then deciding on the shape of the sampling distribution attached to those parameters. Let's take my example of whether thinking about chocolate is related to consumption. You could measure how much people think about chocolate during the day and how much of it they eat in the same day. If the null hypothesis is true (there is no effect), then there should be no relationship at all between these variables. If it reduces consumption, then we'd expect a negative relationship between the two. One model we could fit that tests this hypothesis is the linear model that I described earlier, in which we predict consumption (the outcome) from thought about chocolate (the predictor). Our model is basically equation (2.3), but I'll replace the outcome and letter X with our variable names:

$$\text{consumption}_i = (b_0 + b\,\text{thought}_i) + \text{error}_i \tag{2.18}$$

11 Fisher, in particular, was a world leader in genetics, biology and medicine as well as possibly the most original mathematical thinker ever (Barnard, 1963; Field, 2005d; Savage, 1976).

The parameter, b, attached to the variable **thought** tests our hypothesis: it quantifies the size and strength of relationship between thinking and consuming. If the null is true, b will be zero; otherwise it will be a value different from 0, the size and direction of which depends on what the relationship between the thought and consumption variables is. It turns out (see Chapter 9) that this parameter has a sampling distribution that has a t-distribution. So, that's what we'd use to test our hypothesis. We also need to establish how much data to collect to stand a reasonable chance of finding the effect we're looking for. This is called the power of the test, and I'll elaborate on this concept shortly.

Now the fun begins and you collect your data. You fit the statistical model that tests your hypothesis to the data. In the chocolate example, we'd estimate the parameter that represents the relationship between thought and consumption and its confidence interval. It's usually also possible to compute a test statistic that maps the parameter to a long-run probability value (the p-value). In our chocolate example, we can compute a statistic known as t, which has a specific sampling distribution from which we can get a probability value (p). This probability value tells us how likely it would be to get a value of t at least as big as the one we have if the null hypothesis is true. As I keep mentioning, this p is a long-run probability: it is computed by working out how often you'd get specific values of the test statistic (in this case t) if you repeated your exact sampling process an infinite number of times.

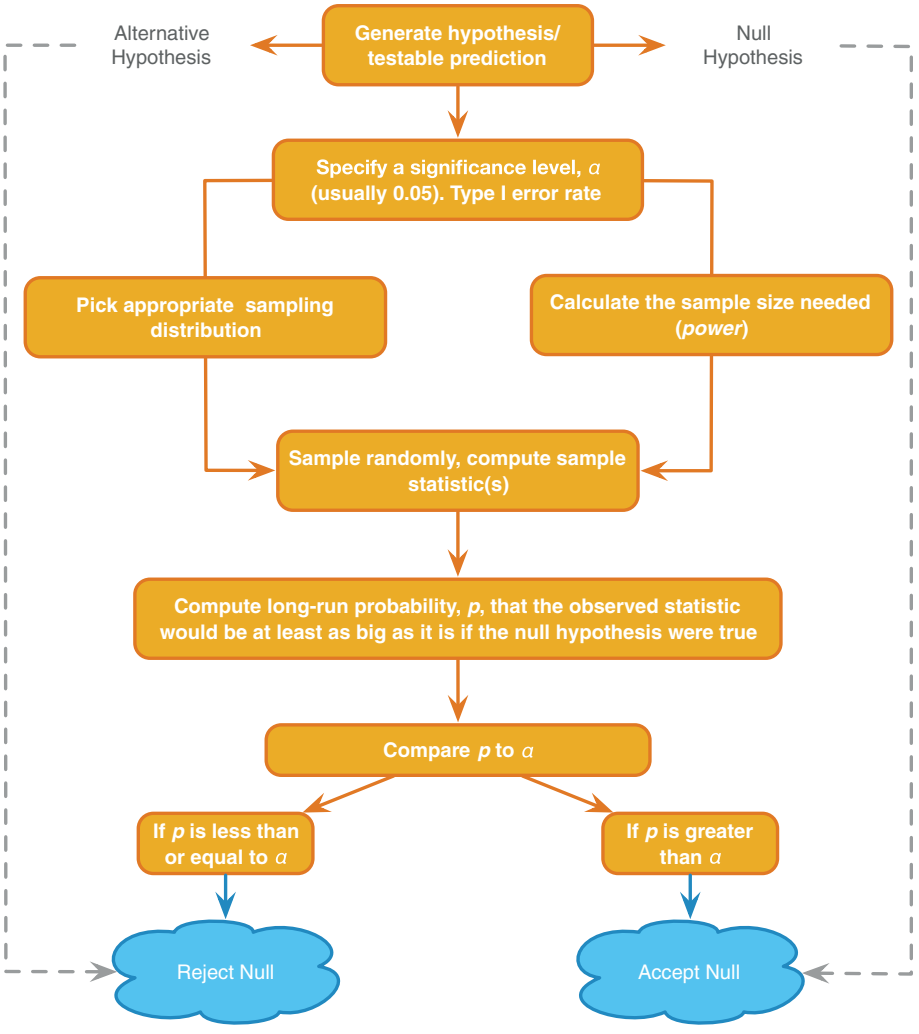
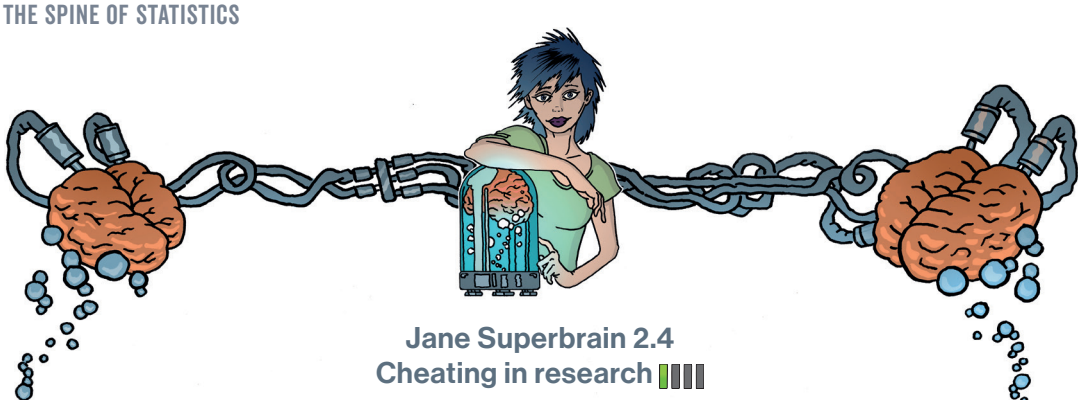


Figure 2.13 Flow chart of null hypothesis significance testing



Jane Superbrain 2.4
Cheating in research

NHST works only if you generate your hypotheses and decide on your criteria for whether an effect is significant before collecting the data. Imagine I wanted to place a bet on who would win the soccer World Cup. Being English, I might bet on England to win the tournament. To do this, I'd: (1) place my bet, choosing my team (England) and odds available at the betting shop (e.g., 6/4); (2) see which team wins the tournament; (3) collect my winnings (or more likely not).

To keep everyone happy, this process needs to be equitable: the betting shops set their odds such that they're not paying out too much money (which keeps them happy), but so that they do pay out sometimes (to keep the customers happy). The betting shop can offer any odds before the tournament has ended, but it can't change them once the tournament is over (or the last game has started). Similarly, I can choose any team before the tournament, but I can't then change my mind halfway through, or after the final game.

The research process is similar: we can choose any hypothesis (soccer team) before the data are collected, but we can't change our minds halfway through data collection (or after data collection). Likewise we have to decide on our probability level (or betting odds) before we collect data. *If* we do this, the process works. However, researchers sometimes cheat. They don't formulate hypotheses before they conduct their experiments, they change them when the data are collected (like me changing my team after the World Cup is over), or, worse still, they decide on them after the data are collected (see Chapter 3). With the exception of procedures such as *post hoc* tests, this is cheating. Similarly, researchers can be guilty of choosing which significance level to use after the data are collected and analysed, like a betting shop changing the odds after the tournament.

If you change your hypothesis or the details of your analysis, you increase the chances of finding a significant result, but you also make it more likely that you will publish results that other researchers can't reproduce (which is embarrassing). If, however, you follow the rules carefully and do your significance testing at the 5% level you at least know that in the long run at most only 1 result out of every 20 will risk this public humiliation. (Thanks to David Hitchin for this box, and apologies to him for introducing soccer into it.)

