The importance of incorporating researcher beliefs into statistical models.

Lauren Ashlee Kennedy

February, 2018

Abstract

In this thesis I consider how statistical assumptions are driven by the assumptions the researcher makes about the data. I focus specifically on assumptions surrounding data generation, namely: a) the shape of distribution expected, b) the process by which data were obtained, c) the shape of the outcome distribution, and d) inferring information about missing data.

Each chapter of this thesis will focus on one of these assumptions using a combination of tools. I use existing methods and propose new models before exploring from a cognitive perspective the types of inference people make. This allows us to explore the concept of researcher assumptions, and to consider where building them in to the statistical model might be beneficial. In three of the four main chapters of this thesis, I use simulation methods to compare models. The models I consider are both Bayesian and frequentist in framework. The aim of these simulations is not to compare frameworks, but to compare different model structures to ascertain the structure that allows the most accurate claims about the data to be made.

There are four main arguments presented in this thesis. First I argue that it is very rare to conduct statistical tests without making some sort of assumption about the data. Second, I demonstrate that for distributional assumptions in a particular type of data, models where the assumptions are not violated can improve the accuracy of the claims made. Thirdly, I present two models that match the assumed generative process of two types of data; contaminated data and data with a heterogeneous effect. I demonstrate that these models are not only more accurate, they also allow the researcher to make richer claims about their data. Finally I experimentally investigate a well-known finding in cognitive psychology—a dislike for ambiguous or missing data. I replicate this preference whilst demonstrating that people are still sensitive to underlying distributional information. Together these findings suggest that the researcher is both sensitive to and makes assumptions about the data. Creating and using statistical models that do not violate the assumptions the researcher makes is important, but building more complicated assumptions into the model can give a richer and more accurate understanding of the data.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968. I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Lauren Kennedy, 8^{th} September, 2017

Acknowledgements

A single name on the cover of a thesis belies the many people who have given it life. I feel very fortunate to have had the support of a large number of remarkable people.

My supervisors Dani Navarro and Amy Perfors, who took a chance and invited an undergrad into their lab. Thank you both for sharing your enjoyment and dedication to science in the many years since then. I can't imagine a better lab to spend so many years in. Dani, thank you especially for taking a 'frequentist' into your lab, and for the many chats about statistics that eventually swayed me over to Bayes. Amy, thank you for your enthusiasm not just for science, but for communicating science to others. Thank you also to Anna, who stepped in as my primary supervisor.

Many thanks to the many other researchers who devoted endless coffee breaks to discussing all manner of interesting topics; Drew, Wai Keen, Emma, Dinis, John, Carolyn, Jess, Keith, Wouter, Sean, Simon, Kristi, Heidi and the many other citizens of level one Hughes and above. What an utterly wonderful collection of people to work with! Thanks especially to Jess, who made the last four years fly past with many tea breaks and laughter.

Many thanks to the people of Australia, who enabled this research through their generous support of the Australian Government Research Training Program Scholarships, and the Australian Research Council. I hope I did you proud.

I would also like to thank the examiners of this thesis, whose correction and advice was gratefully accepted.

Much gratitude to the admin staff, both present and past. Special thanks to Deidre, who was always willing to lend a key when I had locked myself out; Angela, without whom I would surely be still stuck in an airport somewhere; and Lynda, who knew more myths and stories about Hughes than anyone else.

I am exceptionally grateful for the love and support of my family; Justin, Kree, Belinda, Michael, Catherine and Mum; as well as my wonderful in-laws; Chris, Laurie, Laura and 'the Sherwoods'. Thank you all.

I'd like to thank my cat, Jazz, for her company during the long nights when I wrote the bulk of this thesis. Unfortunately, I can't, because she promptly abandoned me for her warm bed. Every. Single. Time. Thanks for being there in spirit, cat.

Lastly my wonderful husband Brad, whose love, humour, and unfailing certainty that I could do it made all the difference.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> An experiment is a question which science poses to Nature, and a measurement is the recording of Nature's answer.
>
> *Max Planck*

Max Planck researched in physics, a world where measurement can map very closely to the truth, and the difference between the two can be quantified. For him experiments and measurement formed the two cornerstones of science. If we in psychology, had the luxury of the measurement capabilities that physics boasted then perhaps we too would only need to ask and measure. Unfortunately in psychology (as in modern day Physics), measurement error is both guaranteed and very difficult to perfectly estimate. Although we can ask nature a question, we need statistics to help us interpret the answer. In this thesis I both apply and propose statistical models to help researchers in psychological and cognitive fields find the answer to the questions they ask.

The central theme of this thesis is to consider the interaction between the question, answer and conclusion (experimental design, measurement and statistics) in psychological and cognitive science. I will emphasise the relationship between researcher *intention* (the question), researcher *interpretation* (of the answer given) and statistical methods. To achieve this I will first briefly discuss traditional measurement theories of classical, generalisability ('G') and item response theory. Each theory makes different assumptions about how the measurement is connected to the *true* score, and so differs in which statistical techniques should be employed to account for it. The relevance and importance of researcher assumptions holds true for all kinds of statistics, from frequentist to Bayesian; no approach is truly impartial, although frequentist statistics is sometimes claimed to be (Reiss & Sprenger, 2017).

This is not to say that researcher assumptions are not contentious. Simmons, Nelson, and Simonsohn (2011) argued that researcher freedom in statistical analysis can lead to dramatically high false positive rates. As a solution to this they argue that if contentious analysis additions are conducted (such as the removal of certain outliers) then the analysis should be reported both with and without this addition. This seems like a plausible solution, but as Gelman and Loken (2013) point out, there is a very large set of potential additions and choices that change analyses, and *most of them are defensible.* Another potential solution is preregistration, where all of the planned analyses are declared before the data are collected, but this limits exploratory analyses (Humphreys, Sanchez de la Sierra, & Van der Windt, 2013). In this chapter I wish to demonstrate that even very simple statistical methods represent a large number of assumptions that are *already* made and incorporated into the model. I argue that assumptions that are explicitly incorporated into the model are beneficial as they allow researchers in the field to define and discuss different beliefs (such as the nature of error in psychological research) in a rigorous fashion. To illustrate this I introduce the interaction between measurement theory, statistical methods and interpretation. I first consider this in a strictly measurement context, before branching out to consider other statistical methods and assumptions.

I will then focus on four assumptions that are exceptionally common within psychological and cognitive research. I describe why these assumptions are important, why they were chosen and when there is a strong probability that they might be violated. The main body of this thesis will build on this with a chapter that focuses on each assumption violation in context. Each chapter considers the evidence for this violation, its potential consequences, and potential solutions. In chapter three we consider established methods of modelling skewed data, whilst in chapters four and five I present alternate models that better match the data assumptions. In chapter six I step back to consider how assumptions are created and acted upon in a specific example.

For each assumption violation, there are a number of potential statistical models that may provide a good alternative. Given this, how should we determine the best model to use? Throughout the chapters I consider three main factors when comparing and contrasting the models. The first is the Type 1 error rate, or the ability of the model to correctly report no effect when one is truly not present. The second is the Type 2 error rate, or the ability of the model to correctly report an effect when one is truly present. The last is the relative informativeness of the model, or how well the claims of the model match the claims the researcher would like to make. Whilst these three factors form a complete picture of many of the concerns researchers have (or should have) when they choose a statistical test, they do not directly consider the problem of model complexity.

I discuss model complexity in the discussion section of this thesis. Model complexity refers to the trade-off between *observed data fit* and *accuracy of future predictions*. The more complex a model, the greater its ability to fit random variation in data that is not reflective of a true relationship between the variables. This *over-fitting* results in the model estimating random variation as if it were true. By the very nature of randomness, future data is unlikely to have the same random variation as the current observed data. When an over-fitted model is used to predict new data, we find that it is much less accurate than when it was fit to the observed dataset. Indeed, a model that fits the observed data slightly less well might predict new data better. I argue that complex models play an important data analysis role if they aid the researcher to make claims about aspects of the data that simpler models cannot, provided this is not at the expense of model description or prediction.

The work in this thesis begins from the understanding that measurement and other data generation theories held by the researcher change the type, intention and interpretation of statistics used. Each of the main chapters highlights one of four different assumption violations, with discussion on what processes produce these violations and potential solutions. Whilst drawing these ideas together is relatively uncommon within the wider body of psychological research, it is not unique. My unique contribution to the field is the novel application of existing models and the proposal of new models that aim to solve these problems.

1.1 Conventions, abbreviations and common terms

For ease of reading, I will summarise commonly used terms and/or symbols in this section.

- **Bold case**: Where possible bold text will be used to represent vector variables, plain text to represent scalar variables.
- **i.i.d.**: Identical and independently distributed.
- α: Internal consistency or Type 1 error rate, as given by context.
- β: The Type 2 error rate.
- **Power**: Defined as $1 - \beta$
- **N**: Denotes the number of observations
- **Observed Value**: This is the measurement for participant i. If there is only one measurement per participant then this is represented as x_i. If there are k items to measure the variable of interest, then this is represented as $x_{i,k}$. If the same measurement is undertaken on each participant j times, this is written as $x_{i,j}$. If there are k items measured on participant i at the j^{th} time point, this is represented as $x_{i,k,j}$.
- **True Value**: The true value is the true value that would be observed if our measurement method was perfect and introduced no error. In reality this is

never the case. In practice researchers infer the true value of x, t, from a set of assumptions about the error that is introduced by measurement.

- **Measurement Error**: This is the difference between the true score and the observed value. It is generally assumed to be drawn from a normal distribution with $\mu = 0$ and represented by epsilon, i.e., $\epsilon \sim N(0, \sigma)$. This implies that the measurement is unbiased but still introduces variance to the scores.

- **Standard Error of Measurement (SEM)**: This is an estimation of the standard deviation of the normal distribution discussed in the previous item. The equation (Equation 1.1) requires the reliability of the measure (estimation of reliability will be discussed in later sections of this introduction), and the standard deviation of the observed scores for the sample of interest.

$$SEM = \sigma\sqrt{1 - r} \tag{1.1}$$

- **Latent Value/Variable**: If a group of k ($k \geqslant 1$) items are all designed to measure one underlying construct, then this underlying construct can be considered a *latent* variable. Like the true value, the latent variable is never directly observed, but instead inferred from a set of observed responses to items[1]. Throughout this thesis, the latent variable will be denoted with a lower case l such that the value for the latent variable for participant i at the j^{th} time point is $l_{i,j}$.

- **Error**: Previously I discussed the concept of measurement error, but error

[1] If $k = 1$, then the latent and true value are equivalent.

can be introduced at many different levels within a model (i.e., the difference between observed Y and predicted $\hat{Y}$ in linear regression). Other sources of error will be discussed as needed throughout this thesis.

- **Bayesian Updating**: Later in this introduction I discuss the difference between two opposing statistical approaches. One of these is Bayesian statistics, where the probability of a set of model parameters (θ) given the observed data is proportional to the product of the probability of the data (x) given the model (the likelihood) and the probability of the model (the prior), as shown formally in Equation 1.2.

$$P(\theta|x) \approx P(x|\theta)P(\theta) \tag{1.2}$$

- **Graphical Models**: In later chapters I will propose novel Bayesian models. To communicate these models, I will employ probabilistic graphical models (PGMs) to describe the likelihood. This is a very common way of visualising the relationship between variables (e.g, see Jordan, 1998). Each variable is represented at a node (either a circle or a square), with the solid lines that join the nodes indicating probabilistic relationships. Nodes that have a double outline are deterministic (i.e. not a random variable itself, but directly calculated from either random or observed variables). Nodes that are square represent discrete random variables, while nodes that are round denote continuous. The colouring of the node determines whether the node was observed (shaded) or not observed (transparent). Large boxes around a group of nodes indicate these nodes a replicated (e.g., for each participant, for each group).

An example of this is shown in Figure 1.1.

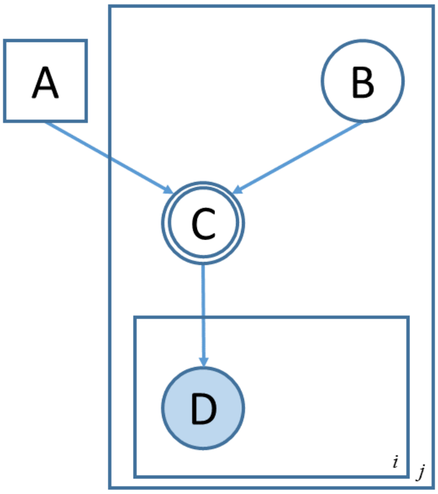


Figure 1.1: **Example Probabilistic Graphical Model (PGM)** This figure is a relatively simple example of the graphical models we use throughout this thesis to demonstrate the likelihood, or the dependencies between the different model parameters. Here A, B and D represent random variables, whilst C represents a deterministic variable (*double outline*) that can be formed from A and B (*direction of the arrows*). As A is *square* this variable is a discrete variable, whilst B, C and D are continuous (*round*). All of the variables are not observed, except for D (shaded). The two large boxes describe that for each j (where j might be experimental group) there is a different B and C. For each i (where i might represent each individual participant) within each group j, a different value for variable D is observed.

- **MCMC** Markov Chain Monte-Carlo sampling methods

- **JAGS** Just Another Gibbs Sampler; a language and code base.

1.2 Measurement theory and inference about error.

If we measured the *entire* population of interest with *complete accuracy* (i.e. observed directly all of the true observed values, t_i) then we wouldn't need to do inference at all. We would know the true value of the unobserved variable(s) for all individuals i. The statistical task would be to focus on describing the latent variable(s) and any relationship between them. Unfortunately psychological constructs (such as scholastic ability and personality factors) always have some error in measurement. This error introduces uncertainty in our estimates of the true score at both the individual level. In addition, ethical and practical constraints limit the nature of experimental manipulation. We also rarely sample the full population. This combines with measurement error to introduce uncertainty at the population level. In the next sections I will discuss different theories of measurement error with the purpose of demonstrating the assumptions made in very simple statistical methods and how these assumptions are often taken for granted or ignored. In later chapters I consider additional and different assumptions that result in models that are more robust and allow for richer claims to be made about the data.

When we consider measurement error, statistics has three main purposes: First, to estimate the size of the measurement error. Second, to use this information and the observed value x_i to estimate the true variable t_i. Third, to estimate the degree of uncertainty of our estimate of t_i, and then to do the same again with any latent variables. With these three purposes in mind, we turn briefly to the three main theories of measurement: classical, generalisability and item response, to see how each one accounts for different levels of measurement error.

1.2.1 Classical Test Theory

Classical test theory is arguably the most common theory of measurement (Gulliksen, 2013; Novick, 1965). Classical test theory is founded on the relationship between a true value of a given variable of interest and the value we observe when we attempt to measure it. In an ideal world the observed and true scores would be exactly the same, but in reality this is almost certainly not the case. This is an assumption (though we have good reason to believe it to be true) because the true value can never be observed. The measurements we make are assumed to be directly related to the truth with some degree of error, as described in Equation 1.3.

$$x_i = t_i + \epsilon \tag{1.3}$$

If the error term, ϵ, is assumed to be normally distributed and with some variance σ_ϵ, then σ_ϵ is the standard error of measurement of the scale. Classical test theory makes three important assumptions about the relationship between the observed value and the (unobserved) true value.

- Firstly, measurement error is drawn from the same distribution for every observation of t at the j^{th} time point and the same distribution for every i^{th} participant's measurement.
- Secondly, the measurement error is normally distributed and centred around

zero, indicating that the scale is not biased.

- Thirdly, all measurement items (if there are more than one) designed to measure the overall latent variable (such as questions in a scale designed to measure depression) have the same measurement error.

If all of these assumptions hold true, we can estimate the standard error of measurement by using estimates of the scale's reliability, or the degree to which the observed measurement varies if the true score remains constant. Reliability can be estimated in two ways.

1. **Test-retest reliability**

   If the first two assumptions hold true, then reliability of the total score for each participant can be used. The participants are measured using the desired scale, and then measured again a short while later. It is presumed that if a short enough time has passed then the true value will not have changed, so any change between the times is due to measurement error.

   Unfortunately in psychology our participants tend to change over time (Hinton-Bayre, Geffen, Geffen, McFarland, & Frijs, 1999). In the field of cognitive testing improvement is often expected even between the first and second test session as participants tend to learn and exhibit practice effects. If this is true, then this is evidence to suggest that the true value has changed with time, and so change between any two time points cannot be strictly attributed to either error or difference in true value, but rather a mixture of both.

2. **Internal consistency**

> If all three of the assumptions listed above hold true, then the standard error of measurement can be estimated using internal consistency. Participants are measured once with a desired scale that has multiple items that all point to t_i with error drawn from ϵ. As all observations from each participant are assumed to have the same ϵ, we can use the correlation between items to estimate error.
>
> Unfortunately these three assumptions do not always hold true in any real psychological testing because often multiple latent traits are being measured by the same scale, and these traits may all be measured with different error from different items. For example, the Depression, Anxiety and Stress Subscales (DASS; P. F. Lovibond & Lovibond, 1995) measures three different latent variables (namely depression, anxiety and stress), all of which are moderately correlated with each other ($r \approx .5 - .7$). One potential solution is to focus on a single sub-scale but factor analysis indicates that some items are more representative of the latent variable than others (Wong, Dahm, & Ponsford, 2013).

The importance of assumptions when calculating reliability is implicated both when estimating error and also when making inference. I will discuss the implications for estimating error further in this introduction, and these implications will be especially salient in chapter five, where we will focus on alternate methods for making inference when it is hard to estimate error. However, one already established alternative is to consider the possibility that there are different sources of error. This is the basis of generalisability theory, which I will briefly discuss in the next section.

1.2.2 Generalisability Theory

Whilst classical test theory assumes that error is drawn from a similar distribution across all possible variations of measurement (i.e. the link between the latent and observed variable is similar across test items, value of latent variable, time etc.), generalisability theory, subsequently referred to as 'G' theory, focuses on segmenting this error into defined sources (Cronbach, Rajaratnam, & Gleser, 1963). This method advocates for research designs that focus on estimating the sources of error and the size of their contributions first before aiming to answer questions of relationship (Shavelson, Webb, & Rowley, 1989).

This emphasis on measuring error first means that the purpose of many studies that are based upon 'G' theory is different to those based upon classical test theory. First, the major contributors to variation are identified (e.g., type of stimulus, time of day, person administering), and then future experiments are designed to maximise the sources of variance (i.e., use many stimuli, test at many different time points, multiple administers) to obtain more accurate estimates. Error is broken down into components that, as described in Equation 1.4, can interact.

$$X = T + \epsilon_{\text{TIME}} \times \epsilon_{\text{STIMULI}} \times \epsilon_{\text{ADMIN}} \times ... \times \epsilon_{\text{OTHER}} \tag{1.4}$$

This approach is useful for the design of experiments to maximise efficiency as an accurate sense of source of variation can aid the researcher to design experiments that best suit their purposes. The concept that error can have different forms is a common feature in many of the chapters throughout this thesis. 'G' Theory extends classical test theory by considering inference over smaller units than just aggregate scores. Item Response Theory (henceforth abbreviated IRT) extends this further by considering item-level effects specifically.

1.2.3 Item Response Theory

Item response theory considers how specific items measure a given latent construct. Whilst classical test theory and 'G' theory consider the relationship between measurement and error, IRT argues that different items *discriminate* between different levels of ability differently. For example, imagine that we are trying to measure mathematical ability. A very difficult question or item might be included in our test to effectively discriminate between an excellent student (who successfully answers the question) and a very good student (who fails to answer the question). However, this item would not discriminate well between the very good, good or poor students as they all would fail to answer the question.

The relationship between item difficulty and individual ability is typically modelled with a sinusoidal function (usually logit or probit). Figure 1.2 shows an example of this function for a single item, with ability (θ) plotted on the x-axis and the probability of success on the item plotted on the y-axis.

If we consider a single item, or a set of many items with relatively similar discrimination curves, we can see that this method violates the first assumption of classical test theory, that the error will be drawn i.i.d. for each item. The uncertainty in estimating the ability of an individual is dependent on the individual's ability. This implies that the error differs for different levels of ability.

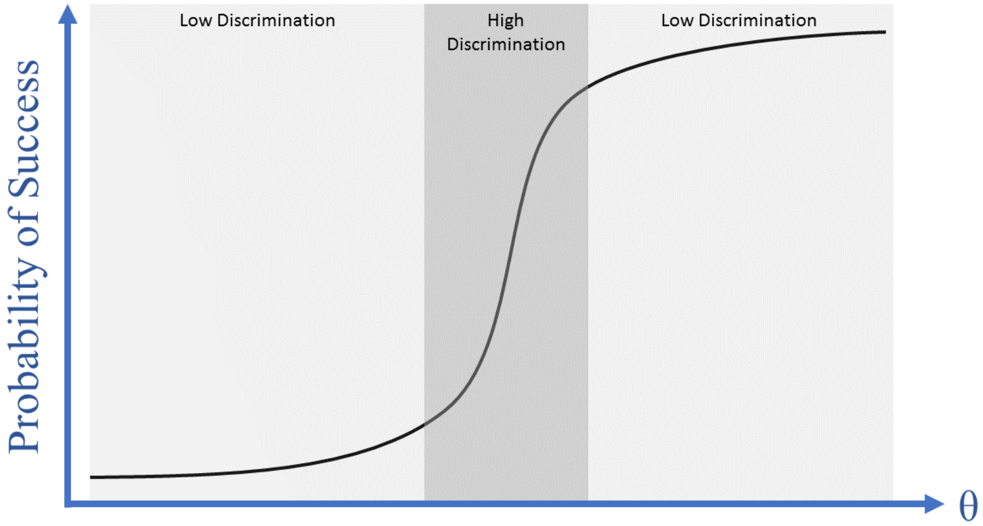


Figure 1.2: **Example of IRT model for single item.** This figure demonstrates the sinusoidal relationship between an individual's underlying ability (θ, x-axis) and their probability of success (y-axis) on an item. Note that the item discriminates well between some abilities (highlighted dark grey), whilst having limited discrimination in others (light grey).

Overall, it is clear that these three measurement theories differ in the assumptions they make about the way the observed measurements are transformed from the true scores of interest. These assumptions result in different understandings of what might otherwise be relatively similar datasets. It is an example of one of the themes underlying this thesis: that data generation processes impact the interpretations one can make from the data. The next section explores further why measurement theory matters.

1.2.4 Identifying Measurement Theory (and why it matters)

Most research papers that do not explicitly focus on measurement do not directly specify the overarching theory of how the items were measured. This might give the appearance that measurement theory does not impact experimental design, conclusions, or the statistics used to the connect the two. I argue, however, that this appearance is false.

The assumptions that we as researchers make about the way our data were generated *should* and *does* impact the way we design, analyse and draw conclusions from experiments. However, practicality dictates that these assumptions be designed to generalise across a wide range of situations, which limits the degree to which they accord with researcher assumptions. In addition, different theories are used with *different aspects* of the data so that the choice of theory relates to the desired outcome. This encompasses questions of measurement accuracy (classical test theory), error estimation (generalisability theory), and individual latent value estimation (item response theory).

For the reasons above, no measurement paradigm is inherently superior but instead must be based on a researcher's *aims*, *intents* and *beliefs*. The next section demonstrates how different measurement theories result in different interpretations of common statistical methods.

1.3 Researcher assumptions and statistics

So far we have seen how different measurement theories make different assumptions about the way data are generated. Here we discuss why this matters in terms of how experiments should be designed and analysed.

This section contains three important take-home points. First, the measurement theories differ in their *emphasis of error*. Second, the researcher's *intention* guides both their underlying measurement theory and the statistics chosen. Third, these both contribute to the *choice of* and *interpretation of* statistical methods. The first point becomes relevant to chapter five, which addresses how to identify individual change from variance due to measurement error. The second and third points relate to the issue of model complexity. In the main chapters I argue that the aspects of the data that the researcher is interested making inference about should impact the choice and interpretation of statistics. This idea is not novel in the abstract, but this thesis identifies several important contexts in which standard approaches fail to sufficiently incorporate this insight and proposes models that do. First, though, it is important to understand how measurement theory relates to assumptions about the data generation process.

1.3.1 Data generation assumptions

One of the major differences between the three measurement theories we discussed in the previous sections is that they all differ in the way error is hypothesised to be introduced to the observed scores. Not surprisingly, this difference results in differences in estimating error, which leads to differences in how experimental design and statistical methods are chosen to elicit error.

Classical test theory

Classical test theory assumes that the observed or measured value represents the *true* variable with some amount of unbiased error. If there are multiple measurements that are designed to represent the same true value (such as multiple items on a test that are designed to measure the same true variable), then the error is assumed to be of the same unbiased distribution but independently sampled for each item (i.e., independently but identically distributed; i.i.d.). Given these assumptions we can estimate error in two main ways; internal consistency and test-retest reliability.

The method for internal consistency follows as such. If the error for each item in a given test is i.i.d. drawn from the same normal distribution with mean of zero, and the underlying true score does not change throughout the test, then we can use the relationship between the items within a individual to estimate the error of measurement. Whilst Guttman (1945) first proposed a number of ways to do this, Cronbach (1951) popularised one of these methods, λ_3, calling it *internal consistency*, or α (Equation 1.5). This measure is an estimator of the mean of all possible split-half reliability estimates (Cortina, 1993).[2]

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_k \sigma_k^2}{\sigma_t^2}\right) \tag{1.5}$$

There are some caveats to the method though. Although some have claimed that

[2] Split-half reliability is the agreement between two random subsets of the items.

internal consistency is the lower bound of reliability (e.g., Novick & Lewis, 1966), McCrae, Kurtz, Yamagata, and Terracciano (2011) provide empirical evidence to support the dissonance between internal consistency and other forms of reliability. Additionally there is a trade-off between having a narrow focus of measurement and the size of Cronbach's alpha (e.g., as discussed in Streiner, 2003). The most obvious example is that tests with a very small number of items risk having an unnecessarily high α. This isn't always the case; tests with many items that are repetitive also risk this (McClelland, 1980).

Although Chronbach's alpha is the most common method of measuring internal consistency (e.g., as noted Zinbarg, Revelle, Yovel, & Li, 2005), it is not the only method of assessing internal consistency. Guttman (1945) discussed six methods that all stem from an underlying assumption of independence between participants, and independence between items within participant, combined with an assumption of an infinitely large population.