It is important that you collect the amount of data that you set out to collect, otherwise the p-value you obtain will not be correct. It is possible to compute a p-value representing the long-run probability of getting a t-value at least as big as the one you have in repeated samples of, say, 80, but there is no way of knowing the probability of getting a t-value at least as big as the one you have in repeated samples of 74, where the intention was to collect 80 but term ended and you couldn't find any more participants. If you cut data collection short (or extend it) for this sort of arbitrary reason, then whatever p-value you end up with is certainly not the one you want. Again, it's cheating: you're changing your team after you have placed your bet, and you will likely end up with research egg on your face when no one can replicate your findings.

Having hopefully stuck to your original sampling frame and obtained the appropriate p-value, you compare it to your original alpha value (usually 0.05). If the p you obtain is less than or equal to the original α, scientists typically use this as grounds to reject the null hypothesis outright; if the p is greater than α, then they accept that the null hypothesis is plausibly true (they reject the alternative hypothesis). We can never be completely sure that either hypothesis is correct; all we can do is to calculate the probability that our model would fit at least as well as it does if there were no effect in the population (i.e., the null hypothesis is true). As this probability decreases, we gain greater confidence that the alternative hypothesis is more plausible than the null hypothesis. This overview of NHST is a lot to take in, so we will revisit a lot of the key concepts in detail the next few sections.

2.9.4 Test statistics

I mentioned that NHST relies on fitting a model to the data and then evaluating the probability of this model, given the assumption that no effect exists. I mentioned in passing that the fit of a model, or the parameters within it, are typically mapped to a probability value through a test statistic. I was deliberately vague about what the 'test statistic' is, so let's lift the veil of secrecy. To do this we need to return to the concepts of systematic and unsystematic variation that we encountered in Section 1.7.4. Systematic variation is variation that can be explained by the model that we've fitted to the data (and, therefore, due to the hypothesis that we're testing). Unsystematic variation is variation that cannot be explained by the model that we've fitted. In other words, it is error, or variation not attributable to the effect we're investigating. The simplest way, therefore, to test whether the model fits the data, or whether our hypothesis is a good explanation of the data we have observed, is to compare the systematic variation against the unsystematic variation. In effect, we look at a signal-to-noise ratio: we compare how good the model/hypothesis is against how bad it is (the error):

$$\text{Test statistic} = \frac{\text{signal}}{\text{noise}} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} = \frac{\text{effect}}{\text{error}} \tag{2.19}$$

Likewise, the best way to test a parameter is to look at the size of the parameter relative to the background noise (the sampling variation) that produced it. Again, it's a signal-to-noise ratio: the ratio of how big a parameter is to how much it can vary across samples:

$$\text{Test statistic} = \frac{\text{signal}}{\text{noise}} = \frac{\text{size of parameter}}{\text{sampling variation in the parameter}} = \frac{\text{effect}}{\text{error}} \tag{2.20}$$

The ratio of effect relative to error is a **test statistic**, and you'll discover later in the book that there are lots of them: t, χ^2, and F, to name only three. The exact form of the equation changes depending on which test statistic you're calculating, but the important thing to remember is that they all, crudely speaking, represent the same thing: signal-to-noise or the amount of variance explained by the model we've fitted to the data compared to the variance that can't be explained by the model (see Chapters 9 and 10, in particular, for a more detailed explanation). The reason why this ratio is so useful is intuitive, really: if our model is good then we'd expect it to be able to explain more variance than it can't explain. In this case, the test statistic will be greater than 1 (but not necessarily significant). Similarly, larger parameters (bigger effects) that are likely to represent the population (smaller sampling variation) will produce larger test statistics.

A test statistic is a statistic for which we know how frequently different values occur. I mentioned the t-distribution, chi-square (χ^2) distribution and F-distribution in Section 1.8.6 and said that they are all defined by an equation that enables us to calculate precisely the probability of obtaining a given score. Therefore, if a test statistic comes from one of these distributions we can calculate the probability

of obtaining a certain value (just as we could estimate the probability of getting a score of a certain size from a frequency distribution in Section 1.8.6). This probability is the p-value that Fisher described, and in NHST it is used to estimate how likely (in the long run) it is that we would get a test statistic at least as big as the one we have *if there were no effect* (i.e., the null hypothesis were true).

Test statistics can be a bit scary, so let's imagine that they're cute kittens. Kittens are typically very small (about 100g at birth, on average), but every so often a cat will give birth to a big one (say, 150g). A 150g kitten is rare, so the probability of finding one is very small. Conversely, 100g kittens are very common, so the probability of finding one is quite high. Test statistics are the same as kittens in this respect: small ones are quite common and large ones are rare. So, if we do some research (i.e., give birth to a kitten) and calculate a test statistic (weigh the kitten), we can calculate the probability of obtaining a value/weight at least that large. The more variation our model explains compared to the variance it can't explain, the bigger the test statistic will be (i.e., the more the kitten weighs), and the more unlikely it is to occur by chance (like our 150g kitten). Like kittens, as test statistics get bigger, the probability of them occurring becomes smaller. If this probability falls below a certain value ($p < 0.05$ if we blindly apply the conventional 5% error rate), we presume that the test statistic is as large as it is because our model explains a sufficient amount of variation to reflect a genuine effect in the real world (the population). The test statistic is said to be *statistically significant*. Given that the statistical model that we fit to the data reflects the hypothesis that we set out to test, then a significant test statistic tells us that the model would be unlikely to fit this well if the there was no effect in the population (i.e., the null hypothesis was true). Typically, this is taken as a reason to reject the null hypothesis and gain confidence that the alternative hypothesis is true. If, however, the probability of obtaining a test statistic at least as big as the one we have (if the null hypothesis were true) is too large (typically $p > 0.05$), then the test statistic is said to be non-significant and is used as grounds to reject the alternative hypothesis (see Section 3.2.1 for a discussion of what 'statistically significant' means).

2.9.5 One- and two-tailed tests

We saw in Section 1.9.2 that hypotheses can be directional (e.g., 'The more someone reads this book, the more they want to kill its author') or non-directional (i.e., 'Reading more of this book could increase or decrease the reader's desire to kill its author'). A statistical model that tests a directional hypothesis is called a **one-tailed test**, whereas one testing a non-directional hypothesis is known as a **two-tailed test**.

Imagine we wanted to discover whether reading this book increased or decreased the desire to kill me. If we have no directional hypothesis then there are three possibilities. (1) People who read this book want to kill me more than those who don't, so the difference (the mean for those reading the book minus the mean for non-readers) is positive. Put another way, as the amount of time spent reading this book increases, so does the desire to kill me – a positive relationship. (2) People who read this book want to kill me less than those who don't, so the difference (the mean for those reading the book minus the mean for non-readers) is negative. Alternatively, as the amount of time spent reading this book increases, the desire to kill me decreases – a negative relationship. (3) There is no difference between readers and non-readers in their desire to kill me – the mean for readers minus the mean for non-readers is exactly zero. There is no relationship between reading this book and wanting to kill me. This final option is the null hypothesis. The direction of the test statistic (i.e., whether it is positive or negative) depends on whether the difference, or direction of relationship, is positive or negative. Assuming that there is a positive difference or relationship (the more you read, the more you want to kill me), then to detect this difference we take account of the fact that the mean for readers is bigger than for non-readers

(and so derive a positive test statistic). However, if we've predicted incorrectly and reading this book makes readers want to kill me *less*, then the test statistic will be negative instead.

What are the consequences of this? Well, if at the 0.05 level we needed to get a test statistic bigger than, say, 10 and the one we got was actually –12, then we would reject the hypothesis even though a difference does exist. To avoid this, we can look at both ends (or tails) of the distribution of possible test statistics. This means we will catch both positive and negative test statistics. However, doing this has a price because, to keep our criterion probability of 0.05, we split this probability across the two tails: we have 0.025 at the positive end of the distribution and 0.025 at the negative end. Figure 2.14 shows this situation – the orange tinted areas are the areas above the test statistic needed at a 0.025 level of significance. Combine the probabilities (i.e., add the two tinted areas together) at both ends and we get 0.05, our criterion value.

If we have made a prediction, then we put all our eggs in one basket and look only at one end of the distribution (either the positive or the negative end, depending on the direction of the prediction we make). In Figure 2.14, rather than having two small orange tinted areas at either end of the distribution that show the significant values, we have a bigger area (the blue tinted area) at only one end of the distribution that shows significant values. Note that this blue area contains within it one of the orange areas as well as an extra bit of blue area. Consequently, we can just look for the value of the test statistic that would occur if the null hypothesis were true with a probability of 0.05. In Figure 2.14,

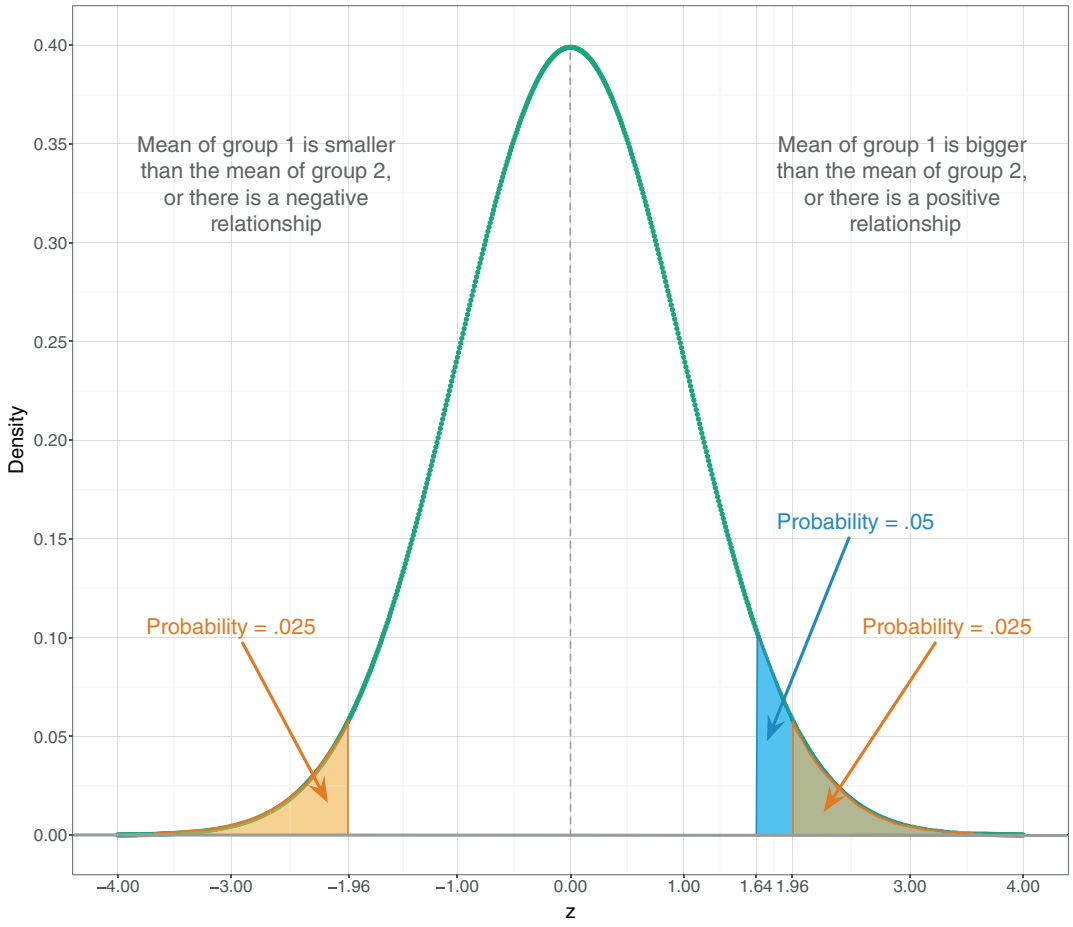


Figure 2.14 Diagram to show the difference between one- and two-tailed tests

the blue tinted area is the area above the positive test statistic needed at a 0.05 level of significance (1.64); this value is smaller than the value that begins the area for the 0.025 level of significance (1.96). This means that if we make a specific prediction then we need a smaller test statistic to find a significant result (because we are looking in only one tail of the distribution), but if our prediction happens to be in the wrong direction then we won't detect the effect that does exist. This final point is very important, so let me rephrase it: if you do a one-tailed test and the results turn out to be in the opposite direction to what you predicted, you must ignore them, resist all temptation to interpret them, and accept (no matter how much it pains you) the null hypothesis. If you *don't* do this, then you have done a two-tailed test using a different level of significance from the one you set out to use (and Jane Superbrain Box 2.4 explains why that is a bad idea).

I have explained one- and two-tailed tests because people expect to find them explained in statistics textbooks. However, there are a few reasons why you should think long and hard about whether one-tailed tests are a good idea. Wainer (1972) quotes John Tukey (one of the great modern statisticians) as responding to the question 'Do you mean to say that one should *never* do a one-tailed test?' by saying 'Not at all. It depends upon to whom you are speaking. *Some people will believe anything*' (emphasis added). Why might Tukey have been so sceptical?

As I have said already, if the result of a one-tailed test is in the opposite direction to what you expected, *you cannot and must not reject the null hypothesis*. In other words, you must completely ignore that result even though it is poking you in the arm and saying 'look at me, I'm intriguing and unexpected'. The reality is that when scientists see interesting and unexpected findings their instinct is to want to explain them. Therefore, one-tailed tests are like a mermaid luring a lonely sailor to his death by being beguiling and interesting: they lure lonely scientists to their academic death by throwing up irresistible and unpredicted results.

One context in which a one-tailed test *could* be used, then, is if a result in the opposite direction to that expected would result in the same action as a non-significant result (Lombardi & Hurlbert, 2009; Ruxton & Neuhaeuser, 2010). There are some limited circumstances in which this might be the case. First, if a result in the opposite direction would be theoretically meaningless or impossible to explain even if you wanted to (Kimmel, 1957). Second, imagine you're testing a new drug to treat depression. You predict it will be better than existing drugs. If it is not better than existing drugs (non-significant p) you would not approve the drug; however, if it was significantly worse than existing drugs (significant p but in the opposite direction) you would also not approve the drug. In both situations, the drug is not approved.

Finally, one-tailed tests encourage cheating. If you do a two-tailed test and find that your p is 0.06, then you would conclude that your results were not significant (because 0.06 is bigger than the critical value of 0.05). Had you done this test one-tailed, however, the p you would get would be half of the two-tailed value (0.03). This one-tailed value would be significant at the conventional level (because 0.03 is less than 0.05). Therefore, if we find a two-tailed p that is just non-significant, we might be tempted to pretend that we'd always intended to do a one-tailed test because our 'one-tailed' p-value is significant. But we can't change our rules after we have collected data (Jane Superbrain Box 2.4), so we must conclude that the effect is not significant. Although scientists hopefully don't do this sort of thing deliberately, people do get confused about what is and isn't permissible. Two recent surveys of practice in ecology journals concluded that 'all uses of one-tailed tests in the journals surveyed seemed invalid' (Lombardi & Hurlbert, 2009) and that only 1 in 17 papers using one-tailed tests were justified in doing so (Ruxton & Neuhaeuser, 2010).

One way around the temptation to cheat is to pre-register your study (which we'll discuss in detail in the following chapter). At a simple level pre-registration means that you commit publicly to your analysis strategy before collecting data. This could be a simple statement on your own website, or, as

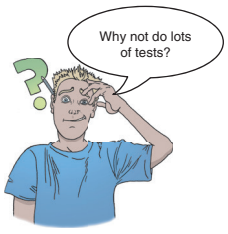
we shall see, a formally submitted article outlining your research intentions. One benefit of pre-registering your research is that it then becomes transparent if you change your analysis plan (e.g., by switching from a two-tailed to a one-tailed test). It is much less tempting to halve your *p*-value to take it below 0.05 if the world will know you have done it!

2.9.6 Type I and Type II errors

Neyman and Pearson identified two types of errors that we can make when we test hypotheses. When we use test statistics to tell us about the true state of the world, we're trying to see whether there is an effect in our population. There are two possibilities: there is, in reality, an effect in the population, or there is, in reality, no effect in the population. We have no way of knowing which of these possibilities is true; however, we can look at test statistics and their associated probability to help us to decide which of the two is more likely. It is important that we're as accurate as possible. There are two mistakes we can make: a Type I and a Type II error. A **Type I error** occurs when we believe that there is a genuine effect in our population, when in fact there isn't. If we use the conventional criterion for alpha then the probability of this error is 0.05 (or 5%) when there is no effect in the population – this value is the **α-level** that we encountered in Figure 2.13. Assuming that there is no effect in our population, if we replicated our data collection 100 times, we could expect that on five occasions we would obtain a test statistic large enough to make us think that there was a genuine effect in the population even though there isn't. The opposite is a **Type II error**, which occurs when we believe that there is no effect in the population when, in reality, there is. This would occur when we obtain a small test statistic (perhaps because there is a lot of natural variation between our samples). In an ideal world, we want the probability of this error to be very small (if there is an effect in the population then it's important that we can detect it). Cohen (1992) suggests that the maximum acceptable probability of a Type II error would be 0.2 (or 20%) – this is called the ***β*-level**. That would mean that if we took 100 samples of data from a population in which an effect exists, we would fail to detect that effect in 20 of those samples (so we'd miss 1 in 5 genuine effects).