This independence assumption does not hold true if the true structure of the test is hierarchical rather than uni-dimensional. A hierarchical test structure indicates that the test items represent or 'load' to a set of latent variables (such as items in a intelligence test loading differently on the different facets of intelligence), but also to an overall latent variable. If this is the assumed structure of the test, McDonald (2011)'s ω_h is more appropriate than Cronbach's α (which is likely to overestimate the test reliability).

As an alternative to internal consistency, test-retest reliability is often used, and is found to be preferable in some cases (e.g., in the measurement of personality;

McCrae et al., 2011). To estimate measurement error using this method, participants are asked to complete a given questionnaire two or more times. If we assume that the underlying true value did not change for the participant between measurements and it is also true that measurement error is independent within participant and across time, the measurement error can be estimated through the correlation between the first and second scores, or *test-retest reliability.* There are a couple of notable phenomena that suggest that these assumptions do not necessarily hold true. It is typical to find a decrease in test-retest reliability over time (for example, comparing short vs long term in McCrae et al., 2011). Additionally, in some contexts change is expected (Dikmen, Heaton, Grant, & Temkin, 1999; Hinton-Bayre et al., 1999). We discuss some of the implications of this in chapter five.

In classical test theory, different assumptions about the data generation process results in different experimental designs (one versus multiple measurement points in time), which then impacts the statistical method chosen to calculate measurement error (internal consistency versus test-retest reliability). From this we see that researcher assumptions impact both experimental design and the statistical methods used to analyse it. But is this just a property of classical test theory? In the next sections we consider whether this holds true for 'G' theory and IRT as well.

Generalisability theory and error

Classical test theory focuses on estimating the error process in transforming some set of true scores to a set of observed scores. It requires that this process be i.i.d. $\sim N(0, \sigma)$ across people, items, time, etc. to accurately estimate what we will now call “population” measurement error. As we discussed in the previous section, these assumptions lead to vastly different experimental designs and methods. Generalisability theory (‘G’ theory) is not so very different, except this theory views error in context of the population.

‘G’ theory requires the researcher to explicitly state what their *intention* is (Cronbach et al., 1963). The researcher needs to have an understanding of the population or set of things that they wish to *generalize* to. Only with this understanding of their intent can they then calculate the expected error using an Analysis of Variance method (ANOVA).

For example, consider a researcher who wishes to estimate the ability of students to undertake an addition process. He provides them with a scenario where addition must take place (for example, “Julie has 10 lollies. Richard has 5 lollies. How many lollies do Julia and Richard have together?). To estimate the measurement error in this problem the researcher must make a choice. Is the priority to generalise to similar problems with different numbers? To problems with the same numbers but different protagonists in the scenario? To problems with the same problems and protagonists but with fruit instead of candy? This problem is similar to the example suggested by Cronbach et al. (1963) and highlights that the researcher must include their own intent in item selection procedures.

These considerations demonstrate that, like classical test theory, ‘G’ theory also requires researchers to make assumptions and express their intent. The next section demonstrates that the same is true of Item Response Theory.

Item Response Theory and error

In classical test theory we saw that the researcher must make assumptions about the underlying error generation process in order to select appropriate experimental designs and statistical methods. In generalisability theory we saw that the population that the researcher intends to generalise to must also be included in the experimental design to create an estimate of expected error. In this section we show that even for similar datasets the researcher's theory of how data are generated leads to a different *statistical methods* AND *intentions.*

Item Response Theory (IRT) proposes that there is an interaction between the discrimination of a given item and the underlying value of the trait being measured for a given individual. Put simply, whilst classical and (to an extent) 'G' theory assume that the error introduced by an item is drawn i.i.d. for each individual, IRT suggests that the error term varies as a function of an individual's underlying trait and the discrimination of the item (Embretson, 1996). This has two main impacts: first, if we have already estimated the discriminability and difficulty of a set of items, Computer Adaptive Testing (CAT) can be used as an experimental design. Secondly, even if the data is collected in a more traditional way it impacts the interpretation that one might make from it.

In general, computer adaptive testing works (although there are much more complex schemes, see Revuelta & Ponsoda, 1998, for a comparison) in the following way. First, a participant is asked a question with high discrimination at a starting level of ability (i.e., an "easy" question). If a participant gets this item correct, they progress to an item of high discrimination at a slightly higher level of ability (i.e., a slightly harder question). If they get a given item incorrect, they are asked a question with high discrimination at a slightly lower level of ability than the item they got incorrect (i.e., they get a slightly easier question). Because of this iterative design and the underlying assumption that the discrimination of each item can be well estimated, it is possible to produce an estimate of the participant's ability with a known standard error. It also allows the researcher to trade-off reducing the standard error of the estimate and the number of items a participant is required to complete.

However, even if the data is collected in a more traditional manner (i.e., a set of i participants all answer the same set of k items) rather than an adaptive design, item response theory can still be used to elicit the discrimination of the items and hence the error in each estimate. Whilst classical test theory is very useful at describing the scale properties as a whole, IRT allows the researcher to infer properties of individual items. From this inference, the researcher can then choose an *alternative experimental design* like CAT, employ alternatives to classical test theory that make different assumptions, and *interpret* these methods to make inference at the level of the individual or item, rather than the group level.

This section discusses the assumptions of three measurement theories to explore how these assumptions impact *experimental design*, *statistics chosen* and the way the researcher *interprets* them. Whilst measurement error in and of itself is not the central focus of this thesis, I chose to include this section for two purposes. Firstly, many of the concepts of measurement error in classical test theory become salient in chapter five (where we consider and propose an alternative to how individual change can be estimated) and chapter three(where we consider skewed data). Secondly, that the researcher *always* makes some inference about the generative process of obtaining data (implicitly or explicitly) and these assumptions impact the way data analysis is undertaken. The next section also demonstrates how data generation assumptions filter through to the *intent* and *interpretations* of higher-level statistical methods. This is important because although this thesis focuses on situations where the statistical assumptions are discordant with the researcher's assumptions, it is important to first understand where they accord.

1.3.2 Researcher intention impacts statistical method

The previous section showed that assumptions about how data are generated affect experimental design and which statistical methods are appropriate. Here we extend this idea further to explore the important role of researcher intention. As the list of potential statistical methods that we could choose to focus on is very large, I choose to highlight two common intentions. The first set are the methods that a researcher might use to estimate *relationships*, which we also consider in chapter three. The second set of methods are those that a researcher might use to estimate *differences between groups*, which we consider in chapters four and five. I focus on how the researcher's underlying data generation assumptions interact with their intentions in the choice of statistical method.

Inference with relationship-driven data

Consider a researcher with two continuous variables. They wish to infer how closely these two variables are related. Already the researcher's *intention* has placed some limitations on the type of analysis used. In this case simple linear regression is generally employed. Although this is a limitation from intention, it is also a limitation imposed by the researcher's *repertoire*. Simple linear regression is almost universally taught as a statistical method in psychology. More complex methods (such as non-linear functions) that make more, less or simply different assumptions about the bivariate relationship are much less common in both classes and research.

Linear regression takes the form of equation 1.6, where a set of predictors X_i are used to predict the outcome variable Y with some prediction error ϵ. In the case of simple linear regression, there is only one predictor X and equation 1.6 simplifies to equation 1.7. Traditionally linear regression assumes that there is no error associated with either the X or Y variables (Spearman, 1904a). Instead the error term represents the error in predicting Y from X.

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_n * X_n + \epsilon \tag{1.6}$$

$$Y = \beta_0 + \beta_1 * X + \epsilon \tag{1.7}$$

So far, other than the decision to investigate the relationship between the two variables, the researcher's intentions have not played a strong part. However, even within the regression analysis there are at least three different components of the analysis that could be used depending on what their intentions are.

First, the researcher could be interested in *prediction.* How well can we predict the outcome variable Y from X? More specifically, for a new observed value of X, x_i can we estimate the corresponding value of Y, y_i? To do this we can employ a prediction interval (Geisser, 1993). Any error in the measurement or observation of X or Y will artificially widen this interval.[3]

Second, the researcher could be interested in *decision making.* In this case the decision would be binary. Can we correctly identify whether X is significantly related to Y or not? If this is the researcher's intent, they have made assumptions that any error in generation (such as scale responding patterns) is independent between X and Y, if present at all.

Third, the researcher could be interested in estimating the *strength* of the relationship. In this case he (or she) is interested in estimating the percentage of the variance in Y that is explained by X. Again if this is the case, the researcher is making similar assumptions to the other aims listed previously, but because the researcher is interested in inferring a different aspect of the relationship, he (or she) uses a different part of the regression analysis.[4]

Different statistical approaches and considerations must be taken if the researcher wishes to *predict* new data, make *binary decisions* or estimate the *strength* of a relationship. For each of these intentions, different underlying assumptions about the underlying sampling procedure, error function and other factors that impact

[3] Provided that this error is not correlated between Y and X. If it is correlated, it could potentially artificially tighten this interval.

[4] In simple regression the second and third aims are similar mathematically. This is a special case. This property does not hold true for other regression-type analyses.

data generation are also made.

Estimating group differences

The important role of researcher intention is not only true for relationship-driven data; it also applies when researchers wish to estimate group differences. Typically when considering group level differences, we tend to use tests that focus on estimating the mean, like t-tests (e.g., paired, Welch, one-sample or students, and two-sample) and ANOVA (e.g., ANOVA, ANCOVA, MANOVA etc.)[5]. Importantly when the focus is on mean estimations, we rely on the mean being a good summary statistic for the distribution at hand. This generally implies that these analyses work best when the distributions are symmetrical, and the tails roughly approximate those of a normal distribution. If this condition holds true, then the research can make claims about the average scores in group A in relation to the average scores in group B.

However in some cases the researcher is not interested in inferring average difference between groups but instead interested in inferring whether the groups differ in terms of the dispersion of the data. Consider, as we do in chapter three, a scale designed to measure depression symptomology. The researcher may not want to compare the groups in terms of average score on this scale, but instead the proportion of individuals who score very highly on this scale. Here the researcher would be interested in analysing the tails of the distribution.

[5] These are, of course, mathematically equivalent to linear regression techniques. I choose to treat these two methods separately in this thesis as they are often used that way in practice.

Summary

In this section I briefly discussed two types of analyses: firstly, analyses that focus on inferring relationships and secondly, analyses that focus on quantifying group differences. For both of these families of analysis I described how the researcher's intentions could impact the *aspect* of the data the researcher wanted to make claims about, and hence the statistical methods employed. I pointed out that, like the estimation of error, these methods *all* make assumptions about how the data are generated (an idea that will be revisited throughout this thesis). However, at this stage, we have only covered how the researcher's assumptions and intentions influence statistical methods. In the next section we consider interpretation.

1.3.3 Interpretation, statistical method and assumptions combined

While the researcher's intention impacts both experimental design and choice of statistical methods, it also impacts how a statistical method is interpreted. As an illustration we consider the various ways the previous two statistical methods can be interpreted.

Simple linear regression

First let's consider the humble simple linear regression. We've already seen we can use it to estimate error and describe relationships. Here consider linear regression of two similar datasets. Both datasets contain two continuous variables that form a pair of observations from a single person.

However these two datasets differ in the way the data was collected. In test-retest reliability the first variable consists of total scores on a given scale, while the second variable consists of total scores on the same scale after some time delay. When the researcher aims to estimate the relationship between two variables, the two variables are measured simultaneously (or close to). The researcher's intention to either estimate test-retest reliability or strength of relationship has impacted the nature of their experimental design.

The researcher then makes a strong assumption about what the error term in equation 1.7 represents. If the researcher is interested in calculating the test-retest reliability, they assume that if the paired observation x_i and y_i are not equal, this difference is due to some *measurement error* ($\epsilon_{\text{MEASURE}}$). If the researcher is interested in estimating the strength of the relationship between the paired observation x_i and y_i, they assume this difference is due to some *prediction error* ($\epsilon_{\text{PREDICT}}$). This demonstrates that two identical analyses on (objectively) similar datasets can produce two completely dissimilar interpretations of the same parameter in the statistical model.

Analysis of variance

Now suppose that a researcher is not interested in doing a linear regression but instead wishes to make inferences about a 2x3 design. Suppose the researcher is analysing a trial where each participant is allocated to one of two different clinicians administering one of three different treatments. An ANOVA is conducted, and it appears that there is a significant effect of the researcher administering the treatment. How does the researcher's assumptions change their interpretations?

If the researcher views this study within the lens of 'G' theory, and their aim is to generalise across a wide range of clinicians (e.g., they are interested in the application of this treatment in a wider population), they might then choose to add many more administering clinicians in order to better estimate the effect of the treatment.

However, if the researcher views this difference as an underlying difference between the two clinicians, they might be interested in identifying why this difference occurs. With this in mind they might conduct post-hoc analyses to identify when and why these difference occur (e.g., is the effect the same for all participants?) and potentially future experiments varying clinician characteristics.

Summary

Through examples of linear regression and ANOVA, I have shown that the *assumptions* and *aims* of the researcher effect the interpretation of statistical methods. This concludes the first half of this introduction, which demonstrated that statistical methods are *always* linked to the researcher. It is linked to their underlying *assumptions* of how the data were generated, to their *intentions* with analysing the data and to the way they *interpret* these analyses. The main body of my thesis considers four different common occasions where the underlying assumptions of the researcher are at odds with the statistical tools that they are using. In each occasion we argue why we believe this to be the case, what the reasons behind this are and propose alternatives that directly link to the underlying assumptions.

In the remaining pages of this introduction I discuss the four assumptions underlying each of the four chapters and conclude with an overview of the contributions this thesis makes to the greater body of research as a whole.

1.4 Statistical assumptions that rarely hold true

Each chapter incorporates a different assumption, but all of the assumptions conform to three main criteria. Firstly, they are chosen because they were relatively common practises or occurrences in psychological data analysis. Secondly, the current methods usually used in each do not match the underlying assumptions of the researcher. Thirdly, the solution in two of the four examples involve a relatively simple Bayesian mixture model.

1.4.1 Assumption 1: Distributional assumptions

Chapter three considers the dissonance between the assumptions the statistical method makes about the underlying distribution (e.g., normality) and the actual sample distribution of the data. This chapter explores data where clinical-type symptoms were measured in a non-clinical sample. This data is skewed, discrete and bounded, yet a literature review suggests that the most common analyses are t-tests and correlation type analyses. While the chapter focuses on a specific application, here we focus more broadly on where in the data generation process skew can occur.

How is skew introduced into a sample?

There are two potential ways that a sample can be observed as skewed. The most obvious (and hopefully the most likely) reason is that the underlying true distribution is skewed. This is the most desirable as it suggests that the process of measurement does not distort the overall shape of the true distribution. If the measurement error was truly random, i.i.d. sampled from $\sim N(0, \sigma_{\text{SEM}})$ and the measurement error is relatively small, then the process of measurement would not bias the overall asymmetry of the true score distribution.[6]

The second is much less likely. In this scenario the error term is both very skewed and contributes a very large amount of variation to the observed score. If both these are true then even if underlying true distribution is normally distributed, then the process of measurement will introduce a skew to the observed distribution. We hope that this is less likely because it implies the measurement is drastically biasing the underlying distribution. If this is the case, it is very difficult to differentiate between any skew in the true distribution and skew introduced by the measurement error.

Whilst the first way relates very closely to the assumptions of classical test theory, the second assumption relates much more closely to the assumptions of item response theory, when error and the underlying true value interact. In this method a scale may be roughly normally distributed for some samples (depending on their underlying true value), but very skewed in others. A good example of this is the previously introduced example of measuring mathematical ability in 10 year olds and university

[6] As the variance of the measurement error grows large it would begin to impact the width of the observed distribution, and eventually swamp the underlying population distribution shape. If there is enough measurement error to change the distribution shape, then there are probably larger issues with the data than skew.

graduates. A normally distributed scale for the university graduates might appear very skewed in the 10 year old sample, even though their true ability is roughly normally distributed.

It is clear that skew can be introduced into data in a variety of ways. What implications does this have for how such data can be interpreted?

One implication is that the mean may no longer be an appropriate estimator. Furthermore, any credible (or confidence) intervals drawn around such an estimate that assume normality would be wider than necessary to account for the longer tail on one side of the distribution (potentially allowing for a large portion of the interval to contain values that are highly unlikely, as we see in chapter four). Traditional t-tests produce intervals that are symmetric around the estimate, which is inappropriate for skewed distributions. The increased width of the intervals can reduce the power of the analyses, impacting sample size calculations.

By contrast, if the error distribution is skewed then any distortion introduced by measurement error will need to be accounted for. When the error is normally distributed with a mean of zero then this is relatively easy to do. The error of the measurement can be estimated through estimates of the reliability of scores (either internal consistency or test-retest reliability). When the error is not normally distributed this is much more difficult. The issue is compounded if the researcher has some belief that the underlying true distribution is also skewed.

If measurement error is related to the underlying true scores in some way, this violates assumptions of independence present in linear regression. As we have already discussed, linear regression models assume that the error term is *independent* and identically distributed. We previously discussed that because measurement error is not directly modelled in a simple linear regression model, any measurement is accounted for by the regression residuals. If we have a variable where the measurement error is related in some way to the underlying true scores, then a regression model that includes this variable may have its assumptions of homoskedasticity violated. Loss of homoskedasticity can have drastic implications on the regression analysis.

1.4.2 Assumption 2: Is the sample contaminated?

Measurement theories all have a method to connect the observed measurement with true scores. What the theories do not consider is the potential for contamination of the sample, where some portion of the sample is not representative of the variable of interest at all. In chapter four, I develop a model that accounts for contamination, but first how does contamination occur, and how would we know if it does?

How does contamination occur?

One way to see contamination is as a function of what is perhaps an unwritten rule of psychological research. Whatever aspect of the human mind you are interested in, there will always be some portion of the sample you select that will be responding in a way that is completely contrary to the way you expected them to, but consistent with some unexpected interpretation of the task.[7]

[7] Although this is not included in this thesis, I began this PhD with a many-armed bandit task experiment. The results were not particularly interesting, partly because a proportion of my sample had begun to optimise their mouse movement during the task to minimise the amount of movement required between trials. They chose the bandit arm that was closest to the 'next' button that still had an acceptable pay-off, even optimising for slight bias to right-left mouse movements compared to left-right

There are two main ways this type of error can be introduced into the data. The first is through an experimental design that allows some proportion of the sample be from some population that is not the population of interest (e.g., interested in the lexicon of native English speakers, but some non-native speakers also participate). The second is through errors in the experimental manipulation (e.g., interested in learning how participants make decisions given different environments, but some proportion of participants choose to randomly select options).

The methods we use in chapter four can be used for both samples; if the *sample* contains contamination (such as calculating scientific ability of high school students where a certain proportion had their parents complete their project for them) and if the *process* contains contamination (e.g., when measuring the number of depression symptoms experienced in the last two weeks, and a certain proportion reporting the maximum number of symptoms they have ever experienced).

How do we identify it?

Consider a relatively common sample size of 125. If the distribution has a relatively high skew, we would expect the majority of the data to be close together, with a few datum in the long tail (see the second panel in Figure 1.3). With such a small sample size, we might have trouble estimating the heaviness of the tail of this distribution (Hoaglin, Mosteller, & Turkey, 1985). Now consider the first panel of Figure 1.3. Here we have simulated a distribution where the majority (96%) of the data is drawn from a standard normal distribution and a very small proportion (4%)is drawn from a contaminate distribution ($N(10, 1)$). Notice that the two distributions look

mouse movements. Needless to say, we did not expect that!

relatively similar.

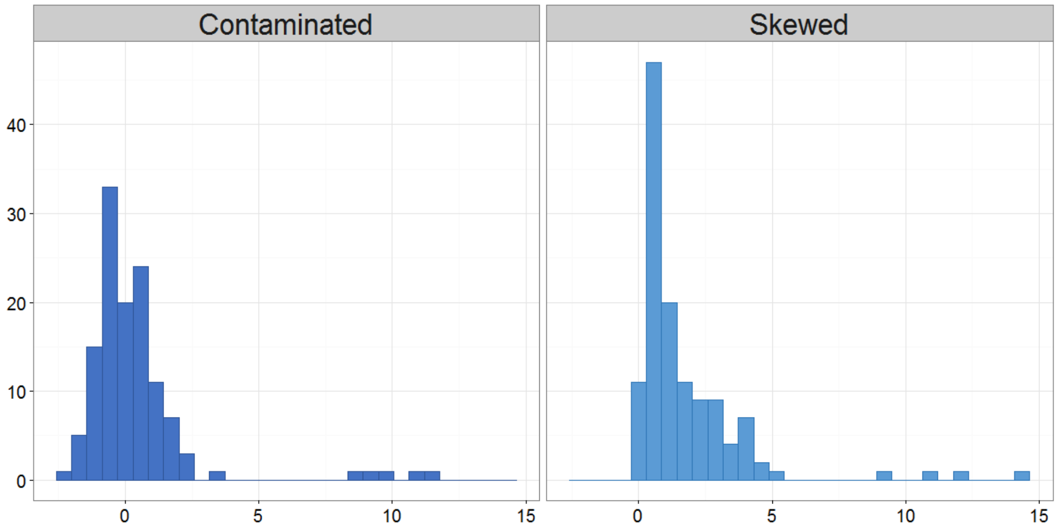


Figure 1.3: **Skewed and contaminated data comparison.** This figure demonstrates the similarity in distribution shape with a contaminated and skewed distribution. The contaminated distribution(*dark blue*) contains 120 random samples from a standard normal distribution, and 5 random samples from a $N(10, 1)$ distribution. The skewed distribution contains 125 samples from a $lnN(0, 1)$ (*light blue*).

While in chapter three we consider the impact of distributional assumptions, in chapter four we consider the impact of contamination when calculating the mean. In both chapters we propose models that can account for these assumption violations. Chapter three proposes a model that provides good mean estimates for a skewed normal distribution, while chapter four uses a Bayesian mixture model to identify and discount points that are potential contaminants that that influence the estimate of the mean. The difference between the two models is when estimating measures of scale. Here the two models diverge; the skewed model estimates scale as reflected by the full width of the distribution, but the contaminated normal model only estimates the width of the target distribution.

Naturally with two models that agree on estimating one parameter but diverge when estimating another, we should question which is the most appropriate to use. It does however make these assumptions more explicit, which we discuss in later chapters. First, we turn our attention to another pressing assumption. If we are willing to consider heterogeneity of sample and response in simple estimation of the mean, could we consider heterogeniety in response? The following section addresses just this.

1.4.3 Assumption 3: Is the response homogeneous?

Chapter five considers an experimental design that compares an experimental manipulation against a control condition. If we are willing to entertain that each condition might result in a clustered effect (i.e., not all participants respond in the same manner), then we must consider whether we should expect the proportion of individuals in each condition to be similar. If they *are* different, wouldn't the proportion of individuals who improve be as important as the size of the improvement?

The concept of considering different proportions of participants is not a new idea. In the clinical psychology literature, a tool known as Reliable Change Indices (RCI; Jacobson & Truax, 1991) and other variations such as the Wyrwich Standardized Difference (WSD; Wyrwich, 2004) and the Hageman Arrindell (HA; Hageman & Arrindell, 1999) have been used. These methods identify individuals who have experienced a significant change but do not the easily compare the proportion of individuals experiencing an effect. All of these methods require an accurate measure of the standard error of measurement (or the error in linking the observed to the latent variables).

These methods compare the change experienced by any given individual to the standard error of measurement. They all aim to do two things: First, to identify the amount of variability or the *standard error of measurement* that should be expected if no change occurred; and second, to compare the amount of change experienced by the individual of interest to decide whether this was significantly different or not.

The Bayesian mixture model introduced in chapter five, unlike existing methods, does not require a direct estimate of variability, but instead compares the relative difference in the proportion and size of effect between two conditions.

One aspect of experimental design that we do not explicitly consider in this section, but is very common in trials that are interested in measuring reliable change, is accounting for missing data. The next section addresses this issue, with a focus not on the statistical model involved but on evaluating people's intuitions in this context.

1.4.4 Assumption 4: Is there omitted data?

Previous empirical research suggests that humans make complex and perhaps unintuitive decisions when there is missing data. Researchers are humans, and humans demonstrate a number of biases of numerical reasoning (e.g., Ellsberg, 1961; Epley & Gilovich, 2006; Landy, Silbert, & Goldin, 2013). Given that the central message of this thesis is that researcher assumptions can and should be incorporated into statistical models, is non-optimal human cognition a problem?

In chapter six I address this question, presenting original research investigating how people make inferences given missing data. It argues that since we cannot avoid assumptions entirely, it is important to be cautious and explicit about which assumptions we make. Moreover, the experiment demonstrates that although people have certain reasoning biases, in many cases such biases are sensitive to the underlying data generation process in an interesting way.

Statisticians draw the distinction between three different types of missing data; missing *completely at random*, missing *at random* and missing with some *underlying process* (i.e., missing not at random). Importantly the choice of which process is most appropriate for the researcher to model lies with the researcher, and their assumptions about the data generative process. To illustrate the difference between these three terms, I provide the following example.

Consider a lecturer who wishes to obtain a sense of their quality of teaching. To do so they send an email to all students inviting them to special lecture. Following the lecture the students are invited to fill out a survey rating the lecturer's efficacy. Some data were missing in the survey. Conceptually there are three different ways this data could be theorised to be missing.

Missing completely at random refers to occasions where the data are missing with no relationship to either the variable of interest or any other part of the dataset. In our example some of the surveys may have inadvertently had coffee spilled over them, rendering them unreadable.

Missing at random refers to occasions where the data are missing with no *direct* relationship to the variable of interest but are related to other aspects of the dataset. In our example this could be occur through some students using a certain email client that is particularly aggressive in identifying junk email (reducing the probability of participation). The students never realised there was a lecture and so never attended to fill out the survey, but the choice of email client is unlikely to be related to the student's opinion of the lecturer.

Lastly the missing data can be assumed to be missing due to some underlying process. In our example if the students already find the lecturer unpleasant in some way (e.g., boring, ineffectual etc.) they may choose not to waste their time by attending an optional lecture by the teacher in question and so the sample will under-represent individuals who dislike the lecturer, which inflates positive scores.