There is a trade-off between these two errors: if we lower the probability of accepting an effect as genuine (i.e., make α smaller) then we increase the probability that we'll reject an effect that does genuinely exist (because we've been so strict about the level at which we'll accept that an effect is genuine). The exact relationship between the Type I and Type II error is not straightforward because they are based on different assumptions: to make a Type I error there must be no effect in the population, whereas to make a Type II error the opposite is true (there must be an effect that we've missed). So, although we know that as the probability of making a Type I error decreases, the probability of making a Type II error increases, the exact nature of the relationship is usually left for the researcher to make an educated guess (Howell, 2012, gives a great explanation of the trade-off between errors).

2.9.7 Inflated error rates



As we have seen, if a test uses a 0.05 level of significance then the chances of making a Type I error are only 5%. Logically, then, the probability of no Type I errors is 0.95 (95%) for each test. However, in science it's rarely the case that we can get a definitive answer to our research question using a single test on our data: we often need to conduct several tests. For example, imagine we wanted to look at factors that affect how viral a video becomes on YouTube. You might predict that the amount of humour and innovation in the video will be important factors. To test this, you might look at the relationship between the number of hits and measures of both the humour content and the innovation. However, you probably ought to also look at whether innovation and humour

content are related too. Therefore, you would need to do three tests. If we assume that each test is independent (which in this case they won't be, but it enables us to multiply the probabilities), then the overall probability of no Type I errors will be $0.95^3 = 0.95 \times 0.95 \times 0.95 = 0.857$, because the probability of no Type I errors is 0.95 for each test and there are three tests. Given that the probability of no Type I errors is 0.857, then the probability of making at least one Type I error is this number subtracted from 1 (remember that the maximum probability of any event occurring is 1). So, the probability of at least one Type I error is $1 - 0.857 = 0.143$, or 14.3%. Therefore, across this group of tests, the probability of making a Type I error has increased from 5% to 14.3%, a value greater than the criterion that is typically used. This error rate across statistical tests conducted on the same data is known as the **familywise** or **experimentwise error rate**. Our scenario with three tests is relatively simple, and the effect of carrying out several tests is not too severe, but imagine that we increased the number of tests from three to ten. The familywise error rate can be calculated using equation (2.21) (assuming you use a 0.05 level of significance):

$$\text{familywise error} = 1 - 0.95^n \tag{2.21}$$

In this equation n is the number of tests carried out on the data. With ten tests carried out, the familywise error rate is $1 - 0.95^{10} = 0.40$, which means that there is a 40% chance of having made at least one Type I error.

To combat this build-up of errors, we can adjust the level of significance for individual tests such that the overall Type I error rate (α) across all comparisons remains at 0.05. There are several ways in which the familywise error rate can be controlled. The most popular (and easiest) way is to divide α by the number of comparisons, k, as in equation (2.22):

$$P_{\text{Crit}} = \frac{\alpha}{k} \tag{2.22}$$

Therefore, if we conduct 10 tests, we use 0.005 as our criterion for significance. In doing so, we ensure that the cumulative Type I error remains below 0.05. This method is known as the **Bonferroni correction**, because it uses an inequality described by Carlo Bonferroni, but despite the name its modern application to confidence intervals can be attributed to Olive Dunn (Figure 2.15). There is a trade-off for controlling the familywise error rate and that is a loss of statistical power, which is the next topic on our agenda.

Carlo Bonferroni

Olive Dunn

Figure 2.15 The king and queen of correction

2.9.8 Statistical power

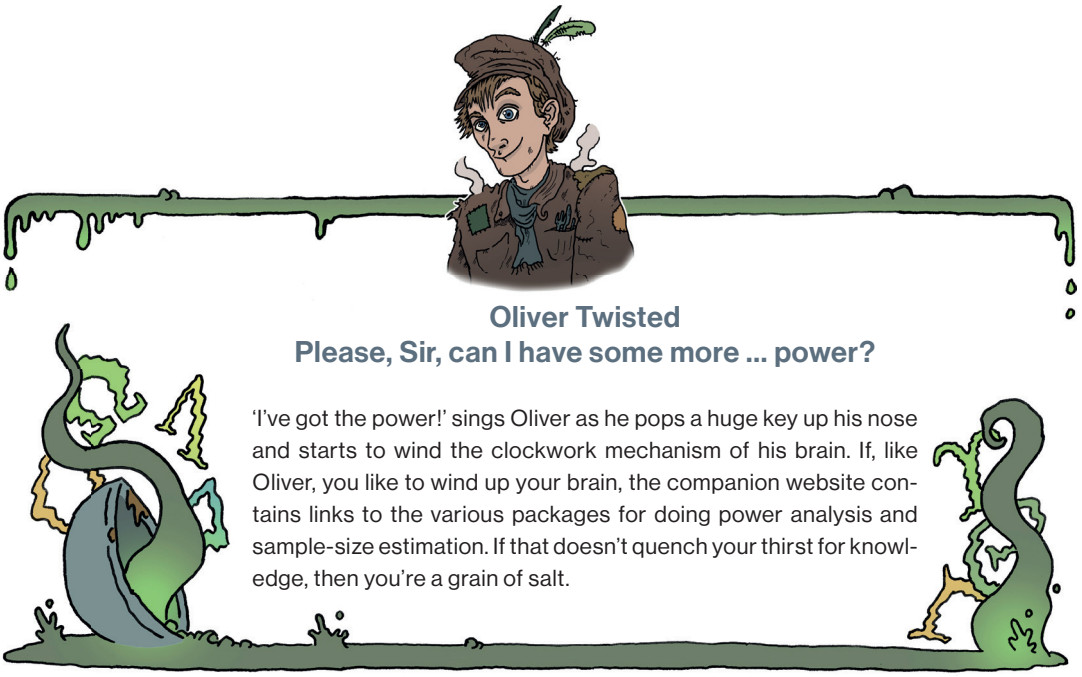
We have seen that it is important to control the Type I error rate so that we don't too often mistakenly think that an effect is genuine when it is not. The opposite problem relates to the Type II error, which is how often we will miss an effect in the population that genuinely exists. If we set the Type II error rate high then we will be likely to miss a lot of genuine effects, but if we set it low we will be less likely to miss effects. The ability of a test to find an effect is known as its statistical **power** (not to be confused with statistical powder, which is an illegal substance that makes you better understand statistics). The power of a test is the probability that a given test will find an effect assuming that one exists in the population. This is the opposite of the probability that a given test will *not* find an effect assuming that one exists in the population, which, as we have seen, is the β-level (i.e., Type II error rate). Therefore, the power of a test can be expressed as $1 - \beta$. Given that Cohen (1988, 1992) recommends a 0.2 probability of failing to detect a genuine effect (see above), the corresponding level of power would be 1 – 0.2, or 0.8. Therefore, we typically aim to achieve a power of 0.8, or put another way, an 80% chance of detecting an effect if one genuinely exists. The power of a statistical test depends on:[12]

1. How big the effect is, because bigger effects will be easier to spot. This is known as the effect size and we'll discuss it in Section 3.5).
2. How strict we are about deciding that an effect is significant. The stricter we are, the harder it will be to 'find' an effect. This strictness is reflected in the α-level. This brings us back to our point in the previous section about correcting for multiple tests. If we use a more conservative Type I error rate for each test (such as a Bonferroni correction) then the probability of rejecting an effect that does exist is increased (we're more likely to make a Type II error). In other words, when we apply a Bonferroni correction, the tests will have less power to detect effects.
3. The sample size: We saw earlier in this chapter that larger samples are better approximations of the population; therefore, they have less sampling error. Remember that test statistics are basically a signal-to-noise ratio, so given that large samples have less 'noise', they make it easier to find the 'signal'.