Chapter six reports on an experiment that investigates the inferences individuals make when faced with missing data that is ambiguous in nature. We show that despite having a relatively good grasp on the underlying distributional information, individuals still tend to favour decisions that did not match the information they had been given.

Although is a slightly different approach to the rest of this thesis, I include it as evidence that researcher assumptions and inference may not necessarily be as straight-forward as made out in chapters three to five. Taken together, one implication may be that the process of choosing more appropriate models (as in chapter three through five) may help researchers to clarify and communicate their underlying assumptions.

1.5 Contributions of this thesis

In previous sections I identified and described how the researcher's *assumptions*, *intent* and desired *interpretation* drive the statistics used by that researcher. I described four different occasions where the *researchers* and *statistical* assumptions currently do not match, and described why this is not desirable. I then explored how changing the basic assumptions might require more complex models, and how this benefits the researcher by allowing them to directly make claims about certain aspects of the data that are important (even if no additional variance is explained), and also make explicit implicit assumptions about the data. This thesis uniquely contributes to this argument in three main ways.

- Firstly, I use key modelling features taken from the cognitive science domains. Models should be based and explored upon their underlying *representation* of the data as well as the way they fit and predict data. I use this concept within the broader field of data analysis in a psychological context.

- Secondly, I achieve this through the suggestion of alternate models that directly achieve the underlying goals central to common data analysis but are currently not achievable (e.g., the removal of outliers before calculating the mean of a distribution and the allowance for heterogeneity in response to an intervention). To the best of our knowledge, whilst the underlying traits of these models are not unique (e.g., Bayesian hierarchical and mixture models are well known), these applications are unique.

- Lastly, I also take considerable pains to explore and demonstrate the reliabil-

> ity of these models in a variety of situations. Unique models are relatively easy to propose in the current climate of Bayesian modelling through MCMC sampling. For each new model I take significant space to demonstrate its properties in terms of power, Type 1 error and other important properties of interest to the researcher. In some cases I also suggest extensions to the models that may be possible, but do not have space to devote to such an in depth analysis.

It is my hope that these three core contributions will help the wider field of psychological research (and others who collect, collate and analyse data from people) to connect closer to some of the core strengths of the field of cognitive science. By modelling the process and incorporating experience that the researcher brings to data analysis it is my hope that data analysis will have greater finesse and allow for stronger, and perhaps more importantly *different* claims to be made about the data.

1.6 Outline of approach

In chapter two I outline and defend the statistical methods that are proposed in this thesis. I will also outline the metrics by which the models are compared.

To address the hypotheses in chapter three I will demonstrate that removing (violated) distributional assumptions can increase the power and false acceptance rate when considering skewed distributions.

Although this suggests that statistical analysis should move towards making fewer assumptions about the data, we show in chapter four that this is not the case. By adding in a theory of contamination as part of the data analysis I show that we can still achieve good accuracy whilst adding in a greater richness of the conclusion that can be made.

Chapter four focuses on the use of Bayesian mixture models where the data are assumed to contain some proportion of contamination that impacts our ability to make inference about the expected value. Chapter five demonstrates that the same family of data analysis techniques can be applied in a completely different context to make different but equally rich claims about individual differences in outcomes in randomised control trials.

In chapter six I explore a cautionary example in a study that asks participants to make a judgement given missing information. Researchers often have to make a judgement about their believed underlying generative process that results in missing data, this chapter demonstrates why these judgements should be explicit.

Chapter 2

Methodology

> Torture numbers, and they'll confess to anything.
>
> *Gregg Easterbrook*

So far I have discussed the basic goals of this thesis, namely investigating the researcher's assumptions, intent and interpretation, and connecting these to modelling approaches. Before we can investigate this, first I will discuss how these models are to be expressed, estimated, evaluated and compared.

The overarching message of this chapter is to provide assurance that the estimation and model analysis methods that I use throughout the forthcoming chapters are not, as Gregg Easterbrook would say, simply torturing numbers to tell a particular story. Instead I aim to demonstrate the principled manner in which the models that I wished to compare were constructed, fit to the data and evaluated.

I outline four main aspects of statistical methodology that are relevant: First, the difference between Bayesian and frequentist modelling approaches, focusing primarily on the underlying differences in underlying probability theory and interpretation. Second, how different parameters within the model are estimated, weighing up the differences between explicit and several sampling-based methods. Third, assessing the estimations produced, focusing on assumption checking and methods of assessing approximate Gibbs sampling methods. Given the accuracy of the estimation procedure, I then consider the accuracy of the model itself through Type 1 and Type 2 error. Finally I summarise the outcomes of these four section to provide a justification of the statistical tools used throughout this thesis.

Before this, I first wish to preface this chapter. The previous chapter contained many examples of common frequentist-style tests. They were chosen to illustrate how common and popular statistical models make strong assumptions about the data. Since frequentist-style statistics are still very common, frequentist statistics were chosen as examples. It is important to note that Bayesian alternatives can be produced that still preserve the same model structure and underlying assumptions (for example in Morey & Rouder, 2015). This thesis is *not* designed to be an example of Bayesian statistics proving to be better than frequentist. My interest lies with the interaction between model and researcher assumptions. In this sense I take the view of Bayarri and Berger (2004) that the most convenient method should be chosen. With this in mind we proceed to the first consideration I made in the methods produced in this thesis: The choice of Bayesian or frequentist statistical methods.

2.1 Bayesian and frequentist statistics: The underlying differences

In recent years statistics, especially statistics in the psychological and cognitive literature, have broken into two rather distinct groups: Bayesian (e.g., Kruschke, 2010; Navarro, Griffiths, Steyvers, & Lee, 2006; Rouder & Morey, 2012; Wagenmakers, 2007) and frequentist (e.g., Cumming, 2008; Smithson & Shou, 2017), with a considerable amount of discussion of the relative merits of each (e.g., Efron, 1986; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008). Despite each side making strong arguments in favour of the efficacy of their chosen method, they differ in underlying probabilistic theory that the statistical inference is founded upon.

In practice there is also a divide between the most commonly used estimation method. In the frequentist framework, analytic solutions (such as least squares estimation for regression) dominate, with simulated solutions (such as bootstrapping procedures) less prevalent. In Bayesian literature, sampling and approximate methods dominate with more complicated models, whereas simpler models tend to be estimated analytically. The next two sections address these two differences in turn.

2.1.1 Differences in Probability

The two main statistical methods differ in the probabilistic theory that underpins them. This difference would perhaps not be as salient and important to researchers if it did not also results in differences in the interpretation of the statistical method as well.

First we will consider frequentist inference, which could be considered the more traditional statistical method in psychological data analysis. The concept of frequentist inference was developed by Neyman (Neyman, 1937), Pearson (Pearson, 1920) and Fisher (Fisher, 1925). Fisher and Neyman-Pearson both based their work off classical probability theory, but had strongly opposing views on the interpretations. Fisher was concerned with issues estimating evidence. He suggested a statistical methodology where findings were quantified in terms of the likelihood of obtaining a result as rare or rarer if a null hypothesis (H_0) was true.

The Neyman-Pearson interpretation was concerned with the balance between the control of Type 1 error, α (the probability that the null hypothesis was falsely rejected) and Type 2 error, β (the probability of incorrectly retaining null hypothesis). This interpretation introduces a second, alternative hypothesis H_A. The set of null and alternative hypotheses encompass the full hypothesis space, for example, the following null H_0 and alternative H_A to test whether the population mean is equal to zero.

$$\begin{aligned} H_0 &: \mu = 0 \\ H_A &: \mu \neq 0 \end{aligned} \tag{2.1}$$

The power of the method can be derived easily from the Type 2 error rate as $1 - \beta$, the probability of correctly rejecting the null hypothesis. These two methods were subsequently integrated within the statistical literature to obtain a method that neither Neyman nor Pearson nor Fisher would be comfortable using, but is the prevailing interpretation (Hubbard & Bayarri, 2003).

Frequentist inference is based on classical probability theory, or the probability of an event occurring in a sequence of opportunities where it is possible it could occur (i.e., number of times a coin flip returns heads). Bayesian inference relies on the estimation of the probability of an event given some prior. This subjective definition allows us to quantify single occurrence events, like the probability it will rain tomorrow. Whilst frequentist inference is concerned with the quantification of the data given a null hypothesis, i.e., $p(x|H_0)$, Bayesian inference is concerned with the quantification of *any* hypothesis, H, given the data observed (i.e., $p(H|x)$). This distinction allows one of the largest benefits of Bayesian statistics, the ability to quantify evidence for the null (Wagenmakers et al., 2008).

The quantification of $p(H|x)$ was first introduced by Laplace and popularised by Bayes with Bayes theorem. This theorem (Equation 2.2) describes the relationship between the conditional probability of event A given event B. The benefit is that it can be used to estimate the probability of a hypothesis given some observed data, as in Equation 2.3. In this equation, $p(H|x)$ is referred to as the *posterior*, $p(H)$ the *prior*, and $p(x|H)$ the *likelihood*. The posterior describes the updated probability of the hypothesis H from the prior probability of the hypothesis $p(H)$ given the observed data x and the likelihood of observing the data x given the hypothesis H, i.e., $p(x|H)$. $p(x)$ is the a normalising constant, and is the probability of x given all hypotheses H. In Equation 2.4, this requires a complicated integral which is not trivial to estimate.

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{2.2}$$

$$p(H|x) = \frac{p(x|H)p(H)}{p(x)} \tag{2.3}$$

$$p(x) = \int p(x|H)p(H)dH \tag{2.4}$$

Generally the hypothesis H relates to testing a set of parameters, denoted $\boldsymbol{\theta}$ such that $\boldsymbol{\theta} = \{\theta_1, \theta_2, ...\theta_n\}$. If this is the case then the problem at hand would relate to estimating the probability of these parameters $\boldsymbol{\theta}$ given some set of observed data x. Formally this means that we would substitute $\boldsymbol{\theta}$ for the generic H in Equation 2.2 to produce Equation 2.5.

$$P(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{x})} \tag{2.5}$$

How do the underlying differences between Bayesian and freqentist approaches in how probability is defined lead to differences in interpretation?

2.1.2 Differences in Interpretation

Perhaps unsurprisingly, when two methods differ dramatically in their underlying theory, the interpretation of these methods is also different.

First we consider the frequentist viewpoint, which encompasses both the Fisherian and Neyman-Person approaches. The Fisher viewpoint suggests that the observed data x should be interpreted in terms of how likely it would be to observe this data given the null hypothesis H_0. Under this view, a p value can be interpreted as the probability that we would observe data as rare as what we had observed or rarer given the null is true.

Unlike the Fisherian view, the Neumann-Pearson weighs evidence for a competing alternative hypothesis H_A. Although we cannot formalise evidence for the null hypothesis H_0, we reject the null hypothesis in favour of the alternative hypothesis if there is sufficient evidence that the null is unlikely. Here we can view the p value in as a long-run probability. If the null is true and a 95% confidence interval is drawn to reflect uncertainty about our estimate of a given parameter of interest, we would expect that 95% of all confidence intervals drawn would contain the true parameter of interest. In both the Fisherian and Neumann-Pearson views, the evidence is viewed as the likelihood of a data x given some null hypothesis H_0 (i.e, $p(x|H_0)$).

Bayesian methods have an alternative interpretation because they estimate the probability of a given hypothesis given x directly (i.e, $p(H|x)$). If we focus on parameter estimation, then a credible interval can be formed that reflects uncertainty on the estimation of a given parameter. While the traditional 95%-confidence interval can be interpreted in that 95% of all 95%-confidence intervals contain the true population parameter, a 95%-credible interval can be interpreted as containing the true parameter with 95% confidence. If we consider hypothesis testing, the Bayesian approach allows for the estimation of the probability of the null hypothesis directly, which frequentist methods do not.

Furthermore, if we wish to compare which of two models is more likely given the data at hand, the frequentist would calculate the likelihood ratio (Equation: 2.6) and interpret it as the ratio of likelihoods of observing the data given hypothesis 1 and 2. The Bayesian would calculate a Bayes Factor (Equation: 2.7), and *even if* $p(H_1) = p(H_2)$ and so the formula was exactly Equation 2.6, would interpret the ratio as the ratio of the conditional probability of hypothesis 1 and hypothesis 2 given the observed data.

$$\frac{p(x|H_1)}{p(x|H_2} \tag{2.6}$$

$$\frac{p(H_1|x)}{p(H_2|x)} = \frac{p(x|H_1) * p(H_1)}{p(x|H_2 * p(H_2)} \tag{2.7}$$

Rather than adopting one approach or the other – only frequentist or only Bayesian – this thesis takes a toolbox approach. While in chapter three we consider mostly frequentist models, in chapters four and five I propose Bayesian mixture models as potential solutions. I am primarily concerned with finding the most appropriate model, and so attempted not to be bound too tightly in idealistic notions of the "true" probability theory. However, the differences between interpretation make it difficult to directly compare frequentist and Bayesian methods. This means that I am careful throughout this thesis to compare frequentist with frequentist methods (chapter three) and Bayesian with Bayesian methods (chapters four and five).

To do this comparison, we must be able to estimate the parameters of the model, or 'fit' the model to the data we have observed, as discussed in the next section. There we will again see differences between frequentists and Bayesians in the methods of model fitting.

That being said, there are a number of intuitive benefits to Bayesian methods (Wagenmakers, 2007). These benefits result in a general preferences for Bayesian methods in this thesis (e.g., see chapters four and five). One benefit that is displayed in chapter four is the creation of intuitive credible intervals, also discussed in the next section. These intervals do not reflect long range probabilities but rather the confidence we have in our estimate interval.

2.2 Estimating parameters and uncertainty

Statistical estimation broadly falls into two main aims: Firstly, to estimate parameters from a limited subset of observed data, and secondly, to compare different hypotheses e.g., comparing H_0 to H_A. Here I focus primarily on *parameter estimation*, with little attention paid to comparing the likelihood of different models. Later in this chapter, when I argue for comparing models by the reliability and accuracy of decisions made from the data, I use an approach similar to Kruschke (2010) and Cumming (2008).

Before discussing parameter estimation, we must first consider how we will describe the parameter of interest. There are two types of information we will consider. The first is a single point estimate that is most representative of the parameter of interest (e.g, the sample mean, $\bar{X}$, estimating the population mean. μ). The second is an interval that represents the uncertainty we have about that population estimate (e.g., a credible or confidence interval). These two types of information are either estimated directly through explicit likelihood functions, or through a sampling procedure.

2.2.1 Explicit

Initially statistical methods were constrained by the limits of computational power. This meant that many of the oldest methods such as correlation (Galton, 1886; Pearson, 1895), student's t-test (Student, 1908) and linear regression (Galton, 1894) had analytic or explicit solutions. Explicit solutions involve deriving a symbolic solution to a general problem given a set of assumptions. This large initial time cost is offset by the generality and speed of the computation after an explicit solution has been found.

Throughout this section we will consider estimating a relatively simple statistic, the population mean μ, which we continue to attempt to estimate throughout this thesis. If the sample is independently drawn from a normal distribution, then the mean of this distribution can be approximated with the sample mean $\bar{X}$. We can also create a 95%-confidence interval[1] that represents uncertainity in this estimate. Given the assumption of normality, the formula for this confidence interval only relies on the sample mean $\bar{X}$, the sample standard deviation sd and the quantile of a t-distribution that corresponds to an 95% level confidence interval (Equation 2.8). The sample mean, sample standard deviation and size of the sample are quick to compute, and there are tables of the relevant quantiles for the t-distribution with different degrees of freedom, which makes this equation very quick to compute.

[1] It is trivial to create an interval that contains the true mean $\alpha\%$ of the time, but 95% is the most common.

$$\left(\bar{X} - t_{df,\frac{\alpha}{2}} \frac{sd}{\sqrt{n}}, \bar{X} + t_{df\frac{\alpha}{2}} \frac{sd}{\sqrt{n}}\right) \tag{2.8}$$

This equation relies on the central limit theorem, which states that for a set of samples $S = \{s_1, s_2, s_3, ...s_k\}$, drawn i.i.d. from some distribution with mean μ, the distribution of means calculated from this set of samples ($\bar{S} = \{\bar{s_1}, \bar{s_2}, \bar{s_3}, ..., \bar{s_k}\}$) will be normally distributed with mean μ provided that k is sufficiently large. This finding is independent of the shape of the distribution from which the sample is drawn.

However, the mean of the distribution may not be the most desirable summary statistic if the distribution is not symmetric. Other estimators might be used, but these estimators add additional complexity to the explicit solution, making the explicit solution difficult if not impossible to find. Explicit solutions are not always tractable and generally quite hard to find. Relying on models that do have explicit solutions results in the use of models which make very strong assumptions about the population distribution, and limits what estimators can be used. In this thesis we only consider explicit solutions in chapter three, whereas in chapters four and five employ sampling solutions.

2.2.2 Sampled solutions

While explicit methods expend time to *create* a solution and then capitalise on this time with when the solution is *calculated*, sampling based methods have the opposite focus. It is relatively easy to define a model, complete with underlying distributional assumptions and assumptions about the relationship between parameters[2]. However, these models can take a much longer time to estimate parameters in the model because a procedure of sampling must be undertaken. This thesis incorporates two methods of sampling: bootstrap and Markov-chain Monte Carlo (MCMC). Note that the choice of solution method is not dependent on the underlying statistical approach. Both Bayesians and frequentist can and do use both, with preferences for different types of approaches.

Bootstrap

This method was first developed by Efron (1992). We first consider this method in a frequentist sense. Say we have N observations of a variable $\boldsymbol{x}$, where the i^{th} variable is denoted x_i such for N observations of x_i, $\boldsymbol{X}$ forms a set such that $\boldsymbol{X} = \{x_i, x_2, ..., x_N\}$. From this sample we would like to make some inference a certain statistic about the population from which the sample was drawn. For simplicity we will again discuss estimating the population mean μ from the sample mean $\bar{\boldsymbol{X}}$. The simplest bootstrapping process takes a sample $\boldsymbol{X_s}$ of size N from $\boldsymbol{X}$ with replacement and calculates the mean for this sample, $\bar{\boldsymbol{X}}_{\boldsymbol{s}}$. This sampling procedure is completed B times, where B is generally very large ($B > 1000$). If the sample is representative of the population, then this set of sample means $\bar{\boldsymbol{X}}_{\boldsymbol{B}} = \{\bar{X}_{s1}, \bar{X}_{s2}, ..., \bar{X}_{sB}\}$ can be used to make inference about the population mean μ.

Like with the analytic method, bootstrapping uses the sample mean $\bar{X}$ as an es-

[2] It is, as we will see, not trivial to demonstrate that the parameter estimates are trustworthy or the model suitable. We discuss this in the next section.

timate for the population mean μ. However, rather than relying on an analytic solution for the 95%-confidence interval, we can use the bootstrapped distribution to estimate an interval that, with 95% confidence contains the population mean μ. As we see in equation 2.9, this does not rely on the sample standard deviation or t-distribution quantiles, but instead the $\left(\frac{1-.95}{2}\right)$ and $\left(1-\frac{1-.95}{2}\right)$ quantiles of the bootstrap distribution.

$$\left(\bar{X}_{b,.025}, \bar{X}_{b,.975}\right) \tag{2.9}$$

The Bayesian approach can also utilise bootstrapping methods, although they are somewhat less common. In chapter four we discuss Bayesian bootstrapping, as first proposed by Rubin et al. (1981). For simpler Bayesian models Efron (2012) suggests that parametric bootstrapping might also be suitable. Parametric bootstrapping uses the sample estimate parameters that describe the population distribution (i.e., if the population is normal, the mean μ and variance σ^2 is estimated using the sample mean $\bar{X}$ and variance sd^2). We would then generate B samples from this distribution and calculate the sample mean $\bar{X}_b$. From this set of sample means, we calculate our interval as in the more traditional bootstrap method.

Whilst we will consider bootstrapping in a Bayesian sense briefly in chapter four, it is not appropriate for many of the more complex models that we will consider in other chapters. For these models we will need to consider an alternate sampling strategy.

Markov-chain Monte Carlo

Markov-chain Monte Carlo (MCMC) methods are a class of simulation procedures used to sample from a distribution of interest. They can be considered an extension of a random walk, where each new proposal point is dependent only on the last accepted point. This feature comes from the Markov-chain part of the algorithm. The Monte Carlo part refers to the process of generating suitable proposal points in order to numerically estimate something that is analytically difficult to calculate.

Why are MCMC methods used so extensively throughout Bayesian statistics? The primary reason for their use in this thesis simply relates to the complexity of the problem at hand. In equation 2.3, we saw the probability of the model H given the data x is equal to an equation of which the denominator is $p(x)$. This denominator can be considered as the probability of x under given all possible models H_i, or the marginal distribution of x. As we see in equation 2.4, this results in a complicated integral, which can make the quantification of $p(x)$ relatively difficult. MCMC methods allow the user to sample from the posterior directly. They promise convergence to the posterior in infinite samples.

MCMC forms a class of a number of different sampling procedures. They all share the same desirable traits, but with small differences between them. Throughout this thesis I will use Gibbs sampling (Gelfand & Smith, 1990; Geman & Geman, 1984), implemented through Just Another Gibbs Sampler (JAGS; Hornik, Leisch, & Zeileis, 2003). Other methods that we could have chosen include the Metropolis Hastings or Hamiltonian Monte Carlo (Neal, 2011) algorithms, as implemented in Stan (Stan Development Team., 2016) and a number of methods implemented by webppl (Goodman & Stuhlmüller, 2014).

For the models implemented in this thesis, many of the MCMC methods mentioned above would have been suitable. Gibbs sampling was chosen primarily because it is very popular. This popularity is partly due to longevity, one of the earliest easily accessible MCMC methods was an implementation of Gibbs sampling in WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) in the late 1980s. JAGS was chosen as the preferred Gibbs sampler and had easy interface with R (R Core Team, 2015) through the package rjags (Plummer, 2016). The popularity was important because I, and the other researchers involved in the projects presented in this thesis, believe the models presented are useful to the wider community. To aid their dissemination we chose to use the most accessible method of implementation. There are also a number of examples that illustrate the suitability of Gibbs sampling for complex hierarchical models (e.g., Growth Curve models, Gelfand, Hills, Racine-Poon, & Smith, 1990). This is the why, but how does Gibbs sampling ensure good estimation of the posterior?

Gibbs Sampling

Bayesian analysis requires some estimate of the posterior $p(H|x)$. This is difficult because to directly estimate the posterior we would need to estimate the marginal distribution $p(x)$. As we have already discussed, this is quite difficult to do for complex hierarchical models (such as those used in this thesis). Gibbs sampling provides us with a method of sampling from the posterior without having to estimate the marginal density.

Gibbs sampling involves iteratively sampling the conditional density instead. Say we have a model parameter space $\boldsymbol{\theta}$ such that $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_k\}$ is the full set of parameters that we wish to estimate. We also have a set of N observations $\boldsymbol{X} = \{x1, x2, x3, ..., x_n\}$ that we hypothesise is generated by the model. Provided we have the conditional distribution $p(\theta_i|\theta_1, \theta_2, ...\theta_{i-1}, \theta_{i+1}, ..., \theta_k, \boldsymbol{X})$, we can use a Gibbs sampling procedure. We describe this process (see Casella & George, 1992; Coro, 2013; Lynch, 2007, for full explanation) next before considering when the conditional density does not have a good solution.

1. Obtain a set of starting values for theta, $\boldsymbol{\theta^{(1)}}$. These can be chosen randomly from priors for $\boldsymbol{\theta}$, or chosen through more advanced methods such as maximum likelihood estimation.

2. Update $\theta_1^{(2)}$ using the conditional distribution $p(\theta_1|\theta_2^{(1)}, ..., \theta_k^{(1)}, \boldsymbol{X})$

3. Sample $\theta_2^{(2)}$ using the conditional distribution $p(\theta_2|\theta_1^{(2)}, \theta_3^{(1)}, ..., \theta_k^{(1)}, \boldsymbol{X})$

4. Continue to sample the i^{th} parameter in the model, $\theta_i^{(2)}$ using the conditional distribution $p(\theta_i|\theta_1^{(2)}, \theta_2^{(2)}, ...\theta_{i-1}^{(2)}, \theta_{i+1}^{(1)}, ..., \theta_k^{(1)}, \boldsymbol{X})$

5. The process outlined in steps 1 through 4 details how to sample one full set of parameters. This process is repeated J times.

Whilst this is a good description of Gibbs sampling, JAGS does not restrict the user to models where the conditional distribution can be expressed (Coro, 2013). Instead a number of sampling algorithms are used (Plummer, 2015).

One such algorithm is the Metropolis Hastings algorithm (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), described by Chib and Greenberg (1995). This algorithm is used in conjunction with Gibbs sampling for steps 1, 3 and 4 where it is not possible to sample directly from the conditional distribution. If $p(x)$ is the distribution that we desire to sample from but cannot directly, and $g(x)$ is some symmetric distribution that we can sample from, then the Metropolis Hastings algorithm allows us to sample from $p(x)$ as follows:

1. Draw a sample from the distribution $g(x'|x_t)$, where x_t is the current sample. This is our proposal as a new member of $p(x)$, but it hasn't been accepted yet.
2. Create an acceptance ratio for this sample, such that $\alpha = \frac{P(x')}{P(x_t)}$, where $P(x)$ is the density of p.
3. The generate a random number u from a Uniform distribution $U[0, 1]$. If $u < \alpha$ then accept the proposed x', otherwise reject and repeat.

For space purposes we will not consider the various virtues of different methods of sampling algorithms here, but instead trust Hornik et al. (2003) that JAGS can appropriately sample from the conditional distribution. In this section we discussed the process of Gibbs sampling in relationship to Bayesian inference. Gibbs sampling, a member of the MCMC family, allows the researcher to sample from the conditional distribution without requiring estimates of the marginal $p(x)$ Unlike bootstrapping procedures, MCMC methods can handle relatively complicated models efficiently.

2.2.3 Type of interval

The previous sections, where I discussed analytic and bootstrapped solutions, demonstrated that different methods can require different types of intervals to represent confidence. While the analytic solution had a formula that suggested the interval with which, with 95% confidence, the true parameter fell within, the bootstrapped 95% confidence interval was derived from quantiles of the distribution of the parameter of interest. Although the intervals are estimated in different ways, they are both estimated using distribution quantiles (the analytic solution indirectly using the t-distribution). In the past two sections we have discussed obtaining a set of samples from the posterior of interest $p(\theta|X)$. Given that we have this set of J samples of the posterior for θ, how should we estimate our uncertainty in our estimate of θ? In this section I will consider two alternatives.

Option 1: Reuse the Quantile method

Perhaps the most obvious option at this stage is to use a similar method as the explicit and boot-strapped intervals and employ a quantile method. To do this we would take the bottom 2.5% quantile and upper 97.5% quantile for a given parameter θ_i. This interval represents the interval that we are 95% confident contains the true parameter estimate for θ. A graphical representation can be seen in the first row, first column of Figure 2.1.

While this method makes sense for a posterior that is uni-modal, for a multi-modal (e.g., tri-modal posterior, first row, second column, Figure 2.1) this interval will contain areas that are very unlikely to contain the parameter of interest. Given this issue, what other options do we have?

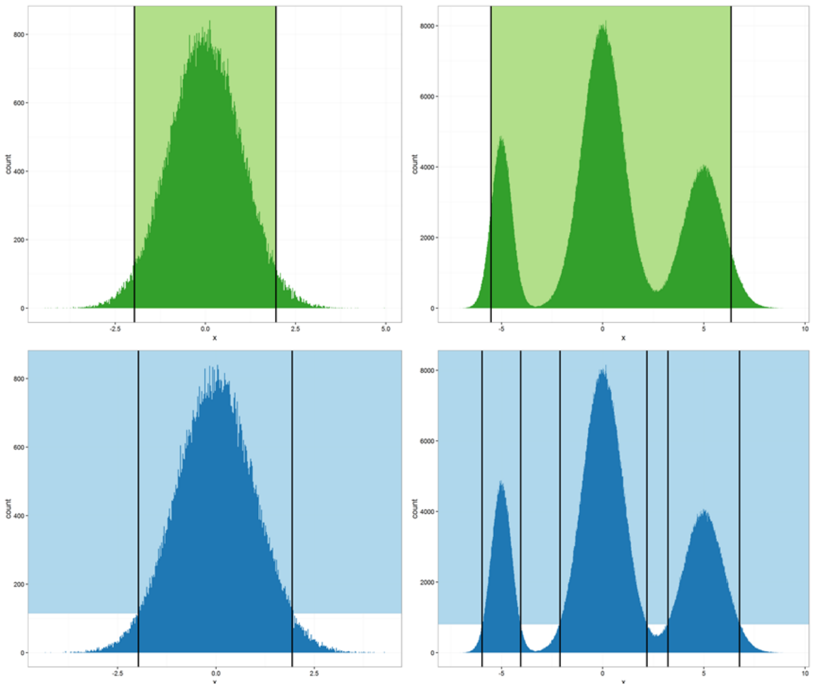


Figure 2.1: **Two different methods of estimating uncertainty** In this figure we demonstrate two different methods of creating a credible interval. The *green* represents the quantile or 'equal-tailed' credible interval, the *blue* represents the highest density interval. Note that while the quantile distribution focuses on estimating the interval that encompasses 95% of the posterior, the HDI estimates the interval that is most likely. For a uni-modal distribution they are relatively similar, but for a tri-modal distribution the HDI returns an interval that is not continuous.

The Highest Density Interval

The Highest Density Interval (HDI) (Hyndman, 1996) seeks to estimate an interval in which the parameter is *most likely*. While the quantile method measures 95% of the mass of the posterior from the x-axis, the HDI measures 95% of the mass of the posterior from the y-axis. The graphical representation of the difference between the two is demonstrated by comparing the two columns in Figure 2.1.

There is very little difference between the interval drawn using the quantile method and the HDI where the distribution is uni-modal. However in a tri-modal posterior (second column, Figure 2.1), there were substantial differences. In this instance the quantile method returns a single, continuous interval that contains two regions of very low probability (i.e, the two troughs between the peaks). The HDI, however, returns a set of intervals that contain the highest probability for that parameter, but are not discrete.

This thesis uses highest density intervals[3]. As we demonstrated above, these intervals are more informative when the posterior is not uni-modal and are more intuitive to interpret (Kruschke, 2014).

2.2.4 Summary

In this section I discussed the problem of estimation. Both frequentist and Bayesian statistics can employ explicit or sampled solutions. While frequentist methods tend to be analytic or estimated via bootstrapping, Bayesian methods tend to be analytic or estimated through MCMC methods. This is reflected in their use throughout this thesis, with Bayesian methods (chapters four through to six) estimated through Gibbs sampling, and frequentist methods (chapters three and six) using analytic or bootstrapped methods.

I also discussed estimating uncertainty through intervals. In frequentist methods uncertainity is represented with confidence intervals. If we draw a countably infinite number of 95% confidence intervals, we would expect that 95% of this set of intervals would contain the true value of the population parameter of interest. In Bayesian methods we *could* draw an interval that reflects the lower and upper quantile of the

[3] In chapter four we also considered quantile intervals with no significant difference in qualitative findings

posterior, but instead choose to estimate the highest density interval. A 95% HDI can be interpreted to contain 95% of the most likely values of the posterior. The differences in interpretation mean that, in this thesis, I will not directly compare frequentist and Bayesian methods. Instead I will compare models developed within a single statistical framework.

2.3 Assessing the estimation

In the previous section I discussed how statistics can be used to estimate parameters of interest. In this section I will consider the methods of understanding whether the estimation is good. This can encompass quite a wide variety of techniques, including model comparison techniques. I will focus primarily on two main areas. The first is used primarily in explicit frequentist methods. It involves methodologically checking whether any assumptions made in constructing the explicit solution hold true for the data of interest. The second method relates primarily to the use of Gibbs sampling to estimate the posterior distribution. Here we question how to check that the procedure has converged to the posterior.

2.3.1 Assumption Checking

Explicit solutions rely on assumptions which, if violated, indicate that the solution may not hold. For every assumption that could be made, there are a number of possible assumption checks that could be conducted. Assumptions like contamination, response heterogeneity and missing data do not necessarily require assumptions checks because they are either obvious (in the case of missing data, chapter six) or difficult to test for (as we will see in the case of contamination and response heterogeneity presented in chapters four and five). Distributional assumptions, however, underpin explicit solutions and can be tested for.

Here I discuss how to test for a common distributional assumption, normality (which is particularly relevant to chapter three) using the Shapiro-Wilk test and quantile-quantile plot.

Shapiro-Wilk Test

The Shapiro-Wilk test of normality (Shapiro & Wilk, 1965) is a null hypothesis significance test that questions whether a given sample distribution was drawn from a population that was normally distributed. The null hypothesis for this test is that the sample is from a normally distributed population, the alternative that it is not. It relies on the W statistic, which calculates the order[4] of the samples and compares them to the order that we would expect if the sample was drawn from a normal population.

The Shapiro-Wilk relies on comparing the observed order of the sample, and the expected order of a normal distribution. This is computationally complex (though once the order is calculated, the analytic solution is quick to calculate), so samples greater than 50 cause difficulty (Rahman & Govindarajulu, 1997). As a solution to this, approximate solutions have been proposed (Rahman & Govindarajulu, 1997;

[4] The order is a list of the indices of the smallest to the largest number. It differs from the rank when there are duplicate numbers, which the rank attempts to average, whilst the order counts separately.

Royston, 1992). Figure 2.2 shows the application of a Shapiro-Wilk test to random samples from three different distributions. These distributions were chosen to illustrate the ability of the Shapiro-Wilk to identify samples drawn from a normal distribution (*blue*), a distribution that very slightly deviates from normality ($t(15)$ distribution, *green*) and a distribution that drastically differs from normality (uniform distribution, *orange*). Each column of this figure corresponds to a different distribution.

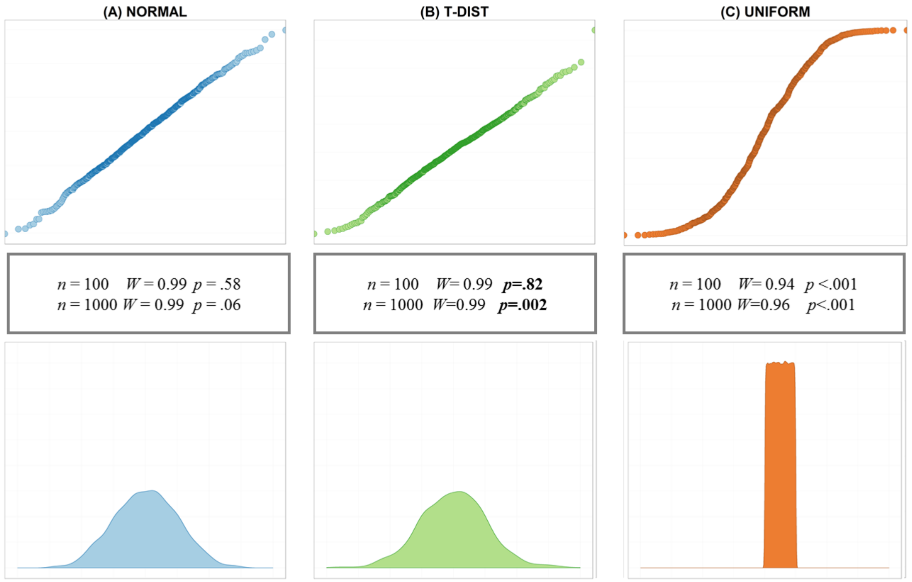


Figure 2.2: **Common methods of testing for normality** The columns represent three different underlying distributions. The bottom row shows each distribution, the top row shows the Normal Q-Q plot, and the text indicates the results of a Shapiro-Wilk test for samples size of 100 and 1000.

The Shapiro-Wilk is considered to be one of the highest powered methods of testing for normality (Yap & Sim, 2011), but, as seen in Figure 2.2 there is still potential to make an error. This means that it is also important to consider other methods of inspecting normality, such as quantile-quantile plots.

Quantile-quantile plot

Normal quantile-quantile plots can be used in addition to the Shapiro Wilk distribution to investigate normality. To create these plots the quantiles of the observed distribution are plotted against the quantiles of a normal distribution of similar size. If the observed distribution is normal, we would expect the relationship between to be linear, as seen in the upper right plot of Figure 2.2. If the observed distribution slightly deviates from normal, we see the slight deviance that can be observed in the middle Q-Q plot (*green*). If the observed distribution highly deviates from a normal distribution, we see a non-linear relationship like in the top left Q-Q plot (*orange*). Although the non-linear relationship is quickly differentiated from the other two; it is more difficult to differentiate between a linear trend (*blue*) and a slight deviation (*green*).

Assumption checks in the context of this thesis

In this section we briefly considered two different methods to test for normality. Testing for normality is important in the context of chapter three. However, it is also important to understand that explicit solutions involve the creation of estimators through assumptions about the data, and these estimators only hold true if the assumptions also hold true. These checks are a method to consider whether the assumptions of the model are supported by the observed data.

The simplicity of the calculations for explicit solutions belies that there are two issues at hand. The first is the main concern of this thesis, the degree to which the assumptions of the model match the true assumptions of the data. The second relates to our ability to calculate model estimates given the data. With explicit solutions, this is often an exercise in simple arithmetic, and can be considered separately to the appropriateness of the model. With sampling-based MCMC methods, the ability to calculate model estimates is of prime concern. In addition, while analytic solutions largely do not depend on the model being appropriate to produce some estimate (regardless of whether this is a good estimate or not), MCMC methods will often struggle to achieve the true posterior distribution for ill-specified models. In the next section I will consider methods of investigating and solving problems associated with this. These methods will be used extensively throughout chapters five and four.

2.3.2 Methods of Assessing Gibbs Sampling procedures

So far we have seen that, provided all necessary assumptions are met, an explicit solution has certain guarantees attached to it. This leads to assumption checks as a method of assessing the 'trustworthiness' of an estimate. Gibbs sampling does not need to make these assumptions, and instead updates a prior with respect to observed data. Provided the prior allows for the true parameter value, the prior can be updated to estimate the posterior for a given parameter. However this convergence is not guaranteed in any particular time frame. How might we assess whether a given procedure has accurately estimated the posterior?

Burning in

When the MCMC chain is initialised, it requires a starting value for the parameters, $\theta^{(1)}$. As we previously discussed, this starting value should be a plausible and likely member of the posterior distribution. Oftentimes, this is achieved by either randomly sampling from the prior, or taking a MLE estimate of the prior. If the prior is close to posterior, this is an acceptable procedure. However, sometimes the prior overlaps the posterior with only low probability. In this instance the initial values might be very rare members of the posterior, and hence the sampling procedure will take a number of iterations in order to converge to sampling from the posterior. Figure 2.3 illustrates exactly this scenario. It depicts a Gibbs Sampling routine, implemented through JAGS, to estimate the population mean, μ of a sample. The prior distribution for μ is considerably different to the posterior. This means that the initial values (highlighted in yellow) are part of the burn in of the model and should be discarded.

In this example it is relatively simple to see through visual inspection where the sampling method appears to have converged. More formally though, sufficient burn in time is not a trivial claim to make. Often this is completed through two methods. Firstly, through visual inspection, similar to the process for Figure 2.3. JAGS also has an automatic procedure, using adaptive procedures until convergence is obtained. Samples are only kept once convergence has been achieved. A third option is to simply set a long burn in and risk discarding viable samples from the posterior. Whilst this is not the most efficient of methods (it can increase the time taken to fit the model), it does not impact the quality of the posterior samples.

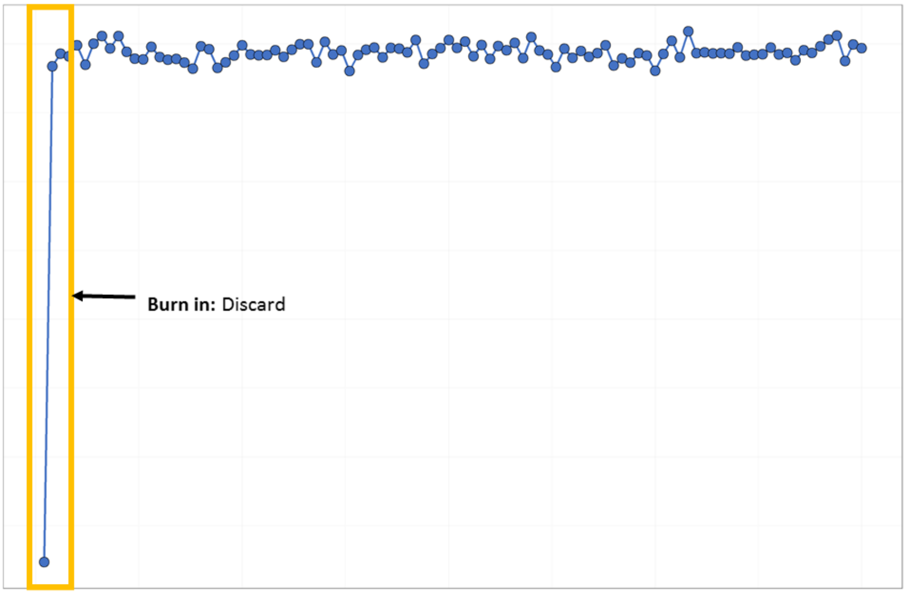


Figure 2.3: **Demonstrating the importance of a burn in period** This figure demonstrates the importance of a burn in period to obtain good samples from the posterior. The x-axis represents the sample iteration, while the y-axis represents the corresponding sample for this iteration. The first samples from the MCMC chain are a very long way from the posterior that the chain converges to. These are highlighted in *yellow* to indicate that these samples would probably be discarded.

In the simulations presented in chapters four and five we use a combination of the automatic adaption implemented through JAGS and a long burn in for conservative reasons. The scale of the simulations presented meant that it was not possible to visually inspect each of the chains, but examples of each model type were usually inspected.

Mixing of chains

In the previous section we considered the burn in period: What it is, how to detect it, and why it should be removed. However, simply removing the burn in period does not guarantee the remaining samples have converged to the posterior. Another common metric of convergence is the mixing of multiple chains.

Imagine we start two Gibbs chains so that they run independently for the same number of samples. If they both converge to the true posterior, we should expect that the chains should intermingle, and not get caught in local minimum or maximum. But what happens if two independent chains do not mix?

Figure 2.4 demonstrates the answer to this question for a rather specific example. In this instance I deliberately miss-specified a model as a mixture of two normal distributions, whilst giving observed data of data drawn from a simple normal distribution. These two chains are the Gibbs samples for the first mean of the mixture. While they oscillate roughly around the true mean of the sample, there are two problems at hand. The first is that the two chains do not mix appropriately. The yellow highlighted areas are examples of poor mixing. Here the two chains do not appear to be sampling from the same posterior.

This is, in part, caused by a failure of one of the founding rules of MCMC sampling; a memory-less feature corresponding to the formal claim that $p(\theta_{i+1}|\theta_i, \theta_{i-1}, \theta_{i-2}, ..., \theta_1, x)$ is equivalent to $p(\theta_{i+1}|\theta_i, x)$. In Figure 2.4, we can see that this is *not* the case. In the middle of the figure the chain is monotonically increasing, indicating that we could predict the i^{th} sample by the $i_{i-2}nd$ sample or the $i_{1-3}th$ etc.

In this example we see improper mixing of chains due to model misspecification, but the problem may also be created when there is a strong correlation between parameters (Lynch, 2007). Both cases have the same potential solution, thinning, discussed in the next section.

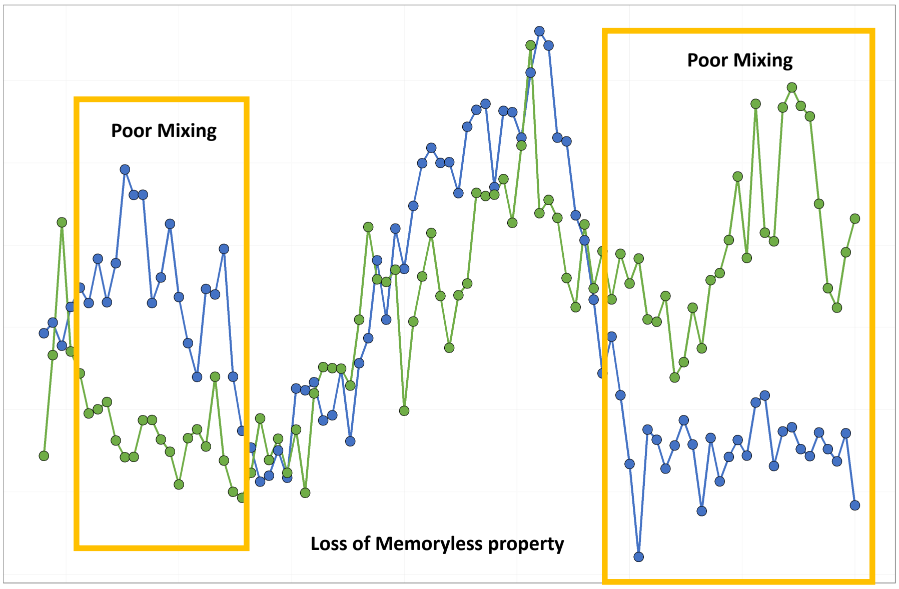


Figure 2.4: **The importance of checking chains** This figure demonstrates that two MCMC chains that were starting in the same manner may not necessarily sample from the whole posterior. Yellow boxes highlight areas of divergence between the chains.

Thinning of chains

One solution to poor mixing is to employ thinning: keeping one sample and discarding the next n samples, before keeping the $n + 1$ sample. The purpose is to ensure the samples that are kept are not strongly correlated with any other samples in the chain, which allows the researcher to obtain a reliable estimate of the posterior.

Figure 2.5 shows two chains constructed for the same model and data as Figure 2.4. The inset of Figure 2.5 is of the same scale as figure 2.4, for comparison purposes. Note that in this figure, the chains mix well. The difference in Figure 2.5 is that I thinned out the MCMC chain by choosing to keep every 100^{th} sample. This was perhaps a dramatic and inefficient method of ensuring that the chains would mix, but it was effective.

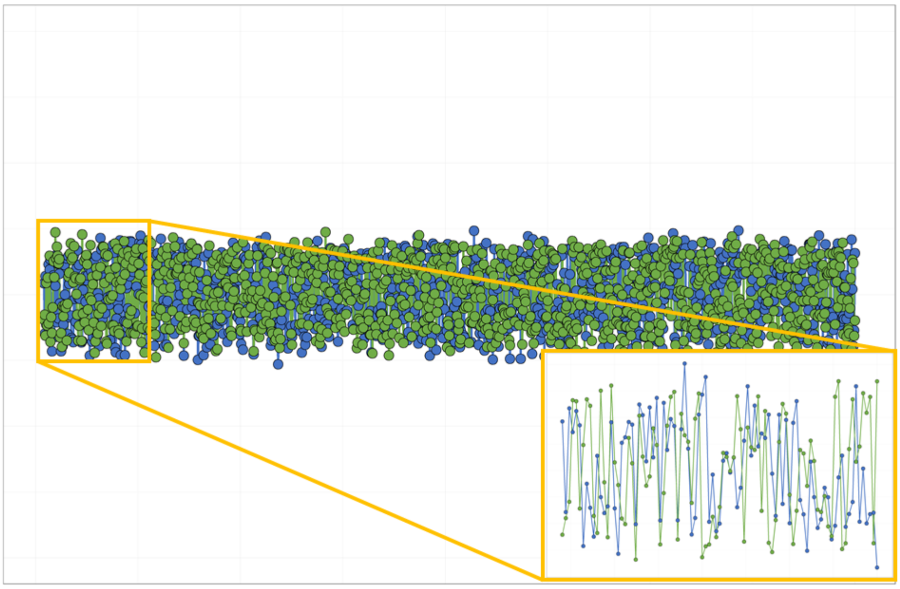


Figure 2.5: **The importance of thinning in MCMC** Two Gibbs-sampled chains for the same model and data as Figure 2.4, but with thinned samples to obtain proper mixing. The magnified section is analogous to Figure 2.4, but 1000 samples are shown to demonstrate the effects are consistent throughout the entire chain.

2.3.3 Summary

This section discussed how to evaluate the parameter estimates obtained from a MCMC procedure. In the case of explicit solutions, the assumptions upon which the explicit solutions were derived must hold true. In the case of sampling solution methods, it is important to analyse the convergence of such methods. Different chapters of this thesis rely on different evaluations. Chapter three relies on analytic solutions; chapters four and five rely on sampling solutions that ensure proper convergence by running for a long time; and chapter six uses both Bayesian and frequentist methods where appropriate.

2.4 Assessing the model

Having considered the difference between Bayesian and frequentist statistics as well as how to estimate parameters and evaluate the goodness of that estimation, it is also important to know how to evaluate entire models. Which of several models is most trustworthy given the data at hand?"

Oftentimes when research into the 'best' model for a given data-situation is done it focuses on model comparison techniques. These methods describe the degree to which each model fits the data, and quantify which model is *most likely* to match the one that generated the data. These methods include information-criterion methods such as the AIC (Akaike, 1974), BIC (Schwarz et al., 1978), DIC (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002), WAIC (Watanabe, 2010) and evidence weighting methods like the Bayes Factor (Kass & Raftery, 1995).

This is a worthwhile and useful way of analysing data. However, this thesis is not interested in which model best fits the data. We, like Box (1976), take the view that "all models are wrong but some are useful." I focus on weighing up the usefulness of models, and this section will focus on how we can weight these models up in terms of usefulness.

While usefulness is not necessarily a statistical criterion, it is a practical question that will directly impact the functionality of a model. A model that is perfectly accurate but completely uninterpretable is of limited use when it comes to creating, considering and extending hypotheses in a psychological research setting.

My approach mirrors the Neumann-Spearman school of thought. It is my belief that a good model is one where the claims that can be made from it can be trusted for the widest set of data situations possible. The model comparisons undertaken in this thesis thus focus on the relative ability of each model to correctly identify and estimate a relationship where there is one (i.e., power) and to correctly claim there is no relationship when there is not (i.e., Type 1 error).

Figure 2.6 shows the four possible options for a statistical decision. First, we consider the options if there is *not* an effect. If the model correctly makes this claim (*light blue*), then we would consider this a desirable trait. If there is not an effect and the model incorrectly claims there *is* an effect (*light green*), then we would *not* consider this a desirable trait. This is known is a Type 1 error.

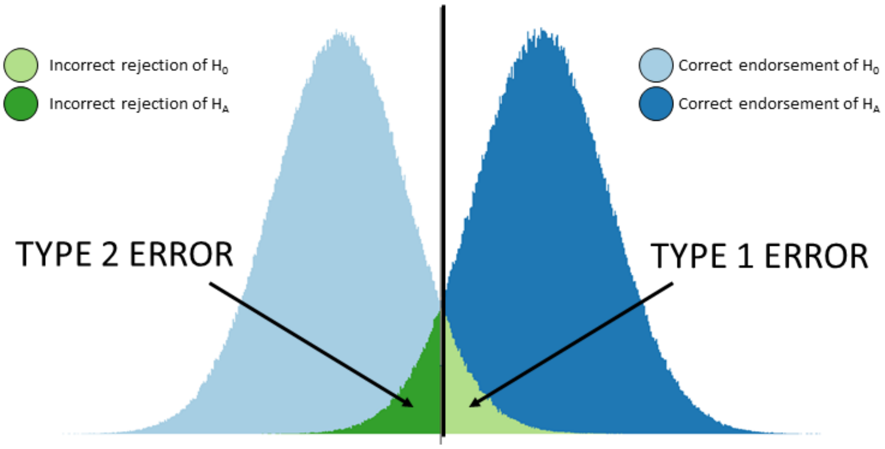

Incorrect rejection of H_0
Incorrect rejection of H_A
Correct endorsement of H_0
Correct endorsement of H_A
TYPE 2 ERROR
TYPE 1 ERROR

Figure 2.6: **Possible outcomes for detecting an effect.** The first distribution represents the possible outcomes if no effect is found, whilst the second distribution represents the possible outcomes if there is an effect found. There is some overlap. An ideal statistical method minimises the green area (errors) in two ways: Firstly, by adjusting the criterion (black bar) to an optimal position; and secondly, through lowering the relative overlap. This is often achieved by increasing the sample size.