Given that power ($1 - \beta$), the α-level, sample size, and the size of the effect are all linked, if we know three of these things, then we can find out the remaining one. There are two things that scientists do with this knowledge:

1. **Calculate the power of a test**: Given that we've conducted our experiment, we will have already selected a value of α, we can estimate the effect size based on our sample data, and we will know how many participants we used. Therefore, we can use these values to calculate $1 - \beta$, the power of our test. If this value turns out to be 0.8 or more, then we can be confident that we have achieved sufficient power to detect any effects that might have existed, but if the resulting value is less, then we might want to replicate the experiment using more participants to increase the power.
2. **Calculate the sample size necessary to achieve a given level of power**: We can set the value of α and $1 - \beta$ to be whatever we want (normally, 0.05 and 0.8, respectively). We can also estimate the likely effect size in the population by using data from past research. Even if no one had previously done the exact experiment that we intend to do, we can still estimate the likely effect size based on similar experiments. Given this information, we can calculate how many participants we would need to detect that effect (based on the values of α and $1 - \beta$ that we've chosen).

12 It will also depend on whether the test is a one- or two-tailed test (see Section 2.9.5), but, as we have seen, you'd normally do a two-tailed test.



Oliver Twisted
Please, Sir, can I have some more ... power?

'I've got the power!' sings Oliver as he pops a huge key up his nose and starts to wind the clockwork mechanism of his brain. If, like Oliver, you like to wind up your brain, the companion website contains links to the various packages for doing power analysis and sample-size estimation. If that doesn't quench your thirst for knowledge, then you're a grain of salt.

The point of calculating the power of a test after the experiment has always been lost on me a bit: if you find a non-significant effect then you didn't have enough power, if you found a significant effect, then you did. Using power to calculate the necessary sample size is the more common and, in my opinion, more useful thing to do. The actual computations are very cumbersome, but there are computer programs available that will do them for you. *G*Power* is a free and powerful (excuse the pun) tool, there is a package *pwr* that can be used in the open source statistics package R, and various websites, including powerandsamplesize.com. There are also commercial software packages such as *nQuery Adviser* (www.statsols.com/nquery-sample-size-calculator), *Power and Precision* (www.power-analysis.com) and *PASS* (www.ncss.com/software/pass). Also, Cohen (1988) provides extensive tables for calculating the number of participants for a given level of power (and vice versa).

2.9.9 Confidence intervals and statistical significance

I mentioned earlier (Section 2.8.4) that if 95% confidence intervals didn't overlap then we could conclude that the means come from different populations, and, therefore, that they are significantly different. I was getting ahead of myself a bit because this comment alluded to the fact that there is a relationship between statistical significance and confidence intervals. Cumming & Finch (2005) have three guidelines that are shown in Figure 2.16:

1. 95% confidence intervals that just about touch end-to-end (as in the top left panel of Figure 2.16) represent a p-value for testing the null hypothesis of no differences of approximately 0.01.
2. If there is a gap between the upper end of one 95% confidence interval and the lower end of another (as in the top right panel of Figure 2.16), then $p < 0.01$.
3. A p-value of 0.05 is represented by *moderate* overlap between the bars (the bottom panels of Figure 2.16).

These guidelines are poorly understood by many researchers. In one study (Belia, Fidler, Williams, & Cumming, 2005), 473 researchers from medicine, psychology and behavioural neuroscience were

shown a graph of means and confidence intervals for two independent groups and asked to move one of the error bars up or down on the graph until they showed a 'just significant difference' (at $p < 0.05$). The sample ranged from new researchers to very experienced ones but, surprisingly, this experience did not predict their responses. In fact, only a small percentage of researchers could position the confidence intervals correctly to show a just significant difference (15% of psychologists, 20% of behavioural neuroscientists and 16% of medics). The most frequent response was to position the confidence intervals more or less at the point where they stop overlapping (i.e., a p-value of approximately 0.01). Very few researchers (even experienced ones) realized that moderate overlap between confidence intervals equates to the standard p-value of 0.05 for accepting significance.

What do we mean by moderate overlap? Cumming (2012) defines it as half the length of the average margin of error (MOE). The MOE is half the length of the confidence interval (assuming it is symmetric), so it's the length of the bar sticking out in one direction from the mean. In the bottom left of Figure 2.16 the confidence interval for sample 1 ranges from 4 to 14 so has a length of 10 and an MOE of half this value (i.e., 5). For sample 2, it ranges from 11.5 to 21.5 so again a distance of 10 and an MOE of 5. The average MOE is, therefore $(5 + 5)/2 = 5$. Moderate overlap would be half of this value

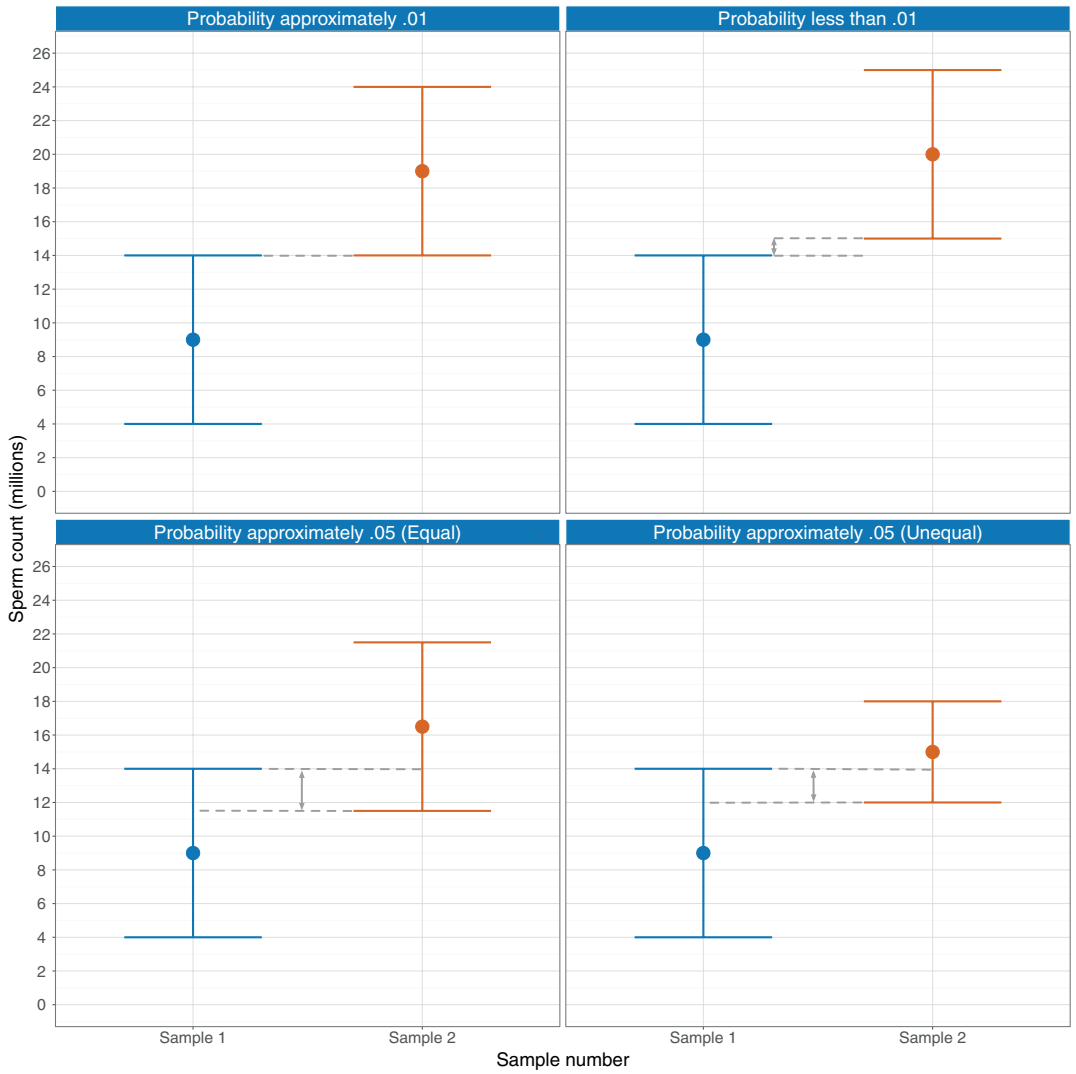


Figure 2.16 The relationship between confidence intervals and statistical significance

(i.e., 2.5). This is the amount of overlap between the two confidence intervals in the bottom left of Figure 2.16. Basically, if the confidence intervals are the same length, then $p = 0.05$ is represented by an overlap of about a quarter of the confidence interval. In the more likely scenario of confidence intervals with different lengths, the interpretation of overlap is more difficult. In the bottom right of Figure 2.16 the confidence interval for sample 1 again ranges from 4 to 14 so has a length of 10 and an MOE of 5. For sample 2, it ranges from 12 to 18 and so a distance of 6 and an MOE of half this value, 3. The average MOE is, therefore, $(5 + 3)/2 = 4$. Moderate overlap would be half of this value (i.e., 2) and this is what we get in the bottom right of Figure 2.16: the confidence intervals overlap by 2 points on the scale, which equates to a p of around 0.05.