If there *is* an effect and the model correctly identifies it (*dark blue*), then we would consider this a desirable trait. If the model does not identify the effect (*dark green*), then this it is considered a miss. Too many misses a non-desirable trait. The ability of a method to identify an effect of a given size is called the power.

Traditionally there are three main factors that impact a method's power: sample size, effect size and the specified acceptable level of type 1 errors (i.e., α). Perhaps unsurprisingly, the larger an effect size, the easier it is to detect the effect. Increasing the sample size decreases the standard error of the measurement, which also helps to detect the effect. These two factors also interact, and thus in my analyses they are varied independently

2.5 Summary: Our approach

In this thesis my interest lies with specific features of the model structure, and how that relates to the researcher's underlying assumptions about how the data were generated. To do this I needed to create, fit and analyse models of differing structure in a way that allowed only the structure to change. This chapter laid out the overarching decisions I made that enabled me to accomplish this throughout the thesis as a whole.

One of the first decisions was the choice of probabilistic theory (i.e., whether to use frequentist or Bayesian methods). Rather than choosing one or the other, I chose instead to use the methods that were most convenient for the problem at hand. This means that while chapter three uses frequentist methods, chapters four and five both use Bayesian methods implemented through JAGS. In these three chapters I deliberately avoided comparing Bayesian and frequentist methods against one another, as this would have introduced confounds (i.e., is a method performing better due to its structure or due to being Bayesian?). Chapter six, which does not directly compare models against one another but instead uses them as tools for data analysis, uses both.

Even within each modelling framework, there were still choices that needed to be made about how the models were fit to the data. Where possible I used already derived explicit solutions because of their speed of computation. This meant that many of the frequentist methods in chapters three and six were implemented through explicit solutions. Where explicit solutions were not easily available, I implemented those methods through more computationally intensive techniques (bootstrapping in chapter three and Gibbs sampling in chapters four to six).

Using different types of methods to fit models requires different methods to evaluate the quality of the estimation. For explicit solutions, this meant considering the underlying assumptions of the method. Chapter three focuses on the issues that can occur when this is not completed, while chapter six relies on these assumption checks to ensure the claims being made from the data are trustworthy. Gibbs sampling, implemented through JAGS, required more work. The sampling chains had to be assessed for appropriate burn in time and appropriate convergence.

Lastly, and perhaps most importantly, in this thesis I am interested in the impact of the structure of the model on the nature and “trustworthiness” of the claims made by the model. This meant that I needed to to compare and contrast on three main attributes: the ability to reject the null when it is false (i.e., power), the ability to retain the null when it is true (i.e., Type 1 error) and the ability to make claims about the aspects of the data that are of theoretical interest to the researcher. While these first two points proved to be particularly important to chapter three and four, this last point proved to be important to chapter five[5].

With these important methodological issues concluded in this chapter, in the next chapter I turn my attention to the first assumption of interest in this thesis, the assumption of normality.

[5] Whilst this is strongly linked to frequentist ideology, these traits are also desirable from a Bayesian standpoint.

Chapter 3

When the distribution is skewed

> It is impossible to escape the impression that people commonly use false standards of measurement
>
> *Sigmund Freud*

3.1 Preface

The goal of this chapter is to explore how specific uses of a scale can lead to difficulties with statistical analysis. I argue that these problems are due to the properties of an interaction between the scale and sample. Freud considers individuals who use the wrong comparison units to weigh up merit, which seemed apt for a chapter that devotes itself to a problem that is somewhat invisible to the traditional metrics of scale comparison. This problem is the violation of normality.

Normality violations are common across a wide range of fields. This chapter focuses on a common problem in psychology: attempting to measure some underlying weak signal of clinical symptomatology (in this case depression) in a sample where the overwhelming majority of individuals do not display strong indicators. When described this way it seems to be a bizarre practice, but this type of research often yields a variety of beneficial outcomes. These range from describing the degree of symptomatology in a specific population (such as nurses (Glass, McKnight, & Valdimarsdottir, 1993) or students (Olsson & Knorring, 1999)), describing the relationship between symptoms and other covariates to obtain a richer picture of co-morbidity (e.g., depression and stress (Hewitt & Flett, 1993), or depression and exercise (Mackenzie et al., 2011)) or even group differences within a population (e.g., gender differences). Others use this method to obtain a sample large enough to test various features of the scale (e.g., Steer & Clark, 1997), which would not otherwise be possible with rarer clinical populations. Clinical populations can also be hard to access and often introduce difficulties of confounding factors (medication, co-morbidity, etc.). We argue the scale-sample interaction results in data that (almost always) violates the assumption of normality.

This chapter investigates the benefit of using minimally descriptive models (in this case, models that make very few assumptions) as a potential solution to assumption violations. We compare these models to the most popular modelling methods, as well as more complex models with the intention of quantifying both the Type 1 and Type 2 error rates associated with each. Interestingly the results suggest that the more complex models *do not* perform the best. Instead the moderately complex models outperform the others, with the least complex models also performing relatively well.

Because of this, this chapter is at odds with the results of the other chapters of this thesis. As I will discuss, the more complex models presented in subsequent chapters represent some underlying theory of how the data were generated. The most complex models in this chapter represent models that have a large amount of freedom to explain variation in the data. This implies that even if they were equivalent in accuracy and precision to the less complicated models (and they are not), using them does not add anything in terms of the richness of the claims that can be made from the data (in fact they are often more complicated to interpret). The exception to this is the models outlined in the addendum.

This chapter shows that violating statistical assumptions is not desirable, does not produce desirable outcomes, and reduces the accuracy of the claims that can be made. Often the researcher has some underlying belief that this violation might occur given previous experience, and this belief could guide their analysis.

My work shows that this is a common problem that arises as a specific feature of the sample-scale interaction (rather than being a product of the scale itself). I conclude by simulating potential solutions to the problem.

Statement of Authorship

Title of Paper	On the use of clinical scales in non-clinical populations: A discussion of the analysis of.	
Publication Status	□ Published □ Submitted for Publication	□ Accepted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Unpublished	

Principal Author

Name of Principal Author (Candidate)	Lauren Ashlee Kennedy		
Contribution to the Paper	Contributed to the original core concept of this experiment, conducted the literature review and analysed the results. Planned (in conjunction with DN and AP) the simulation procedure, conducted and analysed the results. Wrote first draft of the manuscript and contributed to editing process.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Daniel J. Navarro		
Contribution to the Paper	I helped figure out which analyses and simulations to run and contributed to editing the paper.		
Signature		Date	31/8/2017

Name of Co-Author	Amy Perfors
Contribution to the Paper	I helped figure out which analyses and simulations to run and contributed to editing the paper.

Signature		Date	31/08/2017

Name of Co-Author	Nancy Briggs		
Contribution to the Paper	I helped figure out which analyses and simulations to run.		
Signature		Date	31/8/2017

Please cut and paste additional co-author panels here as required.

On the use of clinical scales in non-clinical populations: A discussion of the analysis of skewed responses

Lauren A. Kennedy

School of Psychology

University of Adelaide

Daniel J. Navarro

School of Psychology

University of New South Wales

Amy Perfors

School of Psychology

University of Adelaide

Nancy Briggs

Mark Wainwright Analytical Centre

University of New South Wales

3.4 Abstract

The administration of clinical scales (e.g. the Beck Depression Inventory, the Depression, Anxiety Stress Scales, the Patient Health Questionnaire 9 and the Centre for Epidemiological Studies Depression Scale) to non-clinical samples is common in psychological research, and such data are usually analysed with the help of statistical tools that assume normality. In this paper we consider whether this practice is sensible. We present a review of the literature, empirical data and simulation studies to argue that by design the use of clinical instruments will produce skewed datasets when applied to a non-clinical population. We argue that while the presence of skewness is not inherently problematic, it becomes so when the statistical methods applied to data are not robust to violations of normality. We show that skewness can have a very severe effect on the power of a study to detect an effect, which can impact how trustworthy the results are and hence impact the ability of practitioners who use these results to base practice upon. Finally, we consider alternative statistical tools available to researchers, and argue that simple nonparametric tools that analyse rank ordering of scores are very effective at addressing this problem.

Keywords: Statistical methods, community samples, skewness, robust statistics, depression

On the use of clinical scales in non-clinical populations: A discussion of the analysis of skewed responses.

The development of questionnaire instruments plays an important role in psychology, both in the clinical and non-clinical domain. In a clinical context, a researcher or practitioner might use a tool such as the Beck Depression Inventory (BDI) to assess the severity of depression; in a non-clinical context, a personality questionnaire might be used to measure extroversion. The statistical tools used to validate new scales tend to be similar in both cases: researchers care about test-retest reliability, internal consistency, discriminability, convergent validity and so on.

However, one critical difference between those instruments designed for a clinical context and those that are not is that clinical instruments are designed to be maximally informative when administered to a clinical population rather than a non-clinical population. The BDI, for example, is very deliberately *not* a general purpose measure of how happy someone is: the BDI scores for an unhappy but non-depressed person will not differ greatly from the BDI score of a happy non-depressed person. The purpose of the BDI is to be diagnostic within the clinical spectrum and to discriminate between depressed and non-depressed people. As a consequence, when administered to the general populace, almost all people end up with very low BDI scores and only a few people will have large scores. In other words, non-clinical samples the distribution of BDI scores should be expected to be *skewed*. The same phenomenon does not occur with instruments that are designed for non-clinical populations: by design, the distribution of scores for most measures of cognitive ability or personality will tend towards normality.

The fact that clinical tools are designed to be most informative in clinical populations is highly desirable; it is after all the very purpose of the instrument. However, when those same instruments are administered to a largely non-clinical population as part of a research project, this creates substantial statistical problems. Most data analysis tools used by psychological researchers (e.g., t-test, linear regression, etc.) rely heavily on an assumption of normality. Yet if a clinical instrument such as the BDI is administered to (say) a sample of undergraduate students, one should *expect* that normality will be violated because – by design – most students will have BDI scores near zero. In other words, even before looking at the data one should expect that if a clinical assessment tool has been applied to a non-clinical population, one of the critical assumptions underpinning most statistical tests will have been massively violated.

In this paper we discuss (1) the prevalence of this problem in psychological research, (2) the potential seriousness of this issue, which differs from test to test, and (3) possible remedies for the problem. We focus in particular on the Beck Depression Inventory, not because we believe it is a bad instrument, but because it is a very commonly applied tool and is generally regarded as a good psychometric instrument. The BDI is a popular scale with a theoretical foundation based on the diagnostic criteria in the Diagnostic and Statistical Manual (DSM). It has well demonstrated psychometric properties, including test-retest (.48 to .86) and internal reliability(.73 to .95), discriminative and concurrent validity (see Beck, Steer, & Carbin, 1988, for a review). It has been translated into several different languages (e.g. Ghassemzadeh, Mojtabai, Karamghadiri, & Ebrahimkhani, 2005; Shek, 1990; Wiebe & Penley, 2005), and now takes on a number of different forms (e.g. BDI-SF (Beck, Rial, & Rickels, 1974, 2), BDI-II (Beck, Steer, Brown, et al., 1996)). As it turns out the issue of skewed data *is* a problem with regards to the BDI, but as we discuss later in the paper the problem is likely to be prevalent elsewhere as well.

3.5 Is skewed data prevalent in the empirical literature?

To motivate the problem, consider the structure of the BDI scoring system, in which the upper half of the scale (scores 31-63) correspond to 'severe' depression, the middle section of the scale (scores 10-30) cover the 'mild to moderate' range, and only a small range at the bottom of the scale (scores 0-9) is used to describe non-depressed individuals. Estimates of the incidence of depression vary somewhat from study to study (e.g. 8% for young adults, 13.5% on average for older adults, Beekman, Copeland, & Prince, 1999; Kashani et al., 1987), but for the sake of argument let us suppose that at any given point in time 10% of the populace falls within the depressed range. In that situation, if the BDI is administered to people sampled from the general population, 90% of BDI scores should fall within the bottom sixth of the scale, and only 10% of the BDI scores will be distributed across the upper five-sixths. This is not a characteristic of normally distributed data and almost certainly implies a skewed distribution. This effect is illustrated in Figure 3.1, and demonstrates that the skewness arises because of the fact that the clinical range is (by design) very wide, yet only a small proportion of people fall within that range.

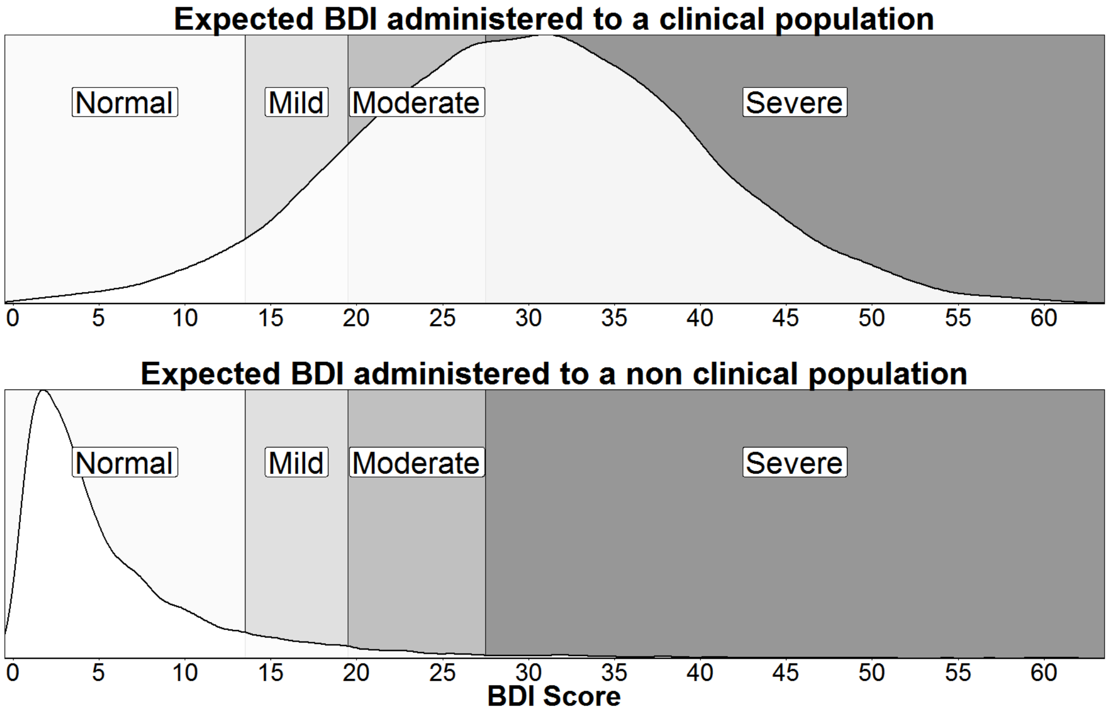


Figure 3.1: This figure demonstrates how sampling a non-clinical population with the BDI will result in a skewed distribution, with 90% of the sample falling into an extremely narrow range on the scale.

Focusing on the characteristics of individuals in the clinical range is quite common amongst scales designed to measure depressive symptoms, and the BDI is by no means the only instrument designed to do so. For example, another popular scale used for this purpose is the Depression and Anxiety Stress Scales (DASS; S. H. Lovibond & Lovibond, 1993), and while our paper will focus on the BDI, we note that the same issue appears when we consider the DASS. To illustrate this, Figure 3.2 plots the distribution of depression scores obtained from 451 middle aged and older adults from the community (Ward, 2015)[1]. Less than 5% of the sample score in the upper half the scale, and 82% (approximately one-sixth) of the sample score in the first quarter. Thirty percent of individuals scored zero, indicating a strong

[1] This data was collected with the approval of the University of Adelaide ethics committee

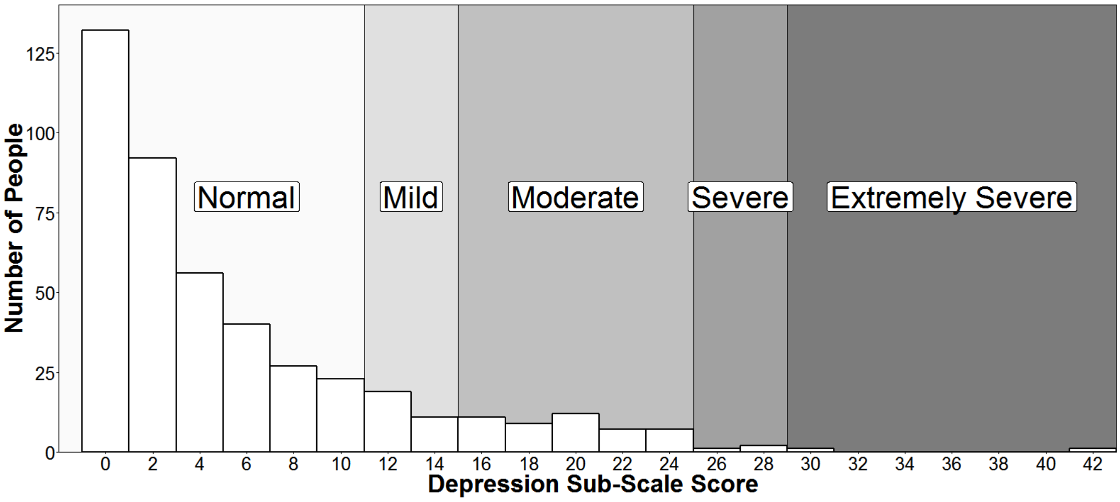


Figure 3.2: Participant scores from a community sample on the depression subscale of the DASS. Over 95% of the scores fall in the lower half of the scale, representing a highly skewed sample, as is typical when a clinical measure is applied to a non-clinical population.

floor effect. Based on these descriptive statistics alone it is hardly a surprise to find that the empirical data from the DASS plotted in Figure 3.2 are more or less identical to the pattern we predicted for the BDI based on a consideration of its structure. The DASS depression scores are very skewed (skew = 1.61) and heavy tailed (kurtosis = 5.68). Not surprisingly, a Shapiro-Wilk test applied to these data found a significant departure from normality (W = 0.80, $p < .001$). In short, although the rest of this paper focuses on the BDI specifically, the problem at hand is a far more general phenomenon, and should be expected to be relevant any time an instrument designed for clinical populations is applied to a community sample. We will return to this topic later.

3.5.1 Skewed BDI scores for non-clinical populations

The preceding discussion suggests that there are good theoretical reasons to expect that administering a clinical measure to a non-clinical population will produce non-normal data, and specifically to expect that most such datasets will be quite skewed. From a clinical perspective this is highly desirable, reflecting as it does the fact that relatively few people report severe depression symptoms. However, from a statistical perspective it presents some difficulties, especially if researchers apply data analysis methods that are not designed to be applied to such data. With that in mind, it is critical to determine whether this represents a problem with current research practices as they are reported in the literature. We consider these isses in some detail in the next section.

To obtain a sense of what current statistical practice in the discipline looks like we conducted a literature search of PubMed, PsychInfo and Embase for all peer-reviewed papers that used any form of the Beck Depression Inventory (i.e., including the BDI, BDI-II, and BDI-SF), which produced 56,777 results.[2] After restricting the search to articles published in English during the years 2009-2014 inclusive that used non-clinical or undergraduate population, we obtained a sample of 2095 papers. After reading titles and abstracts for all 2095 papers, we excluded 700 papers that included some clinical participants, were non-experimental or observational in nature (72 book chapters, 14 conference abstracts and 46 literature reviews), were based on a sample size of one (25 case studies), mentioned the Beck Depression Inventory but did not use it (7 articles) or were duplicates that were previously unscreened (76 articles). The final result was a sample of 1155 published papers in which the BDI was administered to a non-clinical sample. A simple random sample of 250 papers (approximately in 1 in 4.6) was selected for closer analysis.

[2] Full details of this literature search can be found in the Appendix A, included at the end of this chapter.

Each of the 250 papers was read closely. Any papers that should have been rejected during the abstract reading stage were rejected, leaving a total of 148 papers meeting the criteria stated above. Assuming that the remaining papers are similar to those that were read, we estimate that approximately 684 of the 1155 papers in our final sample meet the criteria. Since our sample extended over five years, this exercise suggests that approximately 114 papers are published annually in which the BDI is administered to non-clinical population.

Across the 148 papers there were a total of 153 distinct studies. The mean sample size for these studies was $N = 490$ participants, but this average is strongly influenced by a few very large studies, and the median sample size was much smaller at $N = 244$. The smallest study reported $N = 28$ particiants and the largest reported $N = 7172$. Sixty-nine studies used the BDI, 68 used the BDI-II, 6 used the BDI-SF and 10 used idiosyncratic modifications to the BDI. Although English was the primary language, the studies report 17 other administration languages as well.

Of the 153 studies examined, 104 reported at least one sample mean and 100 reported either the standard error of the mean or the standard deviation. Figure 3.3 shows the results from studies reporting only on the version of the BDI with scores ranging from 0 to 63.[3] While it would be nice to estimate exactly how heavy the tail of a typical distribution is, relatively few studies (only 24 in our sample) report the proportion of individuals at each quartile or criterion level. Nevertheless, it is straightforward to demonstrate that the BDI scores in most community samples *must* have been non-normal and positively skewed, simply on the basis of the reported means and

[3] We plot scores only from the full and standard BDI scale so that they are directly comparable.

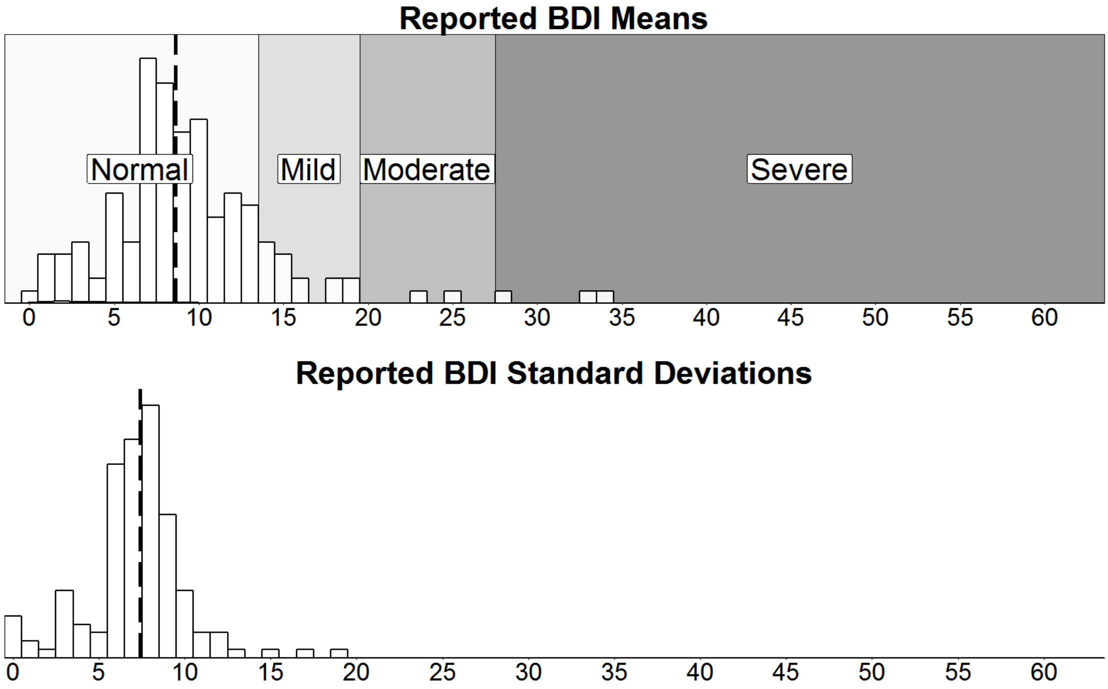


Figure 3.3: Top panel: Distribution of the means from the 103 studies that included at least one sample mean and used the BDI in a form such that the scores ranged from 0 to 63. It is evident that most means fell within the non-depressed range; the median sample mean, shown with the dotted line, was 8.60. Bottom panel: Distribution of the standard deviations from the 97 studies that included at least one measure of variance. In general, standard deviations were similar in size to the means (strongly suggestive of a skewed distribution); the median sample standard deviation, shown with the dotted line, as 7.40.