2.9.10 Sample size and statistical significance

When we discussed power, we saw that it is intrinsically linked with the sample size. Given that power is the ability of a test to find an effect that genuinely exists, and we 'find' an effect by having a statistically significant result (i.e., $p < 0.05$), there is also a connection between the sample size and the p-value associated with a test statistic. We can demonstrate this connection with two examples. Apparently, male mice 'sing' to female mice to try to attract them as mates (Hoffmann, Musolf, & Penn, 2012), I'm not sure what they sing, but I like to think it might be 'This mouse is on fire' by AC/DC, or perhaps 'Mouses of the Holy' by Led Zeppelin, or even 'The mouse Jack built' by Metallica. It's probably not 'Terror and hubris in the mouse of Frank Pollard' by Lamb of God. That would just be weird. Anyway, many a young man has spent time wondering how best to attract female mates, so to help them out, imagine we did a study in which we got two groups of 10 heterosexual young men and got them to go up to a woman that they found attractive and either engage them in conversation (group 1) or sing them a song (group 2). We measured how long it was before the woman ran away. Imagine we repeated this experiment but using 100 men in each group.

Figure 2.17 shows the results of these two experiments. The summary statistics from the data are identical: in both cases the singing group had a mean of 10 and a standard deviation of 3, and the conversation group had a mean of 12 and a standard deviation of 3. Remember that the only difference between the two experiments is that one collected 10 scores per sample, and the other 100 scores per sample.

Compare the graphs in Figure 2.17. What effect does the difference in sample size have? Why do you think it has this effect?

Notice in Figure 2.17 that the means for each sample are the same in both graphs, but the confidence intervals are much narrower when the samples contain 100 scores than when they contain only 10 scores. You might think that this is odd given that I said that the standard deviations are all the same (i.e., 3). If you think back to how the confidence interval is computed it is the mean plus or minus 1.96 times the standard error. The standard error is the standard deviation divided by the square root of the sample size (see Eq. 2.14), therefore as the sample size gets larger, the standard error (and, therefore, confidence interval) will get smaller.

We saw in the previous section that if the confidence intervals of two samples are the same length then a p of around 0.05 is represented by an overlap of about a quarter of the confidence interval. Therefore, we can see that even though the means and standard deviations are identical in both

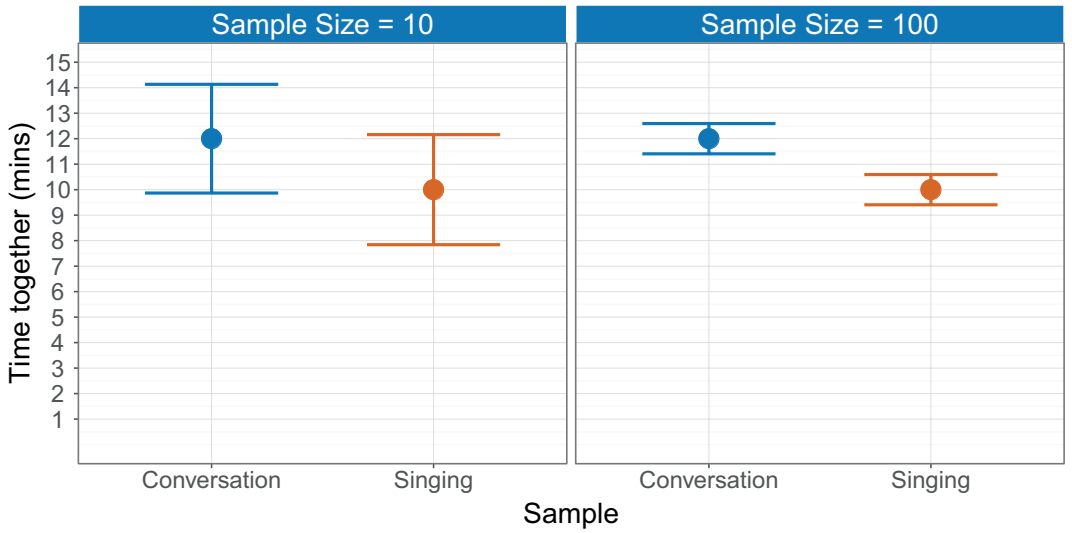


Figure 2.17 Graph showing two data sets with the same means and standard deviations but based on different-sized samples

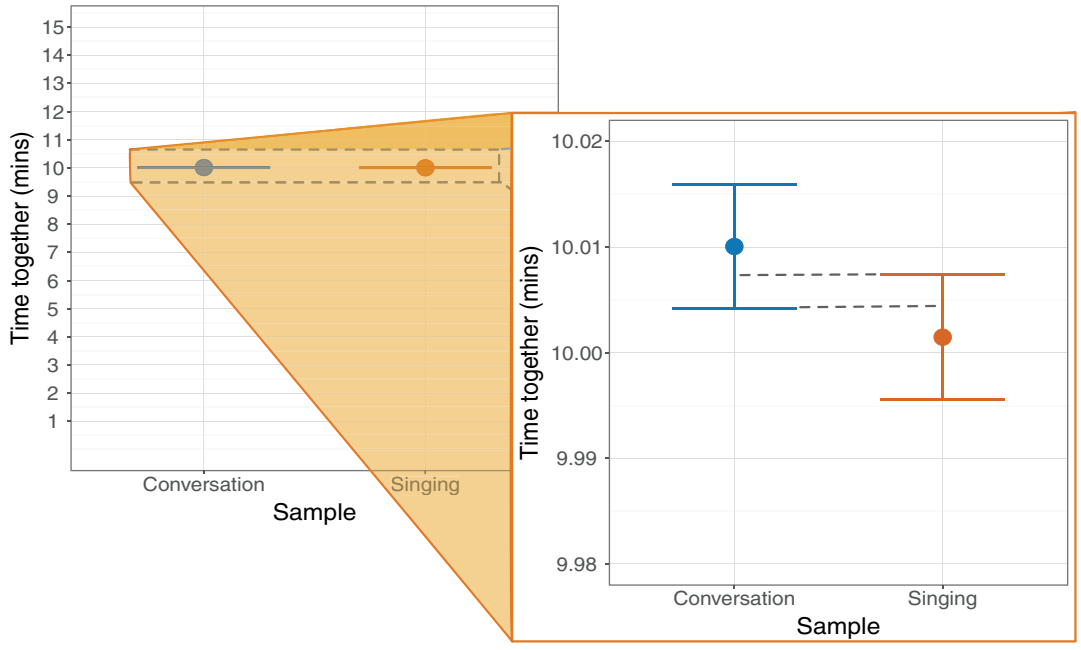


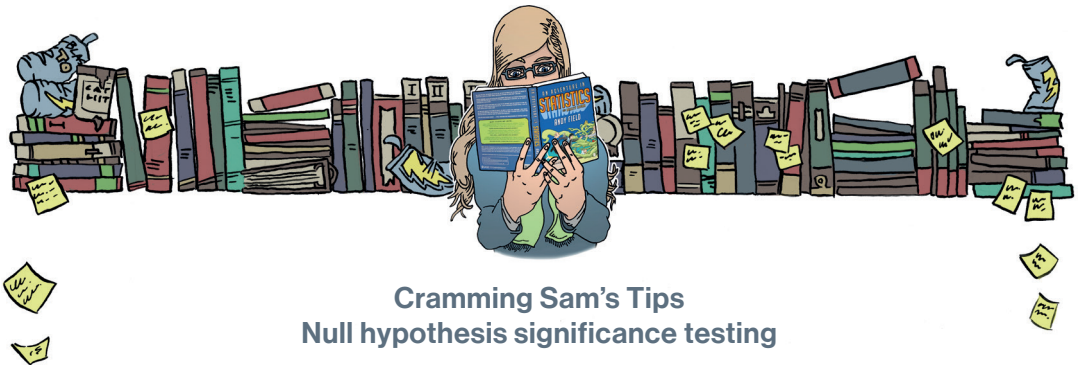
Figure 2.18 A very small difference between means based on an enormous sample size ($n = 1{,}000{,}000$ per group)

graphs, the study that has only 10 scores per sample is not significant (the bars overlap quite a lot; in fact, $p = 0.15$), but the study that has 100 scores per sample shows a highly significant difference (the bars don't overlap at all; for these data $p < 0.001$). Remember, the means and standard deviations are *identical* in the two graphs, but the sample size affects the standard error and hence the significance.