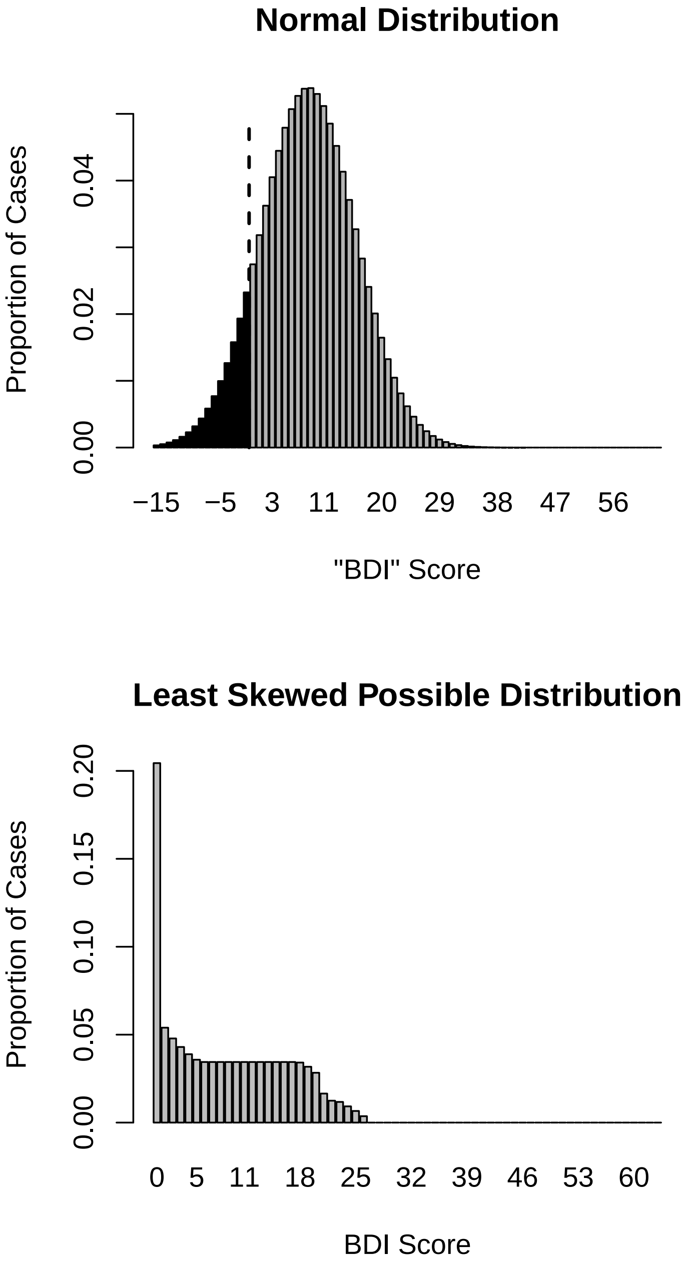


Figure 3.4: BDI scores in community samples are almost certainly skewed. The top panel plots a normal distribution with mean 8.6 and standard deviation 7.4, consistent with the typical values reported in the literature: a substantial proportion (about 12%) of the distribution lies below 0 (black bars), corresponding to impossible values of the BDI. On the bottom panel, we plot the *least* skewed unimodal distribution of BDI scores that could produce a mean of 8.6 and standard deviation of 7.4, and it is clear from inspection that the departures from normality are very substantial.

standard deviations. To illustrate this, consider a study that reports a mean BDI score of 8.6 and a standard deviation of 7.4, which is grossly typical of the values actually reported in the literature. Given that the minimum possible BDI score is 0, it is impossible for these scores to be normally distributed, as shown in the top panel of Figure 3.4. In order for these scores to satisfy normality approximately 12% of the participants must have had BDI scores below 0 – which is of course impossible – and the moment one starts to truncate the lower tail of this distribution the resulting scores are necessarily positively skewed, thus ensuring that the violations of normality are *systematic*. The severity of these violations are impossible to estimate without access to raw data, but it is not too difficult to show that even in the most optimistic case the violations are non-trivial. To demonstrate this, the bottom panel plots the *least skewed* distribution of BDI scores that produces a mean of 8.6 and standard deviation of 7.4 while constraining all the BDI scores to lie between 0 and 63, and forbidding the distribution to become bimodal.[4] It is obvious from inspection that this distribution is very severely non-normal (its skewness of 0.4 is essentially the lower bound on the skewness) and much more closely resembles the empirical distribution of DASS scores in Figure 3.2 than a normal distribution. In short: when applying a clinical tool such as the BDI to a community sample, the *default* assumption should be that data will be non-normal and positively skewed.

[4] Estimating this distribution is not difficult: it is a constrained optimisation problem defined over 63 unknown values, and is easily solved using standard optimisation tools: we used the "L-BFGS-B" method in the `optim()` function in R to do so, using a penalised error function to enforce constraints, but we imagine any standard tool would work well.

3.5.2 How widespread is the skewness problem?

Up to this point our discussion has focused mostly on the application of the BDI to non-clinical populations, but – as foreshadowed by the fact that DASS scores follow the same pattern that we predicted for BDI – we suspect that the phenomenon is much more general. To illustrate this generality, we undertook a smaller scale investigation of two other scales commonly used to measure depression or dysphoria in a non-clinical populations.

The first extension we examine is the Patient Health Questionnaire-9 (PHQ-9). The PHQ-9 was developed as part of a battery of mental health screening tests (PHQ-9; Kroenke, Spitzer, & Williams, 2001) and is similar to the BDI insofar as the items reflect the diagnostic criteria in the DSM-IV (Kroenke et al., 2001). The tool is designed for use by primary care practitioners to diagnose and screen for depression in patients, and as such it is primarily a clinical instrument. This is evident from looking at the ranges of scores corresponding to different diagnostic categories: PHQ-9 scores range from 0 to 27, with scores below 10 indicating a low risk of depression and scores over 15 indicating major depression (Kroenke et al., 2001). Despite the focus on clinical diagnostics, the PHQ-9 has also been administered to non-clinical populations (e.g., Fang, Young, Golshan, Moutier, & Zisook, 2012), and when applied in that fashion we should *expect* it to produce systematically non-normal data with a positive skew. The non-depressed individuals should be contained within 33% of the scale, whilst the (much rarer) depressed individuals should be spread across 48% of the range. This is exactly what one expects of skewed distributions, and quite inconsistent with normality assumptions.

To see if there is empirical evidence that this skewness exists in real data, we looked at two studies that have reported PHQ-9 scores in non-clinical popilations. One reported a mean score of 6.89 and standard deviation of 5.31 (Fang, Young, Golshan, Moutier & Zisook, 2012). In keeping with what we found for the BDI, if these scores were truly normally distributed, then about 10% of the participants should have had negative scores. Given that the PHQ-9 has a floor of zero, it seems most likely that the data were non-normal and positively skewed. As a second example, Zivin, Eisenberg, Gollust & Golberstein (2009) used the PHQ-9 to estimate that somwhere between 12.93 - 15.36% of students were 'depressed'. Again, these statistics imply skewness: if 85% of students restricted to the bottom third of the scale, with the remaining 15% spread across the upper two thirds, it is very unlikely that the data are normal, and are much more representative of a positively skewed distribution.

As a final example we considered the 20-item Center for Epidemiological Studies Depression Scale (CESD; Radloff, 1977), a public domain self-report measure for depressive symptoms. On the surface, there are grounds to be more optimistic that CESD scores might be normally distributed in community samples, because the CESD was specifically designed to be useful and appropriate for non-clinical administration. However, the structure of the scale itself suggests some reasons to be cautious. Scores on the CESD range from 0 to 60, with a cut-off point for increased risk of depression of 16 (Radloff, 1991). Accordingly, the lower 23% of the scale describes depressive symptoms as they occur with a non-clinical population, with the remaining 77% describing possible depression (scores 17 through 28), and probable depression (28 through 60). As such, a substantial majority people should still be

Table 3.1: Differences and estimates among the four scales. Each scale has a reported clinical cut-off, above which scores indicate a clinical status and below which indicate non-clinical status. We used these cut-offs to understand the rough proportion of scale designed to describe the non-clinical population.

Scale	Intended use	Proportion describing non-clinical	Reported mean	Reported SD	Range	Skew?
BDI	Screening and assessment	14.3%	8.60	7.40	0 - 63	Likely
DASS (Depression)	Non-clinical screening	23.8%	5.67	6.72	0 - 42	Likely
PHQ-9	Screening and assessment	33.3%	6.89	5.31	0 - 27	Possible
CESD	Non-clinical population	25.0%	8.97 12.51	8.50 9.67	0 - 60	Likely

expected to score at the lower end of the CESD scale, suggesting that skewness will likely be an issue.

Turning to the empirical evidence, we note that when Radloff (1991) investigated the distribution of CESD scores in an adult population, they reported a mean of 8.97 and standard deviation of 8.50. If these data were normal one would require approximately 15% of adults to have CESD scores below zero, and again the most reasonable inference is that there is a strong floor effect with many participants scoring near zero, and a small number of participants scoring very highly: the data are skewed. This affect appears to be attenuated for their undergraduate population: a mean of 12.51 and standard deviation of 9.67 “only” requires 10% of the population to have impossible CESD scores in order to satisfy normality. As such it is possible that the skewness in the undergraduate population is reduced, but it is difficult to infer too much on the basis of such a cursory analysis.

Obviously, our investigation of the PHQ-9 and CESD scales were not as thorough as our evaluation of the BDI. However, the evidence that we do have all points in the same direction (see Table 3.1 for a summary). It would appear that the skewness problem is not peculiar to the BDI and DASS, but is a general phenomenon that should be expected when clinical instruments are applied to community populations. Moreover, it should be noted that this problem does not appear to be the consequence of "poor" scale design in the traditional sense of the term: all four instruments we have considered have well-demonstrated psychometric properties, but these desirable properties provide no protection against floor effects in non-clinical populations, and as such cannot provide safeguards against the skewness problem. This problem emerges from the very nature of the fact that the scales are designed to measure a (comparatively) rare thing. Most people are not depressed, but people who are depressed vary considerably in the severity of their depression. Any scale designed to measure depression that preserves these characteristics will tend to produced skewed data.

3.6 What statistical tools are currently used?

In the previous section we argued that skewed data should be the expected result whenever an instrument designed for clinical usage is administered to a mostly non-clinical population. Our survey of the literature on the BDI bears this out, with the

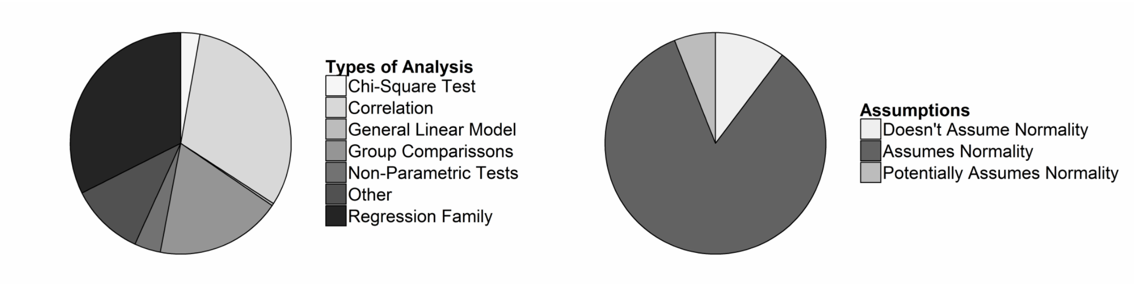


Figure 3.5: Distribution of statistical analyses applied to BDI data from non-clinical samples (left panel). Over three quarters of the analyses assumed that the data was normally distributed (right panel); this includes correlations, ANOVAs, and various forms of regression. Less than one-third of reported studies incorporated other analyses (e.g., non-parametric analyses, SEM, and so forth).

descriptive statistics reported in the relevant papers implying that skewed data is the norm, and a more cursory examination of the behaviour of other scales suggests that skewness is quite typical. However, although it seems clear that skewness is typical, it may not be obvious that this skewness presents a problem with the existing literature. In theory, it need not be problematic at all: the normal distribution is by no means universal in empirical data, and a great many statistical tools have been designed specifically to allow scientists to make inferences when normality is violated. Viewed from a statistical perspective there is no reason to be concerned with skewed data so long as these data are analysed with appropriate tools. With that in mind, our goal in this section is to survey the statistical tools that researchers have reported using to analyse their data.

In the first instance we return to our survey of the literature in order to examine what kinds of analyses are typically reported. To that end, we recorded every analysis that incorporated the BDI. As Figure 3.5 shows, the most common analyses are correlations, multiple regressions, t-tests and ANOVAs: over two-thirds of all reported analyses fell into these categories. All of them rely on some form of normality assumption, so prima facie it seems highly likely that in most of these cases the assumptions of the hypothesis test reported in the papers were violated.[5]

This pattern of results is somewhat worrying. Across the 148 papers we examined, we find pervasive evidence that skewness exists in the data alongside widespread usage of statistical analysis that are not designed for skewed data. Moreover, there is almost no reporting of the shape of the distribution (121 papers) or any discussion of the assumptions of the statistical tests (118 papers). Given this, we turn our attention to the effect of skewed data on most common analyses found in the literature review.

3.7 Why does this matter in practice?

A pragmatic researcher might read the previous section and be unconcerned. Although the normality assumptions of common statistical tests are well known to most researchers, it is not unusual for people to rationalize the use of such tests even when the normality of the data seems unlikely. A common justification is the claim that with a large enough sample, violations to normality have only have a minimal effect on the outcomes. While this is not entirely unreasonable, it opens up

[5] Strictly speaking, it is possible that the raw data did not violate the assumptions of these tests. For example, it is technically possible to have a skewed dependent variable while having normally distributed residuals in a regression analysis (which is what linear regression actually assumes) if it turns out, for example, that the independent variables are also skewed in just the right way. However, it would be a very remarkable coincidence for this to happen on a regular basis. The more likely scenario is that the normality assumptions of the test were indeed violated by the data.

the question of just how large a sample size is required. To put it another way, how badly does skewness affect the performance of the statistical tests typically used in clinical psychology?

In order to investigate this, we look at the effect that skewness has on both Type 1 and Type 2 errors. Type 1 error captures the tendency to incorrectly reject the null hypothesis even if it is true: that is, the tendency to find a significant effect where there is none. Type 2 error reflects the opposite tendency – failure to find an effect when one exists. How likely are these errors when the underlying data are skewed and the test is inappropriate? We can address this question by simulating sample data with varying degrees of skewness[6]. Since the data are simulated, we know the "correct" outcome and the target effect size. Running the simulated data through the appropriate test will reveal how many times that test returns the result that we know is correct. We can also vary sample size to determine whether, as is often presumed, increasing sample size solves the problems associated with non-normal data.

3.7.1 Comparing means (*t*-tests)

We first simulate the situation that arises when there are two groups and the researcher wishes to investigate whether the population means of the two groups are the same. Typically one would use *t*-tests or ANOVAs to determine whether any difference between populations is signified, and with this in mind we ran a simula-

[6] All of the details from the simulations included in this paper, including data included and code to run, are contained within LK's github page https://github.com/lauken13/Community-Samples-and-Skewed-Data.

tion study based on this practice. One thousand data-sets were drawn from a G-H distribution, which is a distribution where a standard normal distribution is transformed using two parameters to change the degree of skew and kurtosis. We focused on the modifying the degree of skew using the 'g' parameter. For each simulated dataset we sampled two groups while manipulating the degree of skew of the data (keeping the population skew the same between the two groups) and the number of participants in each group. Once these were sampled, we added to one group the effect size (which was acceptable as the g-h distribution is a standard normal transformation). The three levels of effect size (small, medium and large) correspond to a pre-transformation Cohen's d of 0.3, 0.5 and 0.8. The aim of these simulations were to answer two main questions: (a) the effect of skew on the power and Type 1 error rate of the test, and (b) whether any loss of power or increase in Type 1 errors could be rectified by simply increasing the sample size.

In Table 3.2 we describe the sample skew characteristics for the first and second groups in the t-tests. Although we expect the skew of the our scales to correspond most closely to the "Medium Skew" condition, it could very well fall in the "High Skew" condition. The "Normal" condition demonstrates that even if the underlying distribution was *not* skewed it is entirely possible to have a sample with a skew as high as 2.0. As we have rationalised before, we believe that the underlying distribution of the is truly skewed, indicating the possible levels of skew could be very high.

Figure 3.6 shows the results of these simulations. Somewhat reassuringly, it suggests that Type 1 error – incorrectly rejecting the null hypothesis when it should be

Table 3.2: Sample skew for group 1 and group 2 in the simulations investigating the effect of skew in the t-test, broken down by simulated skew condition.

Distribution	G-H Dist.	Group 1		Group 2	
	G value	Range	Mean(SD)	Range	Mean(SD)
Normal	0.00	(-2.0, 2.0)	0.00 (0.28)	(-2.13, 1.87)	0.00 (0.28)
Low Skew	0.20	(-1.7, 2.27)	0.52 (0.34)	(-1.65, 3.18)	0.52 (0.34)
Medium Skew	0.50	(-1.93, 8.24)	1.37 (.65)	(-1.58, 9.18)	1.38 (0.64)
High Skew	1.50	(-0.73, 22.11)	4.99 (3.18)	(-0.99, 22.02)	4.99 (3.18)

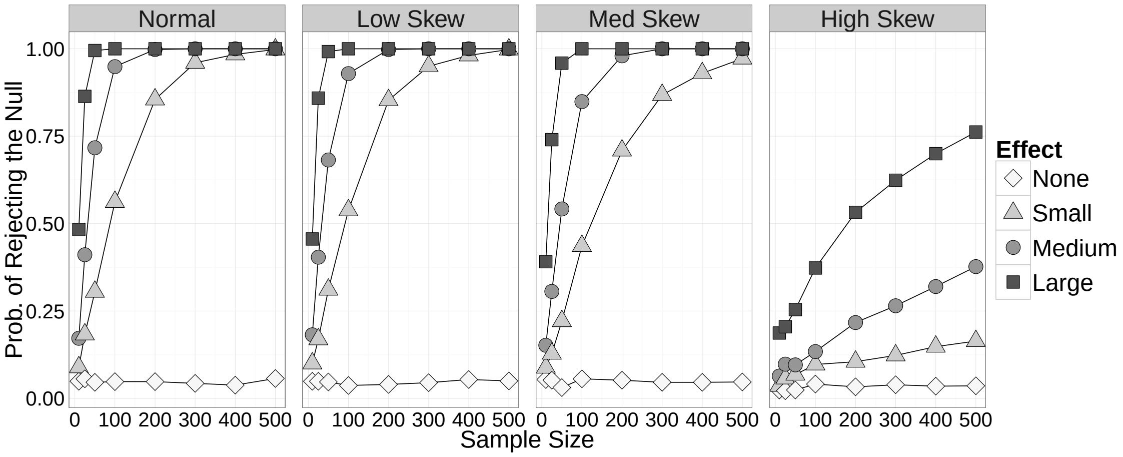


Figure 3.6: Error rates with a t-test as a function of sample size, skewness, and effect size. The y axis shows the proportion of time t-tests gave the correct analysis for our simulated data. In the simulations in which there was truly no effect, a correct analysis would be to retain the null. When there was a small, medium, or large effect, a correct analysis would be to reject the null. The diamond markers indicate that when there is no effect, t-tests typically correctly retain the null: as a result, Type 1 error is low. However, the other markers show that when there is an effect, t-tests often incorrectly report that there is none (i.e., Type 2 error is high) when the data are skewed. The effect of skew is demonstrated by change in shape of the lines across the panels. Where the lines drop lower, there is an effect of skew. This is the case even for large effect and sample sizes if the skew is high enough.

retained – is low even when the data are highly skewed and the sample size is small. However, most *unreassuringly*, it suggests that Type 2 error – incorrectly retaining the null hypothesis when it should be rejected – is high. If the dataset is highly skewed, even when there is a large effect size and $N = 500$, the null hypothesis is incorrectly retained a full 25% of the time; put another way, 25% of these studies would erroneously conclude there is no effect when in fact there is a small one. The problem becomes even more extreme when the samples are a more typical 50-100. In that case, a typical t-test on skewed data will have a Type 2 error 75% of the time, almost double that of one performed on normally distributed data. In other words, more often than not the test will suggest that there is no effect when in fact there is one.

Why does this occur? What effects do these deviations from normality mean? One implication is that when data are skewed, the mean is no longer a measure of central location: the presence of a small number of extreme values pulls the mean higher than the median. Another implication is that when the underlying distribution is skewed, there is a higher variance in the samples. This is because some samples may (just by chance) contain one or more extreme points, while others may not. As a result, variances and outcomes may differ widely, especially when sample size is small. The effect size, as measured by Cohen's d, decreases as we increase the skew of the groups (through transformation). The groups are then much less likely to be found to be significantly different, even though they were designed and simulated to be so. Ranking measures like the Mann-Whitney U test are invariant to these changes.

These results so far suggest that the typical test used to compare two means – t-tests – dramatically lose power and hence increase Type 2 errors when there is a

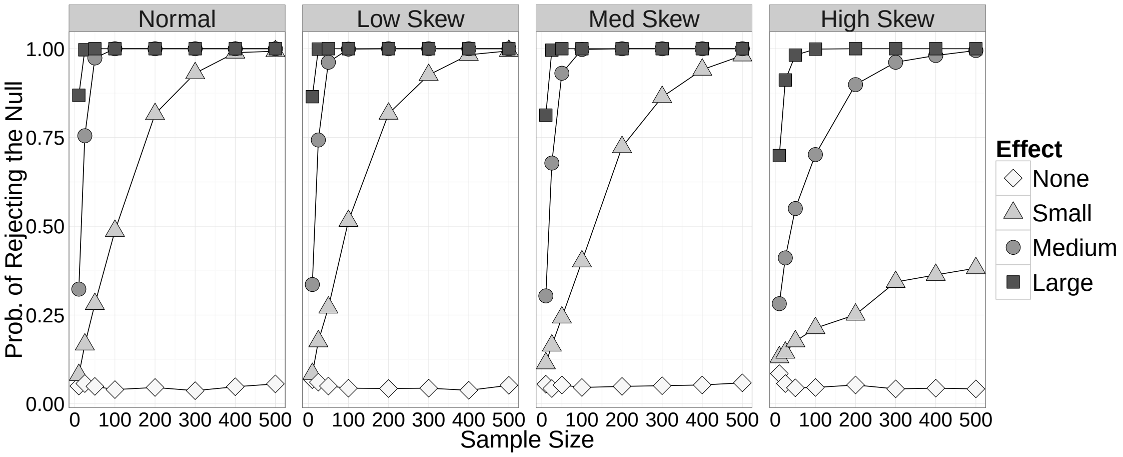


Figure 3.7: Error rates within the linear model as a function of sample size, skewness, and effect size. The y axis shows the proportion of time the linear model (i.e., a simple regression with a single predictor) gave the correct analysis for our simulated data. When there was truly no effect, a correct analysis would be to retain the null. If there was a small, medium, or large effect, a correct analysis would be to reject the null. The diamond markers indicate that when there is no effect, analyses based on the linear model typically correctly retain the null: as a result, Type 1 error is low. However, the other markers show that when there is an effect, the linear model often incorrectly reports that there is none (i.e., Type 2 error is high) when the data is skewed. While this finding is less severe than that found for t-tests, the Type 2 rate is still substantial even when the effect size and sample size are large, as long as there is sufficient skew.

high degree of skew in the data. However, our literature review revealed a heavy reliance on not just t-tests, but another family of tests as well – tests involving linear regression.

3.7.2 Inferring relationships (correlations and linear regression)

This group of tests focuses on determining whether an increase in a variable – or more than one variable if we consider multiple regression – suggests another variable will not change (null hypothesis) or will change in a linear manner (alternative hypothesis). This family of tests includes correlations and regressions, all of which rely on a common set of assumptions and methods. What is the outcome when these tests are applied to skewed data?

We investigate this question as we did with *t*-tests, by simulating sample data with varying degrees of skewness. In our simulations we knew the "ground truth" (i.e., whether there was an effect, and if so how large) and evaluated whether the model reported the correct analysis. For simplicity we consider only simple regressions with a single continuous predictor (see Appendix 3.11.2 for additional detail). As before, we wished to investigate the effects of sample size and the amount of skewness in the data. To that end we manipulated the strength of the relationship between the two variables, as measured by Pearson's r, as well as the sample size and degree of skew. Similar to before with Cohen's d, the small, medium and large effect sizes correspond to a correlation of .3, .5 and .8. In Table 3.3 we describe the sample skew characteristics for the first and second observations in the correlations. As with the t-tests, 1000 data-sets were used for the simulation study. Unlike the t-tests (where the effect size was added after transformation) here we simulated two correlated variables using the mvrnorm function in the MASS package in R, and then transformed each independently using the g-h transformation to obtain the same distribution shape as before.

The results are shown in Figure 3.7, and are very similar to those from the previous study. Type 1 error rates remain low even if the data are highly skewed, but there is a substantial increase in Type 2 errors when there is high or even mild skew.

Table 3.3: Sample skew for the first and second observation in the simulations investigating the effect of skew in the simulated simple linear regression, broken down by simulated skew condition.

Distribution	G-H Dist.	Group 1		Group 2	
	G value	Range	Mean(SD)	Range	Mean(SD)
Normal	0.00	(-2.11, 1.87)	0.00 (0.28)	(-1.98, 1.92)	0.00 (0.28)
Low Skew	0.20	(-1.86, 3.32)	0.52 (0.34)	(-1.82, 3.35)	0.52 (0.34)
Medium Skew	0.50	(-1.69, 10.07)	1.37 (.65)	(-1.41, 9.98)	.1.38 (0.64)
High Skew	1.5	(-.91, 22.14)	4.99 (3.19)	(-0.68, 22.00)	4.98 (3.18)

When the effect size is small and the data are highly skewed, even sample sizes of up to 500 return the incorrect conclusion more often than not. As before, there are many reasons for the poor performance when the data is skewed. One important factor is that skewed datasets tend to include many outlier points that lie far from the main cluster of responses, and these can have a disproportionate effect on the inferred relationship. Specifically, this occurs when they have *high leverage*: the lack of nearby data points means that they have a large effect on the final result. This effect is shown with the two datasets depicted in Figure 3.8. Both datasets have exactly the same points except the one on the right has one additional outlier. This single data point changes the correlation from a moderate, positive one ($r = 0.433$) to a negligible, negative one ($r = -0.144$). As skew becomes more pronounced, the presence of high-leverage outliers like this increases.

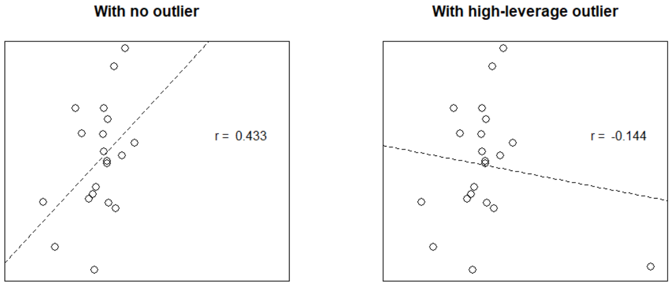


Figure 3.8: Graphical illustration of the problem with a single high-leverage outlier. The datasets in both panels are identical except that the one on the right contains one outlier. This single point has a dramatic effect on the line of best fit and the overall correlation, altering it from a correlation of moderate size ($r = 0.433$) to one of negligible size ($r = -0.144$).

3.7.3 Summary

As this section illustrates, one of the most pervasive problems that skewness introduces is a loss of power when the data are analysed using tools that rely on a normality assumption. From a practical standpoint this is troubling, because it undermines the sample size calculations that researchers typically use to work out how many people they need to recruit, and distorts the estimates of the power that researchers report in the litetature. If we have a sample of 100 individuals drawn from two normally distributed groups with a large difference between them, we would have a very high probability (95% or higher) of detecting that this difference is present (good power). However if these two groups had a high skew, then we would detect the same difference in only 60% of the trials. The need for an larger sample sizes when dealing with skewed data is discussed by Cundill and Alexander (2015), who provide methods for calculating the required sample size for negative binomial and poisson distributions. More generally, if it is indeed the case that most studies have much lower power than they appear to have based on the usual methods for sample size calculation, the rate of false positives in the literature is likely to be much higher than one would hope for, contributing to the poor replication rate that been reported in empirical investigations (Open Science Collaboration, 2015).