Taking this relationship to the extreme, we can illustrate that with a big enough sample even a completely meaningless difference between two means can be deemed significant, with $p < 0.05$.

Figure 2.18 shows such a situation. This time, the singing group has a mean of 10.00 ($SD = 3$) and the conversation group has a mean of 10.01 ($SD = 3$): a difference of 0.01 – a very small difference indeed. The main graph looks very odd: the means look identical and there are no confidence intervals. In fact, the confidence intervals are so narrow that they merge into a single line. The figure also shows a zoomed image of the confidence intervals (note that by zooming in the values on the vertical axis range from 9.98 to 10.02, so the entire range of values in the zoomed image is only 0.04). As you can see, the sample means are 10 and 10.01 as mentioned before,[13] but by zooming in we can see the confidence intervals. Note that the confidence intervals show an overlap of about a quarter, which equates to a significance value of about $p = 0.05$ (for these data the actual value of p is 0.044). How is it possible that we have two sample means that are almost identical (10 and 10.01), and have the same standard deviations, but are deemed significantly different? The answer is again the sample size: there are 1 million cases in each sample, so the standard errors are minuscule.

This section has made two important points. First, the sample size affects whether a difference between samples is deemed significant or not. *In large samples, small differences can be significant,*



Cramming Sam's Tips
Null hypothesis significance testing

- NHST is a widespread method for assessing scientific theories. The basic idea is that we have two competing hypotheses: one says that an effect exists (the *alternative hypothesis*) and the other says that an effect doesn't exist (the *null hypothesis*). We compute a test statistic that represents the alternative hypothesis and calculate the probability that we would get a value as big as the one we have if the null hypothesis were true. If this probability is less than 0.05 we reject the idea that there is no effect, say that we have a *statistically significant* finding and throw a little party. If the probability is greater than 0.05 we do not reject the idea that there is no effect, we say that we have a *non-significant* finding and we look sad.
- We can make two types of error: we can believe that there is an effect when, in reality, there isn't (a *Type I error*); and we can believe that there is not an effect when, in reality, there is (a *Type II error*).
- The power of a statistical test is the probability that it will find an effect when one exists.
- The significance of a test statistic is directly linked to the sample size: the same effect will have different p-values in different-sized samples, small differences can be deemed 'significant' in large samples, and large effects might be deemed 'non-significant' in small samples.

13 The mean of the singing group looks bigger than 10, but this is only because we have zoomed in so much that its actual value of 10.00147 is noticeable.

and in small samples large differences can be non-significant. This point relates to power: large samples have more power to detect effects. Second, even a difference of practically zero can be deemed 'significant' if the sample size is big enough. Remember that test statistics are effectively the ratio of signal to noise, and the standard error is our measure of 'sampling noise'. The standard error is estimated from the sample size, and the bigger the sample size, the smaller the standard error. Therefore, bigger samples have less 'noise', so even a tiny signal can be detected.

2.10 Reporting significance tests

In Section 1.9 we looked at some general principles for reporting data. Now that we have learnt a bit about fitting statistical models, we can add to these guiding principles. We learnt in this chapter that we can construct confidence intervals (usually 95% ones) around a parameter such as the mean. A 95% confidence interval contains the population value in 95% of samples, so if your sample is one of those 95%, the confidence interval contains useful information about the population value. It is important to tell readers the type of confidence interval used (e.g., 95%), and in general we use the format [*lower boundary*, *upper boundary*] to present the values. So, if we had a mean of 30 and the confidence interval ranged from 20 to 40, we might write $M = 30$, 95% CI [20, 40]. If we were reporting lots of 95% confidence intervals it might be easier to state the level at the start of our results and just use the square brackets:

✓ 95% confidence intervals are reported in square brackets. Fear reactions were higher, $M = 9.86$ [7.41, 12.31] when Andy's cat Fuzzy wore a fake human tongue compared to when he did not, $M = 6.58$ [3.47, 9.69].

We also saw that when we fit a statistical model we calculate a test statistic and a p-value associated with it. Scientists typically conclude that an effect (our model) is significant if this p-value is less than 0.05. APA style is to remove the zero before the decimal place (so you'd report $p = .05$ rather than $p = 0.05$) but, because many other journals don't have this idiosyncratic rule, this is an APA rule that I don't follow in this book. Historically, people would report p-values as being either less than or greater than 0.05. They would write things like:

× Fear reactions were significantly higher when Andy's cat Fuzzy wore a fake human tongue compared to when he did not, $p < 0.05$.

If an effect was very significant (e.g., if the p-value was less than 0.01 or even 0.001), they would also use these two criteria to indicate a 'very significant' finding:

× The number of cats intruding into the garden was significantly less when Fuzzy wore a fake human tongue compared to when he didn't, $p < 0.01$.

Similarly, non-significant effects would be reported in much the same way (note this time the p is reported as greater than 0.05):

× Fear reactions were not significantly different when Fuzzy wore a David Beckham mask compared to when he did not, $p > 0.05$.

In the days before computers, it made sense to use these standard benchmarks for reporting significance because it was a bit of a pain to compute exact significance values (Jane Superbrain Box 3.1). However, computers make computing p-values a piece of *p*s, so we have no excuse for using these

conventions. We should report exact p-values because it gives the reader more information than simply knowing that the p-value was less or more than a random threshold like 0.05. The possible exception is the threshold of 0.001. If we find a p-value of 0.0000234 then for the sake of space and everyone's sanity it would be reasonable to report $p < 0.001$:

- ✓ Fear reactions were significantly higher when Andy's cat Fuzzy wore a fake human tongue compared to when he did not, $p = 0.023$.
- ✓ The number of cats intruding into the garden was significantly less when Fuzzy wore a fake human tongue compared to when he did not, $p = 0.007$.