3.8 What solutions are available to researchers?

The previous sections describe the prevalent use of statistical tests with assumptions of normality, despite research designs in which the nature of the scales used dictates that the distribution is likely to be skewed. One of the biggest resulting issues is the lack of power, even with extremely large sample sizes. In this section we consider possible remedies to the problem.

3.8.1 Outlier removal

A straightforward solution to the problem of skewed data is to simply remove outliers from analysis. This method is appropriate if the research question is focused on non-clinical individuals, because in such cases the outliers typically reflect participants in the clinical range and may not be the most relevant to the research problem being addressed. Identifying these outliers is sometimes a challenge, but in many cases the scale itself provides information about the severity of symptoms and it is straightforward to simply set a exclusion threshold based on that (e.g., removing all participants whose scores put them in the ‘severe’ range). One could also remove outliers statistically, either by examining boxplots (as in Hubert and Vandervieren (2008)) or using the MAD-median rule (Davies & Gather, 1993).[7] Regardless of how outliers are excluded, outlier exclusion in general is a valid method as long as the exclusion criterion is set before the data are analysed and there is a theoretical (i.e., psychological) motivation for removing the extreme values.

One of the biggest difficulties with the outlier removal method is the fact that in many cases there *is* no psychological justification for outlier removal. In fact, many studies are interested in investigating the full range of variability across a population, and in most cases that necessarily includes a small proprtion of people who fall within the clinical range. In such cases, one has no psychological reason for exluding those individuals, or any strong statistical justification for doing so *except* for the fact that those participants introduce non-normality to the sample. If one ends up excluding data solely for the sake of statistical convenience, some concerns would seem to be warranted. To address this in a more principled way, it seems wiser to make use of statistical methods that are *robust* to the presence of skewness. With this in mind we consider some of the main approaches.

3.8.2 Comparing groups (t-tests)

We first consider the problem of comparing group means, which typically requires a t-test. As we saw in Figure 3.6, these tests lead to a high Type 2 error rate when the data are skewed, leading to a substantial decline in power. One solution to this is Yuen's t-test (Yuen, 1974). It aims to overcome the problems caused by skewed

[7] Both of these methods are supported through the statistical package WRS developed by Wilcox and Schönbrodt (2015).

data by comparing the 20% trimmed mean rather than the means, as is traditional. By doing so it excludes the most extreme 20% of the data before calculating the means. The other solution we consider is the Mann-Whitney U test, which operates on ranks rather than the raw data values. Each data point is replaced with its rank (e.g., the highest depression score in the sample would get the highest rank score) and the ranks are compared. This avoids the problems associated with skewed data because it removes information about the distances between points.

In order to investigate whether either of these alternatives overcomes the problems associated with the traditional t-test, we replicated the simulations in Figure 3.6 using Yuen's t-test and the Mann-Whitney U test. The results, shown in Figure 3.9, demonstrate that both represent a considerable improvement. When there is a high amount of skew the traditional t-test usually has a Type 2 error rate of about 75%, even when the sample size is extremely large. By contrast, both Yuen's t-test and the Mann-Whitney U have a Type 2 error rate of less than 5% when the sample size is large enough, even when the effect size is small. The Mann-Whitney U is slightly more effective, performing just as well across the board on highly skewed data as on data with no skew at all.

3.8.3 Inferring relationships (correlation and linear regression)

These result so far are reassuring, but are applicable only to comparisons between groups. Often researchers may be interested in evaluating linear relationships including correlations or regression. This also leads to more complicated adaptions

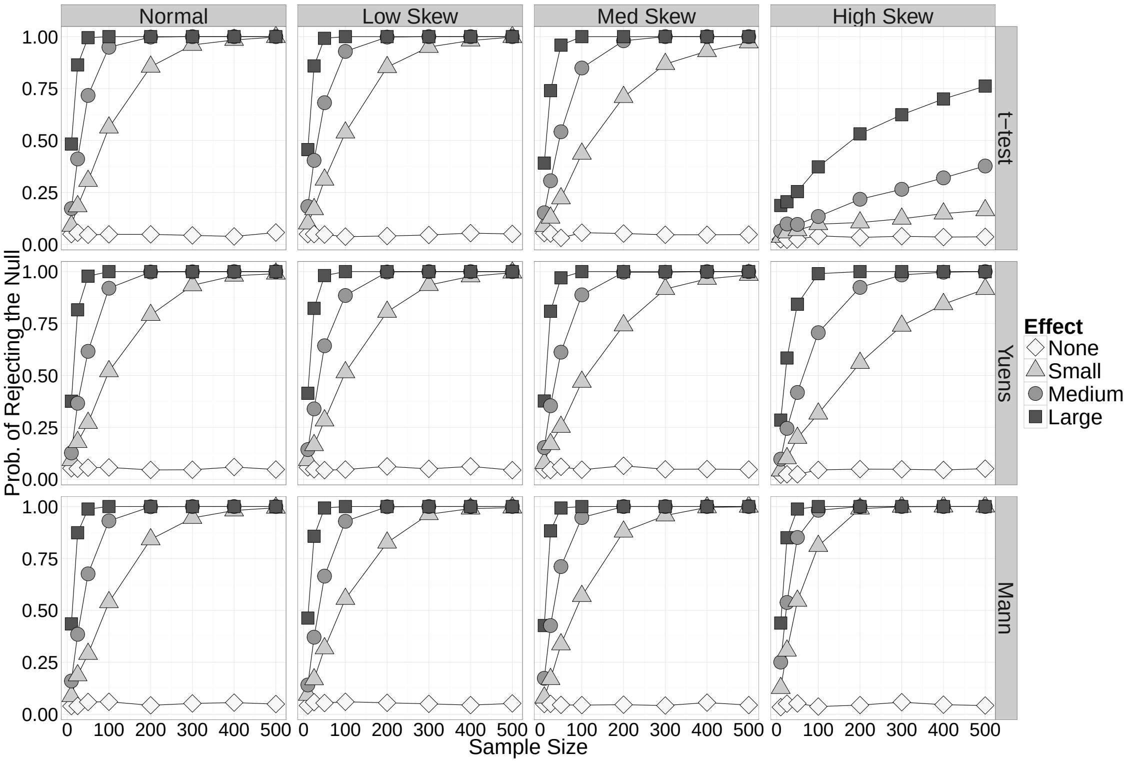


Figure 3.9: Simulation results comparing the performance of two alternatives to the traditional *t*-test: Yuen's *t*-test and the Mann-Whitney U. As before, the y axis captures the proportion of time the test gives the correct answer (i.e., accepting the null when there is no effect and rejecting it when there is one). Both alternatives represent a considerable improvement when the data is skewed, achieving less than 5% Type 2 error rates with large enough sample size. The Mann-Whitney U in particular performs as well on highly skewed data as it does on data with no skew at all.

of regression like structural equation modelling and path analysis. As we saw in Figure 3.7 these approaches have high Type 2 error rates when applied to skewed data. This data has an increased probability of including outliers have high leverage, which can have a strong effect on the outcomes of the test. But what are the alternatives? We consider two distinct ways of counteracting this problem. The first is Spearman's rho, a simple method that ranks the raw data and then takes a correlation of the ranked values. It is a non-parametric alternative to Pearson's r (Spearman, 1904b). The other alternative we consider is HC4 regression (Godfrey, 2006). This form of regression that down-weights high leverage points so that they have less ability to influence the parameter estimates.

In order to investigate whether either of these alternatives overcomes the problems associated with a simple linear model, the simulations from Figure 3.7 were replicated with Spearman's correlation and HC4 regression. The results, shown in Figure 3.10 suggest that while Spearman's method provides improvement, HC4 regression is not much different from the traditional method. When the data are highly skewed, the linear model could detect a small effect only about 40% of the time, even when the sample size was extremely larger. In contrast, Spearman's rho detected even a small effect with remarkable accuracy regardless of the skew of the data. In hindsight this is unsurprising: HC4 regression is designed to reduce the effect of high leverage points, but does not account for a decrease in information in the higher reaches of the scale.

3.8.4 Inferring relationships (multiple regression)

These results suggest that there are promising alternatives to simple linear regression and correlation. However often multiple regression is the chosen tool of choice. How well do these more complicated models identify the relationship between a skewed predictor and outcome variable when other non-skewed and non-related predictors are included in the model? In order to investigate this we replicated the simulations from Figure 3.7 were replicated in a similar fashion, but we also included three additional predictors that had no relationship with the outcome or predictor variable. As with the simulations in Figure 3.10, we included both a linear model

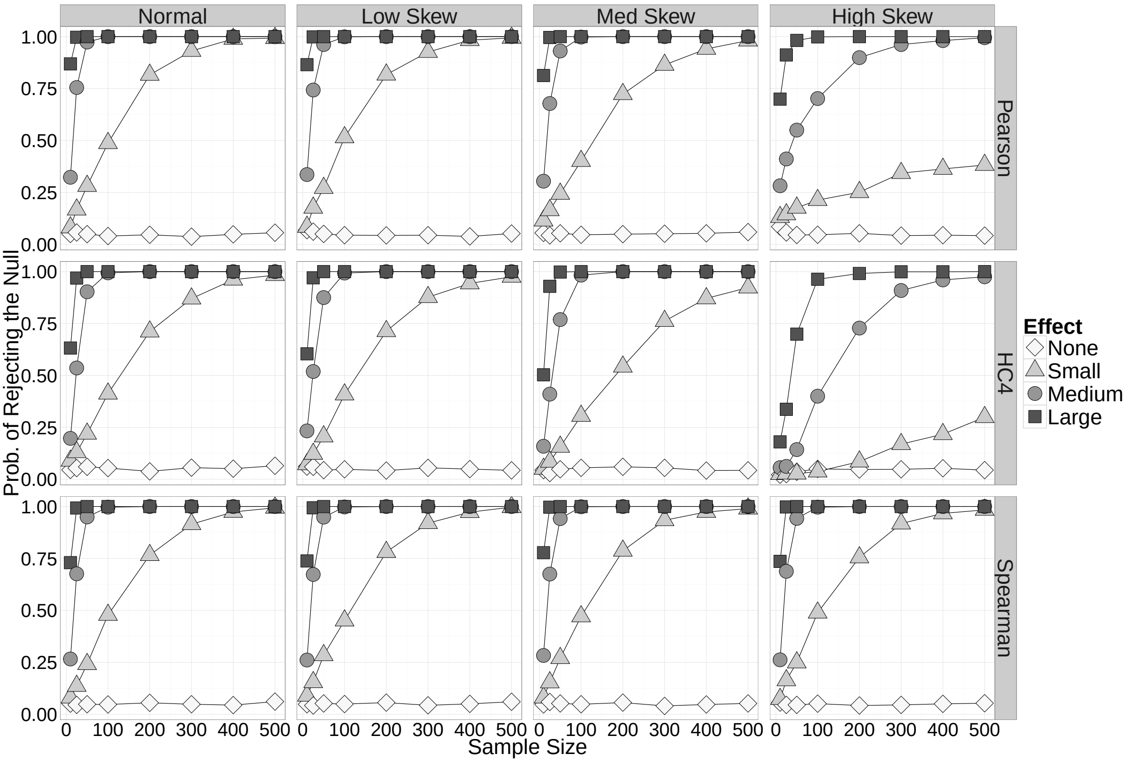


Figure 3.10: Simulation results comparing the performance of two alternatives to the traditional simple linear regression: Spearman's rho and HC4 regression. As before, the y axis captures the proportion of time the test gives the correct answer (i.e., accepting the null when there is no effect and rejecting it when there is one). Unlike the simulations demonstrated with the t-test, only one model provides improvement when the data is skewed. Whilst the Spearman correlation provides a similar level of power regardless of skewness, the HC4 regression does not improve on the traditional linear model.

and the HC4 adaption of a linear model. As Spearman's correlation is not suited for regression with multiple predictors, we relied on an extended model for ranked regression (Kloke & McKean, 2012) that is designed in a similar fashion to Spearman's rho, but allows for multiple predictors. In Table 3.4 we describe the sample skew characteristics for the first and second observations in the multiple regression.

The results of these simulations can be seen in Figure 3.11. As the simulations now have more than one predictor, we will focus on investigating whether the skewed

Table 3.4: Sample skew for outcome and skewed predictor in the simulations investigating the effect of skew in the multiple regression, broken down by simulated skew condition.

Distribution	G-H Dist.	Group 1		Group 2	
	G value	Range	Mean(SD)	Range	Mean(SD)
Normal	0.00	(-2.12, 2.03)	0.00 (0.28)	(-2.46, 1.91)	0.00 (0.28)
Low Skew	0.20	(-1.71, 2.90)	0.52 (0.34)	(-1.65, 3.24)	0.52 (0.34)
Medium Skew	0.50	(-1.90, 8.98)	1.38 (0.65)	(-1.43, 8.77)	1.38 (0.64)
High Skew	1.5	(-.74, 21.96)	4.99 (3.19)	(-0.78, 22.17)	5.00 (3.21)

predictor was (when it was a a significant predictor) correctly identified as a significant predictor (i.e. the power of each method) or (when it was not a significant predictor) correctly identified as not a significant predictor (i.e., the Type 1 error rate of each method[8]). Like before, the small, medium and large effect sizes correspond to a correlation of 0.3, 0.5 and 0.8. As with the simulations for simple linear regression presented in Figure 3.10, both the traditional linear model and HC4 regresion perform quite poorly at identifying small effect sizes in high skewed data. However, similar to Spearman's correlation in the previous simulations, ranked regression performed quite well, correctly identifying about 80% of small effect sizes with a large sample size and highly skewed data.

[8] We could have chosen to focus on identifying whether the collection of predictors explained a significant amount of variance in the outcome variable through R^2 and its corresponding significance test, but for simplicity focused only on the relationship between the skewed predictor and outcome variable given non-related predictors.

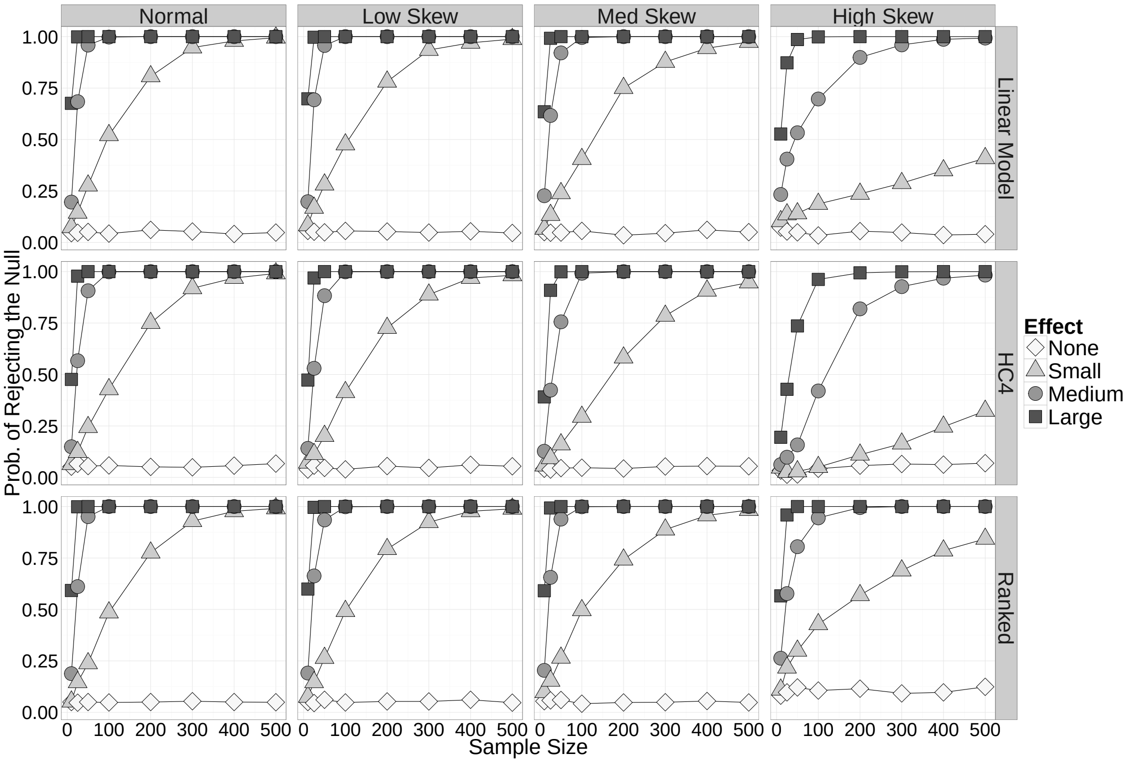


Figure 3.11: Simulation results comparing the performance of two alternatives to the traditional multiple linear regression: ranked regression and HC4 regression. As before, the y axis captures the proportion of time the test gives the correct answer (i.e., accepting the null when there is no effect and rejecting it when there is one). Similar to the simple linear regression results in Figure 3.10, only one model provides improvement when the data is skewed. Whilst the ranked regression was a significant improvement, the HC4 regression is no real improvement on the traditional linear model.

3.8.5 Summary

We considered three common analyses that, from our literature review, are commonly used with skewed data. The t-test was the most sensitive to skew, which resulted in less power at identifying small, moderate and even large effects. While the Yuen t-test improved performance, the Mann-Whitney U test was superior overall. This general pattern was mirrored in our simulations based on regression analyses: approaches based on ranks (e.g., Spearman's rho) outperformed standard linear regression, whereas other methods such as HC4 regression did not.

3.9 Discussion

The sensitivity of scales to the spectrum of non-clinical symptoms provides a unique challenge to our normality reliant statistical models. In this manuscript we present a non-clinical sample of scores as they fall on the DASS depression subscale, a literature review that highlights the population mean and standard deviation of the scores of a non-clinical sample on the Beck Depression Inventory, and extended this logic to two other scales, the CESD and PHQ-9. We summarised these findings in Table 3.1.

We argue that this is a necessary characteristic of the scales. Their central design purpose is to screen for and describe depression in a clinical population. For the purpose they were designed for they perform exceptionally well. However, we argue that it can be problematic when they are applied a non-clinical population and traditional statistics are used to infer relationships and differences (as is commonly done).

This issue predominately does not cast doubt on the significant findings. Through simulation studies we demonstrated that the type 1 error rate (the likelihood of a model to falsely find an effect when in fact there is none) was constant across different levels of skew. However, as the distribution of the sample data grows more skewed, the power of the traditional methods (the ability to detect an effect when there is indeed an effect present) is significantly hampered.

If traditional power calculations cannot be directly used with skewed data, then researchers can no longer trust their estimations of whether they have sufficient a sample size to detect the effect size they expected to find. Because of this, they are less likely to replicate their findings in replications. If this is a global trend, then it decreases the trustworthiness of the research findings overall. With these issues in mind, we sought to find alternative methods that were less sensitive to the shape of the data.

The first method to increase the power of the more traditional methods is to attempt to reduce the degree of skew. One of the most common methods to achieve this is to remove outliers. Removing only a few extreme points in a highly skewed distribution can reduce the uncertainty around parameter estimates and hence increase the power. We discussed methods that are appropriate for extreme outlier removal in a skewed distribution. However, we also acknowledge that often the entire sample is of interest. In these times it is undesirable to remove extreme points because the act of doing so reduces the overall information available.

The second method is to use methods that aim to make a similar inference to the traditional methods, but does not rely on assumptions of normality. The first methods we investigated were alternatives to the traditional two sample t-test. In this instance a method based upon ranking data, the Mann-Whitney U test, produced the best improvement in power. A similar preference for ranking methods was found for simple linear and multiple regression techniques.

One potential limitation to the evidence for the ranking methods advocated for in this paper is that transformation methods were not considered. Part of the reason

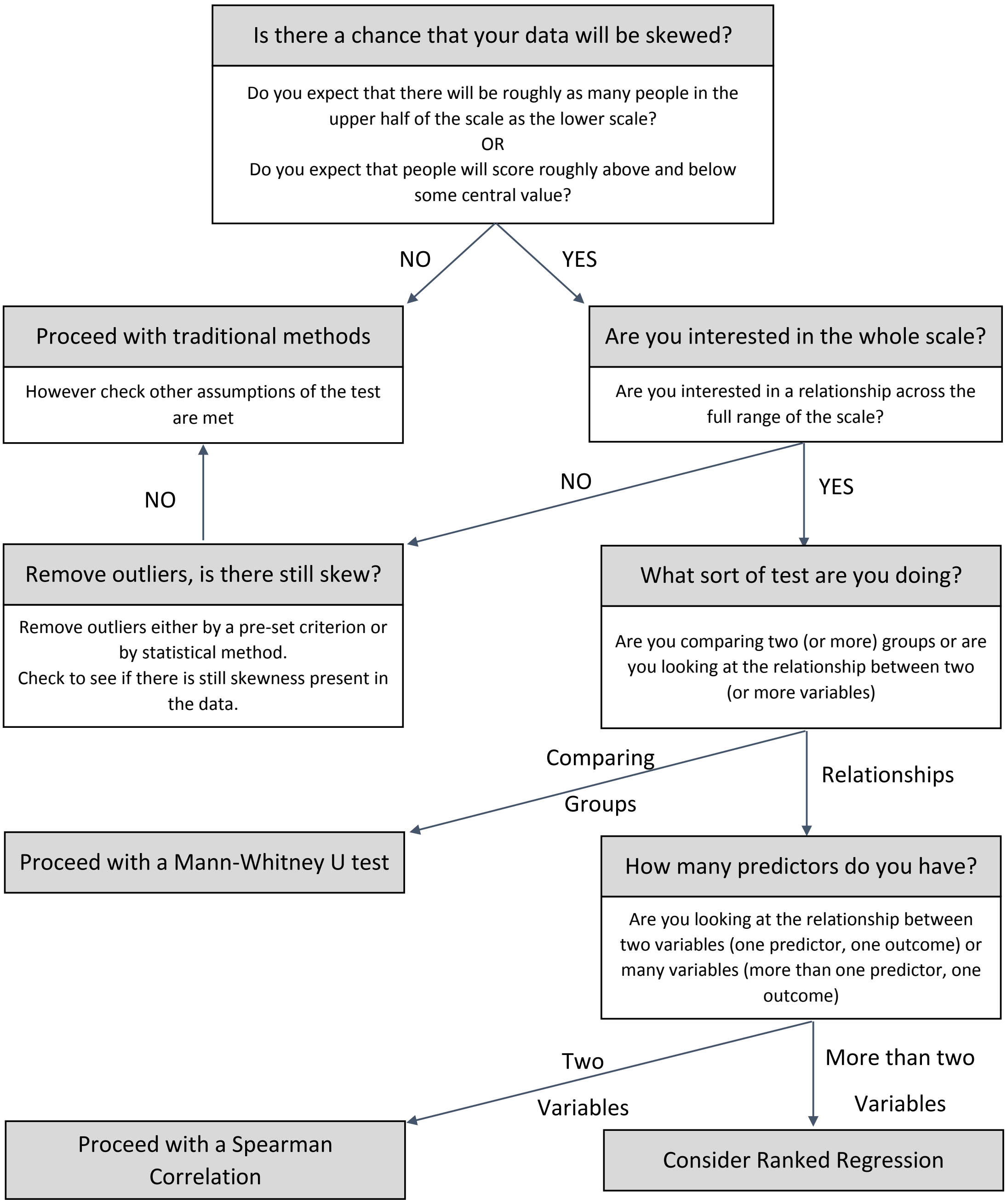

Is there a chance that your data will be skewed?
Do you expect that there will be roughly as many people in the upper half of the scale as the lower scale?
OR
Do you expect that people will score roughly above and below some central value?
NO
YES
Proceed with traditional methods
However check other assumptions of the test are met
Are you interested in the whole scale?
Are you interested in a relationship across the full range of the scale?
NO
NO
YES
Remove outliers, is there still skew?
Remove outliers either by a pre-set criterion or by statistical method.
Check to see if there is still skewness present in the data.
What sort of test are you doing?
Are you comparing two (or more) groups or are you looking at the relationship between two (or more variables)
Comparing Groups
Relationships
Proceed with a Mann-Whitney U test
How many predictors do you have?
Are you looking at the relationship between two variables (one predictor, one outcome) or many variables (more than one predictor, one outcome)
Two Variables
More than two Variables
Proceed with a Spearman Correlation
Consider Ranked Regression

Figure 3.12: Describes the steps required to identify situations where skewed data might be likely, and the methods one might take to analyse skewed data to improve power.

for this was the method we chose for our simulations. All simulation data was created through a transformational method, where data were created as normally distributed data with a given effect size, and then transformed to have the correct degree of skew. We chose this method because it best reflects how a normally distributed sample could become skewed with a scale that is maximally informative for only a few rare (in a non-clinical sample) individuals. It also allowed us to theoretically manipulate the skew and effect size in our simulation without relying on the (normal assuming) analytic methods. As the data were simulated through transformation, it difficult to compare the transformation without the simulation method biasing the results. Transformation of the data is generally completed on a case-specific manner as there is not a one-size fits all solution.

In this paper we reviewed the distribution of clinically aimed scales when used with non-clinical samples. We discussed that the very nature of these scales implied that they would produce skewed data to some degree on with a non-clinical sample. This isn't necessarily an issue, but we showed that the commonly used traditional statistical techniques were considerably less powerful than presumed with this data. One method to mitigate these issue is to use ranking methods like the Mann-Whitney U test (for groups) or Spearman or ranked regression techniques (for relationships). We demonstrated the improved power of these techniques through simulation studies. We refer readers to then flow chart presented in Figure 3.12 for a summary of the findings presented here.

We sought to examine the increase in power that could be obtained from a relatively small change of analysis method. We feel that this change means that both the analysis and interpretation is relatively easily understood and implemented, which increases the ease in which researchers might move to this method. However, this benefit is also one of the major limitations of this manuscript. We did not consider more complex methods (such as M-estimators, quantile differences test, quantile-quantile regression and members of the generalised linear model family like the log normal, gamma and Weibull), nor do we consider the relative benefits of these methods when compared to more simple rank methods for this problem. We also do not consider scale-specific approaches like boundary inflation, nor item specific relationships, which we agree is a both a limitation and a potential area for future research.