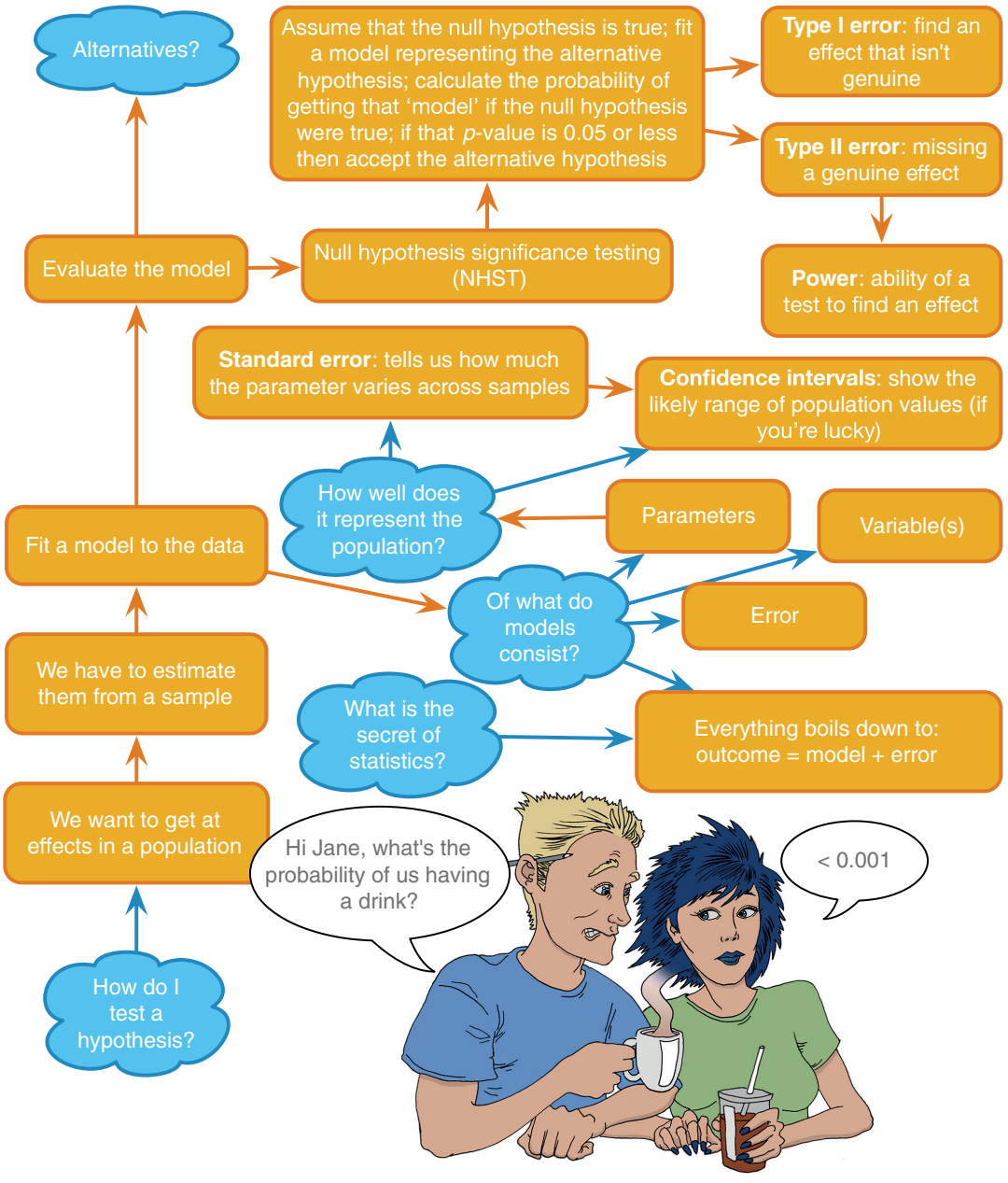


Figure 2.19 What Brian learnt from this chapter

2.11 Brian's attempt to woo Jane

Brian was feeling a little deflated after his encounter with Jane. She had barely said a word to him. Was he that awful to be around? He was enjoying the book she'd handed to him though. The whole thing had spurred him on to concentrate more during his statistics lectures. Perhaps that's what she wanted.

He'd seen Jane flitting about campus. She always seemed to have a massive bag with her. It seemed to contain something large and heavy, judging from her posture. He wondered what it was as he day-dreamed across the campus square. His thoughts were broken by being knocked to the floor.

'Watch where you're going,' he said angrily.

'I'm so sorry ...' the dark-haired woman replied. She looked flustered. She picked up her bag as Brian looked up to see it was Jane. Now *he* felt flustered.

Jane looked as though she wanted to expand on her apology but the words had escaped. Her eyes darted as though searching for them.

Brian didn't want her to go, but in the absence of anything to say, he recited his statistics lecture to her. It was weird. Weird enough that as he finished she shrugged and ran off.

2.12 What next?

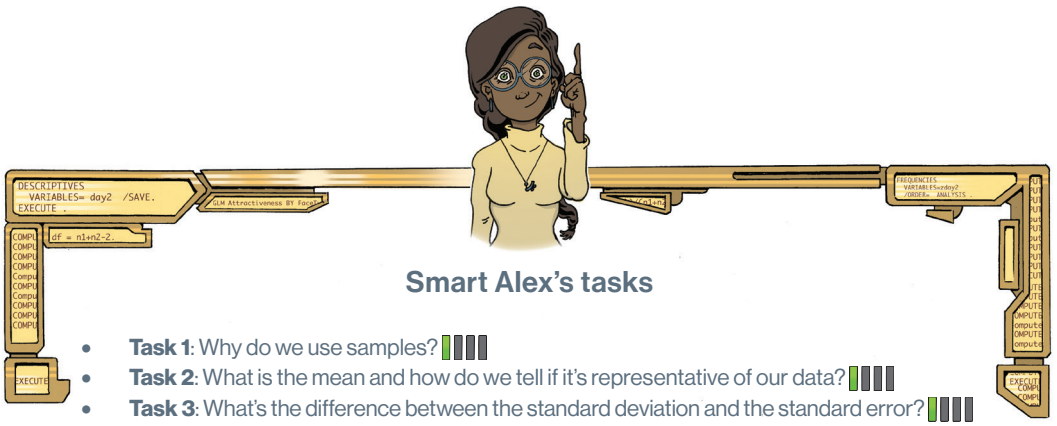
Nursery school was the beginning of an educational journey that I am a still on several decades later. As a child, your belief systems are very adaptable. One minute you believe that sharks can miniaturize themselves, swim up pipes from the sea to the swimming pool you're currently in before restoring their natural size and eating you; the next minute you don't, simply because your parents say it's not possible. At the age of 3 any hypothesis is plausible, every way of life is acceptable, and multiple incompatible perspectives can be accommodated. Then a bunch of idiot adults come along and force you to think more rigidly. Suddenly, a cardboard box is not a high-tech excavator, it's a cardboard box, and there are 'right' and 'wrong' ways to live your life. As you get older, the danger is that – left unchecked – you plunge yourself into a swimming-pool-sized echo chamber of your own beliefs, leaving your cognitive flexibility in a puddle at the edge of the pool. If only sharks *could* compress themselves ...

Before you know it, you're doggedly doing things and following rules that you can't remember the reason for. One of my early beliefs was that my older brother Paul (more on him later ...) was 'the clever one'. Far be it from me to lay the blame at anyone's feet for this belief, but it probably didn't help that members of my immediate family used to say things like, 'Paul is the clever one, but at least you work hard'. Like I said, over time, if nothing challenges your view you can get very fixed in a way of doing things or a mode of thinking. If you spend your life thinking you're not 'the clever one', how can you ever change that? You need something unexpected and profound to create a paradigm shift. The next chapter is all about breaking ways of thinking that scientists have been invested in for a long time.

2.13 Key terms that I've discovered

α-level
Alternative hypothesis
β-level
Bonferroni correction
Central limit theorem
Confidence interval
Degrees of freedom
Deviance
Experimental hypothesis
Experimentwise error rate
Familywise error rate
Fit
Interval estimate
Linear model
Method of least squares
Null hypothesis
One-tailed test
Ordinary least squares

Parameter	Sampling distribution	Two-tailed test
Point estimate	Sampling variation	Type I error
Population	Standard error	Type II error
Power	Standard error of the mean (SE)	
Sample	Test statistic	



Smart Alex's tasks

- **Task 1**: Why do we use samples?
- **Task 2**: What is the mean and how do we tell if it's representative of our data?
- **Task 3**: What's the difference between the standard deviation and the standard error?
- **Task 4**: In Chapter 1 we used an example of the time taken for 21 heavy smokers to fall off a treadmill at the fastest setting (18, 16, 18, 24, 23, 22, 22, 23, 26, 29, 32, 34, 34, 36, 36, 43, 42, 49, 46, 46, 57). Calculate the standard error and 95% confidence interval of these data.
- **Task 5**: What do the sum of squares, variance and standard deviation represent? How do they differ?
- **Task 6**: What is a test statistic and what does it tell us?
- **Task 7**: What are Type I and Type II errors?
- **Task 8**: What is statistical power?
- **Task 9**: Figure 2.17 shows two experiments that looked at the effect of singing versus conversation on how much time a woman would spend with a man. In both experiments the means were 10 (singing) and 12 (conversation), the standard deviations in all groups were 3, but the group sizes were 10 per group in the first experiment and 100 per group in the second. Compute the values of the confidence intervals displayed in the figure.
- **Task 10**: Figure 2.18 shows a similar study to the one above, but the means were 10 (singing) and 10.01 (conversation), the standard deviations in both groups were 3, and each group contained 1 million people. Compute the values of the confidence intervals displayed in the figure.
- **Task 11**: In Chapter 1 (Task 8), we looked at an example of how many games it took a sportsperson before they hit the 'red zone'. Calculate the standard error and confidence interval for those data.
- **Task 12**: At a rival club to the one I support, they similarly measured the number of consecutive games it took their players before they reached the red zone. The data are: 6, 17, 7, 3, 8, 9, 4, 13, 11, 14, 7. Calculate the mean, standard deviation, and confidence interval for these data.
- **Task 13**: In Chapter 1 (Task 9) we looked at the length in days of 11 celebrity marriages. Here are the lengths in days of eight marriages, one being mine and the other seven being those of some of my friends and family (in all but one case up to the day I'm writing this, which is 8 March 2012, but in the 91-day case it was the entire duration – this isn't my marriage, in case you're wondering): 210, 91, 3901, 1339, 662, 453, 16672, 21963, 222. Calculate the mean, standard deviation and confidence interval for these data.

Answers & additional resources are available on the book's website at
https://edge.sagepub.com/field5e