Also worth mentioning is the limitations of ranked methods themselves. One of the challenges for these methods is where multiple points have the same rank (i.e., there are ties). There exists a diverse literature to account for these ties when calculating uncertainty on the estimator (which is used for hypothesis testing), but there will inevitably be a point where there are too many ties to make meaningful inference. Future work should investigate where this point is for scales of this type, where there is good reason to suspect a higher proportion of ties in the lower parts of the scale.

Overall we wish this paper to be a cautionary note not about not using scales such as the BDI, DASS, PHQ-9 etc (which we believe in general are very good scales). Instead we wish to emphasise that the relatively simple substitution of rank order methods for more traditional statistical methods allows significantly greater power than traditional linear regression and t-tests.

3.10 References

Beck, A. T., Rial, W. Y., & Rickels, K. (1974). Short form of depression inventory: Cross-validation. *Psychological Reports*, *34*, 1184–1186.

Beck, A. T., Steer, R. A., Brown, G. K., et al. (1996). Beck Depression Inventory-II. *San Antonio, TX: Psychological Corporation*, 78204–249.

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*(1), 77–100.

Beekman, A., Copeland, J., & Prince, M. J. (1999). Review of community prevalence of depression in later life. *The British Journal of Psychiatry*, *174*(4), 307–311.

Cundill, B. & Alexander, N. D. (2015). Sample size calculations for skewed distributions. *BMC Medical Research Methodology*, *15*(1), 1–9.

Davies, L. & Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association*, *88*(423), 782–792.

Fang, D. Z., Young, C. B., Golshan, S., Moutier, C., & Zisook, S. (2012). Burnout in premedical undergraduate students. *Academic Psychiatry*, *36*(1), 11–16.

Ghassemzadeh, H., Mojtabai, R., Karamghadiri, N., & Ebrahimkhani, N. (2005). Psychometric properties of a Persian-language version of the Beck Depression Inventory-Second edition: BDI-II-PERSIAN. *Depression and Anxiety*, *21*(4), 185–192.

Godfrey, L. (2006). Tests for regression models with heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, *50*(10), 2715–2733.

Hoaglin, D. C., Mosteller, F., & Turkey, J. (1985). *Exploring data tables, trends, and shapes.* Hoboken, NJ: John Wiley & Sons.

Hubert, M. & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, *52*(12), 5186–5201.

Kashani, J. H., Carlson, G. A., Beck, N. C., Hoeper, E. W., Corcoran, C. M., McAllister, J. A., ... Reid, J. C. (1987). Depression, depressive symptoms, and depressed mood among a community sample of adolescents. *American Journal of Psychiatry*, *144*(7), 931–934.

Kloke, J. D. & McKean, J. W. (2012). Rfit: Rank-based estimation for linear models. *The R Journal*, *4*(2), 57–64.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613.

Lovibond, S. H. & Lovibond, P. F. (1993). *Manual for the Depression Anxiety Stress Scales (DASS).* Sydney, NSW: Psychology Foundation of Australia.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

R Core Team. (2015). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401.

Radloff, L. S. (1991). The use of the Center for Epidemiologic Studies Depression Scale in adolescents and young adults. *Journal of Youth and Adolescence*, *20*(2), 149–166.

Shek, D. T. (1990). Reliability and factorial structure of the Chinese version of the Beck Depression Inventory. *Journal of Clinical Psychology*, *46*(1), 35–43.

Spearman, C. (1904b). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.

Ward, L. (2015). Health and ageing. *[Unpublished raw data].*

Wiebe, J. S. & Penley, J. A. (2005). A psychometric comparison of the Beck Depression Inventory-II in English and Spanish. *Psychological Assessment*, *17*(4), 481.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Amsterdam: Academic Press.

Wilcox, R. R. & Schönbrodt, F. D. (2015). *The WRS package for robust statistics in R (version 0.27.5).* Retrieved from https://github.com/nicebread/WRS

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*, *61*(1), 165–170.

3.11 Appendix

3.11.1 A. Details of our literature search

The first author, LK, undertook an initial literature search of PubMed, PsychInfo and Embase to see how common the Beck Depression Inventory is in the literature. The search terms for these three databases is included in full in Table 3.5.

Database	Scale	Students	Language	Year
PubMed	beck depression[tw] OR BDI[tiab]	Students[mh] OR student*[tiab] OR schools[mh] OR universit*[tiab] OR faculty[mh] OR college*[tiab] OR undergraduate*[tw] OR (nonclinical[tiab] AND adolesce*[all])	english[lang]	Filters: published in the last 5 years
PsychInfo	beck depression inventory.sh OR beck depression inventory.mp OR BDI.mp	educational degrees.sh or college graduates.sh or high school graduates.sh or higher education.sh OR colleges.sh or schools.sh or community colleges.sh or college environment.sh or higher education.sh or military schools.sh or universit*.mp or student*.mp	limit 9 to english language	10 and 2009:2014. (sa_year).
Embase	'beck depression inventory'/exp/mj OR 'beck depression inventory'	'student'/exp	[english]/lim	[2010-2014]/py

Table 3.5: Literature search terms, broken down by database

3.11.2 B. Details of simulation studies.

All of the simulations presented in this paper were conducted in similar manner. The simulations can be divided into those that dealt with group comparisons and those that dealt with relationships. For each simulation two samples with a desired correlation or cohen's d and shape were created (details below). The test of interest was then completed. For each simulation this procedure was repeated 1000 times. Details of this protocol are described in this section.

Each reported sample size was the sample of individuals in each group. For example, in a two-sample t-test, a sample size of 10 meant that there are 10 individuals in each group. For the simple correlations described, a sample size of 10 means that there were 10 individuals who each gave two values.

Two samples were created with a given effect size. For the simulations that were focused on comparing two means, first a sample was drawn from a population with a given shape and then sampling from a second distribution with the same shape as the first but displaced by the effect size (as the distributions were standard normal). The simulations linear regression/correlation used a Cholesky-Square Root decomposition to generate two normal variables with a known correlation, and then transforming these two variables with a g-h transformation.

We relied on a transformation method to create an appropriate level of skew. The distribution was created with a g-h distribution (Hoaglin et al., 1985). This distribution can be used to transform a standard normal distribution to be positively skewed to varying degrees. This distribution can also be used to vary the heaviness of the tail (or kurtosis). For the purposes of the simulations presented in this manuscript, the parameter controlling kurtosis was kept constant. The distributions of the low skewed (g=0.2), medium skew (g=0.5) or highly skewed (g=1.5) were used. They are shown in Figure 3.13.

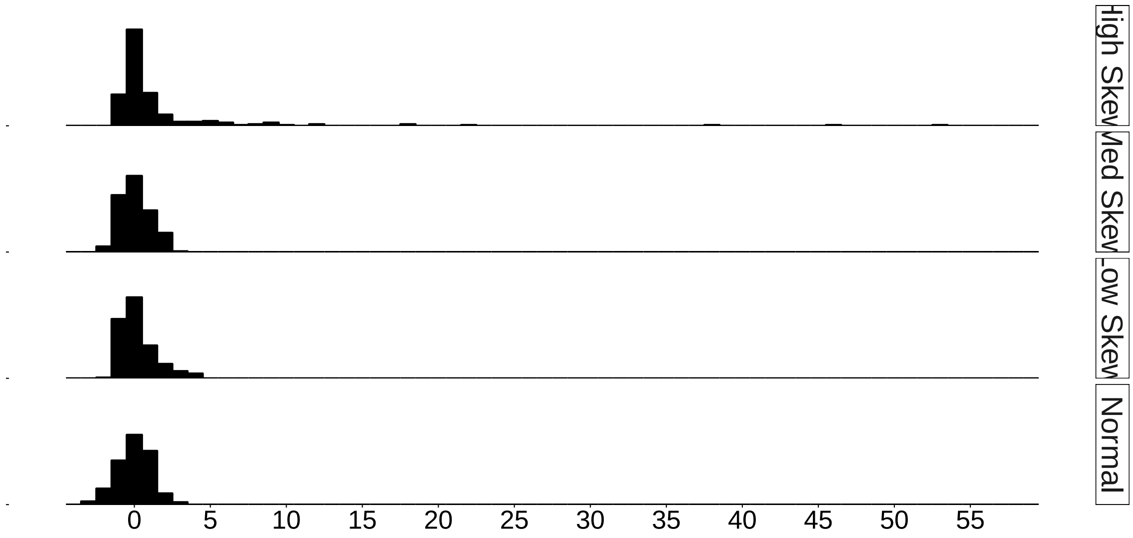


Figure 3.13: This figure demonstrates the three different distributions used in many of the simulations described in this paper. These distributions are drawn with sample of 250 data points. The high skew condition has a much higher probability of outliers.

This method of generating data was chosen for two reasons. The g-h distribution is often used when investigating robustness of statistical tests (Wilcox, 2012). This means that this research fits within a broader research programme. Additionally the g-h distribution can be used to increase the degree of skew in increments to allow for testing.

However, there are some limitations with using such a distribution. The distribution is designed to have median of zero regardless of the skew. To accomplish this, the distribution can have values that are less than zero. This is unlike the scales of interest in this paper, which have a fixed lower bound at zero. We believe that this should not have a large impact on analysis, the tests described in this paper should be robust under translation of scores under the scale.

Another limitation with the g-h distribution is that it is a continuous unbounded distribution. The distribution of the scales described in this manuscript are discrete and bounded, generally at zero with some upper bound. This may have some impact on the applicability of these simulations to these datasets.

The tests investigated were conducted using standard tests in the statistical program R (R Core Team, 2015). The HC4 regression was conducted using the function *olshc4* from the WRS package that accompanies Wilcox (2012). The ranked regression model was conducted using the function rfit (Kloke & McKean, 2012). The remaining functions (t.test for t tests, cor.test for correlations, wilcox.test for the Mann-Whitney U test) were all contained in the stats package released with R. The full code for these simulations can be found at https://github.com/lauken13/Community-Samples-and-Skewed-Data.

In this section details of the simulations presented in this manuscript are described. The distributions of the simulated data were created using a g-h distribution. There are some issues with this distribution, so the simulations should be repeated using random samples from real data.

3.12 Addendum

This paper was designed to demonstrate to researchers in very applied fields that there are substantial perils associated with the use of scales designed to make clinical distinctions with a non-clinical sample. It proposed very simple solutions like ranked methods. It is important to note that we never intended for these methods to be taken as the most appropriate solution for all researchers with this sort of data. Rather this manuscript was a suggestion that simple and relatively well-known alternatives would be more appropriate than the current prevailing practice as they do not involve violating core assumptions.

Realistically our methods should be considered only as an improvement to the current standard. Another possibility requires an elegant family of distributions recently proposed by Smithson and Shou (2017). I will illustrate it using the data from the community sample discussed earlier in the chapter (i.e., the distribution of scores on the DASS subscale for 451 adults)

The study from which this was taken included measures on the other two subscales of the DASS (anxiety and stress) and a number of demographic and health related subscales. In this section we will evaluate current practice (e.g., Pearson correlation), the ranked method proposed in the chapter (e.g., Spearman correlation) and the method proposed by Smithson and Shou (2017) as implemented through the accompanying package cdfquantreg (Shou & Smithson, 2016). Our question of interest is the relationship between depression and stress. This analysis has similar features to a vignette accompanying the cdfquantreg package that considers the relationship between stress and anxiety.

The most common analysis for estimating the relationship between depression and stress, the Pearson correlation, suggests that the relationship between depression and stress is significant (r=.611, p < .001). The Spearman correlation is slightly smaller (r=.607, p < .001), although this method is ill-suited to analyses where there are multiple ties. There are many in this dataset because many individuals score in the very low ranges of the scale.

The alternative proposed by Smithson and Shou (2017) is to use the package cdfquantreg to investigate the relationship between depression and stress in this dataset using a logit-logistic base distribution. This allows us to model the location and dispersion separately. In this case we see that depression grows more severe with higher levels of stress ($\mu = 6.71$, $p < .001$), but also that the relationship has less dispersion as levels of stress grow ($\sigma = -0.669$, $p < .001$).

This more complex model demonstrates that those who score highly on the depression subscale tend to be stressed, whilst those who score lowly on the depression subscale may or may not be stressed. In the next chapter, we see again that increasing complexity can change the richness and informativeness of the conclusions that can be made.

Chapter 4

When the sample is contaminated

> I do not refuse my dinner simply because I do not understand the process of digestion.
>
> Oliver Heaviside

While the previous chapter considered the violation of distributional assumptions, this chapter investigates the impact of impurities, conflicts, or contaminants in the sample. Contamination indicates that a certain proportion of a sample is *not* representative of the population or process of interest and instead represents some other population or process. Most researchers would agree that this is an inevitable event in psychological research. In this chapter I consider potential ways of inferring the population mean given some underlying assumption that there exists some proportion of the data that are contaminants.

As in chapter three the work here compares a number of models that differ in their degree of complexity. The main difference between the complex models in this chapter and those in chapter three is that these models were created to account for an underlying belief of how the the data were generated. This demonstrates again that complex models are potentially beneficial, but only when when they address the researcher's beliefs and aims. This work shows that the most complex models (the contaminated normal model) make the strongest claims about the process in which the data were generated, and the least complex models (Bayesian bootstrap in terms of free parameters, normal model in terms of freedom to account for different shapes) make the smallest claims about the process of data generation.

Our results indicate that when contamination is expected, the most complex model is the most appropriate. This is contrary to chapter three, where the most appropriate model was among the most simplest. Why is there this difference? I argue that when the additional complexity stems from an underlying notion of how the data were generated, it allows for richer and more reliable data analysis.

Like Oliver Heaviside in the quote that begins this chapter, I agree that refusing to model and analyse data because I do not (and will never) fully understand the process in which it was created would be futile. However, Heaviside sells himself short. He understands a great deal about the process of digestion (it is an organic process and starts at the mouth) and that understanding heavily governs the things he is willing to endorse as food and the process by which he feeds himself (namely through his mouth). Knowledge about the process has informed his model of digestion and aids him in making rich claims about the process of dinner. This chapter argues our knowledge about the process in which data are produced *should* be incorporated into the models we use. This allows the researcher to draw more precise, more accurate, and altogether richer conclusions about their data.

Statement of Authorship

Title of Paper	Not every interval is credible: Evaluating robustness in the presence of contamination in Bayesian data analysis
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Kennedy, L. A., Navarro, D. J., Perfors, A., & Briggs, N. (2017). Not every credible interval is credible: Evaluating robustness in the presence of contamination in Bayesian data analysis. Behavior Research Methods, 1-16.

Principal Author

<table>
<tr><td>Name of Principal Author (Candidate)</td><td colspan="3">Lauren Ashlee Kennedy</td></tr>
<tr><td>Contribution to the Paper</td><td colspan="3">Contributed to the original core concept of this experiment (with DN), coded some of the models. Planned (in conjunction with DN and AP) the simulation procedure, conducted and analysed the results. Wrote first draft of the manuscript and contributed to editing process.</td></tr>
<tr><td>Overall percentage (%)</td><td colspan="3">75%</td></tr>
<tr><td>Certification:</td><td colspan="3">This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.</td></tr>
<tr><td>Signature</td><td></td><td>Date</td><td></td></tr>
</table>

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

<table>
<tr><td>Name of Co-Author</td><td colspan="3">Daniel J. Navarro</td></tr>
<tr><td>Contribution to the Paper</td><td colspan="3">I contributed to the main idea of the paper, helped to design the simulations, and helped to frame and edit the paper.</td></tr>
<tr><td>Signature</td><td></td><td>Date</td><td>31/08/2017</td></tr>
</table>

Name of Co-Author	Amy Perfors
Contribution to the Paper	I helped figure out which simulations to run and what analyses were necessary. I also contributed to the editing of the paper.

Signature		Date	31/08/2017

Name of Co-Author	Nancy Briggs		
Contribution to the Paper	I contributed to the planning and framing of the paper		
Signature		Date	31/08/2017

Please cut and paste additional co-author panels here as required.

Not every credible interval is credible: Evaluating robustness in the presence of contamination in Bayesian data analysis

Lauren A. Kennedy

School of Psychology

University of Adelaide

Daniel J. Navarro

School of Psychology

University of New South Wales

Amy Perfors

School of Psychology

University of Adelaide

Nancy Briggs

Mark Wainwright Analytical Centre

University of New South Wales

4.3 Abstract

As Bayesian methods become more popular among behavioral scientists, they will inevitably be applied in situations that violate the assumptions underpinning typical models used to guide statistical inference. With this in mind, it is important to know something about how robust Bayesian methods are to the violation of those assumptions. In this paper we focus on the problem of *contaminated* data (such as data with outliers or conflicts present), with specific application to the problem of estimating a credible interval for the population mean. We evaluate five Bayesian methods for constructing a credible interval, using toy examples to illustrate the qualitative behaviour of different approaches in the presence of contaminants, and an extensive simulation study to quantify the robustness of each method. We find that the "default" normal model used in most Bayesian data analyses is not robust, and that approaches based on the Bayesian bootstrap are only robust in limited circumstances. A simple parametric model based on Tukey's "contaminated normal model" and a model based on the t-distribution were markedly more robust. However, the contaminated normal model had the added benefit of estimating which data points were discounted as outliers and which were not.

Keywords: Bayesian data analysis, robust methods, contaminated data

Not every credible interval is credible: Evaluating robustness in the presence of contamination in Bayesian data analysis

Data analysis plays a central role in scientific discovery, helping researchers organize and interpret the results of experimental and observational studies. In the majority of cases, researchers conduct data analyses with the help of "default" models that are broadly applicable to a wide range of scientific problems, rather than attempt the more arduous (and often unrealistic) task of developing a task specific model for each problem. As a consequence of this pragmatic behavior by scientists, it is important to verify whether these default models behave in a sensible way when applied to real data that may not satisfy the assumptions built into the model. Much of the applied statistics literature has this character (Bamnett & Lewis, 1994; Hoaglin et al., 1985; Wilcox, 2012).

Because the world is complex and data are limited, it is generally recognized that it is never possible to build a model that is detailed enough to capture everything of interest in a particular data analysis problem: it is for this reason that statisticians often remark that "all models are wrong" (e.g., Box, 1976). This is also why examining the performance of models when their assumptions are violated (i.e., when the model is *misspecified*) is critically important for ensuring sound scientific practice. Following the advice of Box (1976), our goal here is to determine which models are most useful when their assumptions are violated.

With this in mind, it is useful to consider the rise of Bayesian data analysis in behavioral statistics. Orthodox statistical inference begins from the frequentist premise, which assumes that the word "probability" describes the results of an *aleatory* process: when a repeatable event such as the flip of a coin is repeated sufficiently many times, in the long run the proportions of cases converge to a specific number (e.g., 50% of coin flips come up heads). The vast majority of analyses in the psychological literature rely on orthodox statistical theory: when reporting confidence intervals and p-values, for instance, one is relying on frequentist methods.

Unlike the orthodox approach, the Bayesian view does not interpret probability in aleatory terms. Instead, the Bayesian approach begins from an *epistemic* assumption, and interprets the word "probability" as the degree of belief that a reasoner should place in an particular proposition. Over the last few decades, the Bayesian approach has become increasingly popular, with many researchers choosing to report Bayes factors rather than p-values and credible intervals instead of confidence intervals. In part this shift has been driven by the well-documented failures of orthodox methods (Wagenmakers, 2007), and in part by the fact that there are now several easily accessible tools that provide default Bayesian solutions to common data analysis problems (for a default t-distribution model see Kruschke & Meredith, 2015, online version; http://www.sumsar.net/best_online/). For instance, the development of the BayesFactor package in R (Morey & Rouder, 2015) and its easy accessibility through user friendly software such as JASP (JASP Team, 2016) has allowed widespread adoption of Bayesian methods that were once available only to specialists. In our view this mainstreaming of Bayesian inference is greatly to be desired, but it carries new problems for Bayesian statisticians to consider. In particular, when Bayesian analyses were more inaccessible, that meant that the only people who implemented Bayesian statistics were technically sophisticated enough not to misuse them; as they become more mainstream – which is a good thing! – it does mean that they are also more likely to be applied inappropriately or misused.

What challenges are these? In an ideal world, the answer would be “none at all”: every researcher would consider the particular constraints imposed by their research problem, work out what those constraints imply about the statistical model to be used, construct the priors that best express the knowledge that the researcher possesses, implement that model, and then report the posterior distributions that result once the data are observed. In the real world, of course, this is not so much a polite fiction as it is a statistician’s fantasy. In our experience researchers (sometimes) do not have rich enough knowledge of the problem or (very often) lack the many years of statistical training required do anything other than rely on default methods.

As default Bayesian tools become mainstream, they will increasingly be used in situations that violate their assumptions. As such, it is critical to consider how well default Bayesian methods perform when they are misspecified, and to consider whether different defaults are required if performance is especially poor with respect to common problems. There are a many different situations in which model misspecification becomes a concern: in this paper we focus on the problem of *contaminated data*. Contaminated data problems arise when a researcher’s observations are composed not just of data that reflects the question of interest, but also data reflecting some other generating process.

4.3.1 The problem of contaminated data

We frame the problem of contaminated data in terms of one of the simplest estimation problems in statistics: using observed data $X = (x_1, \ldots, x_n)$ to construct an interval estimate for a population mean. Very typically we might assume that the observed data are drawn from a normal distribution with unknown mean μ and standard deviation σ. In most cases the quantity of interest is the unknown mean μ. With the help of this model, an orthodox statistician might construct a confidence interval that is designed to ensure that μ falls within the interval for 95% of cases, or a Bayesian statistician might construct a credible interval and claim that there is a 95% probability that the true mean falls within the range. Regardless of whether one is Bayesian or frequentist, the analyses that empirical scientists usually rely upon assume that every observation X has been sampled from the relevant population distribution.

In many – perhaps most – real world inference problems facing psychological researchers, this assumption is untenable. Sometimes participants press buttons randomly in the hope of leaving the study early, sometimes people answer the wrong question, sometimes a legitimate response is miscoded in the data. As a consequence, most datasets end up with some unknown proportion of *contaminants*; observations that arise from some other unknown process and are not informative about the population parameters of interest to researchers. To the extent that one's statistical inferences are driven by contaminants, researchers run the risk of drawing the wrong conclusions from experimental data.

The difficulties posed by contaminants are well known both to empirical scientists and to statisticians, and the issue is very frequently covered in introductory research methods classes. Researchers typically use exclusion criteria to try to identify contaminants and remove them from datasets. Similarly, statisticians have developed a variety of *robust* methods that are designed to produce sensible inferences even when contaminants are present, from both a Bayesian (Berger et al., 1994; Wang & Blei, 2015) and frequentist perspective (Huber, 1964; Huber & Ronchetti, 2009). In this paper we approach the problem from a Bayesian perspective.

To what extent does contamination present a problem for Bayesian data analysts? One possibility is that all Bayesian methods are robust to contamination simply by virtue of being Bayesian. This may seem somewhat implausible to an expert reader – and we imagine that statisticians would recognize this as the strawman hypothesis that it is – but Bayesian approaches often seem so desirable that experimental researchers could certainly be forgiven for thinking that this is a serious possibility. With this in mind, in the next section we will focus on relatively simple Bayesian models that can easily be applied to typical experimental data in the behavioral sciences, and investigate the differences between them when contaminants are present.

4.4 Four ways to get five posteriors

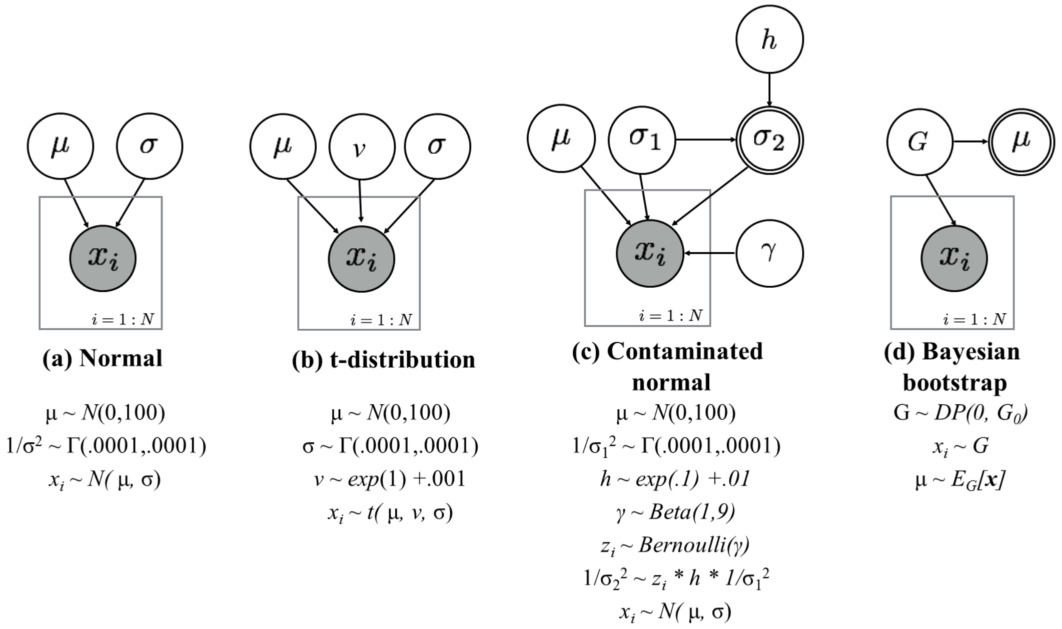
We focus on the relatively simple situation in which a researcher is presented with a sample of n continuous-valued univariate observations $X = (x_1, \ldots, x_n)$, and the goal is to construct an interval that captures the researcher's uncertainty about the population mean. Although this situation is somewhat limited in scope, it corresponds to a very common problem in psychological research, since it is quite typical for researchers to examine and report means and confidence intervals (or the Bayesian equivalent). In this section we describe five different methods[1] that a Bayesian researcher might use to construct a highest density credible interval when presented with this inference problem: the *normal* model, the *contaminated normal* model, the *t distribution model*, and two versions of the *Bayesian bootstrap* method.[2]

4.4.1 The normal model

The first model we consider is a "default" normal model that makes no attempt to address the problem of contaminants. As shown in Figure 4.1a this model assumes that all observations are sampled independently from a normal distribution with unknown mean μ and unknown standard deviation σ. Because these parameters are unknown, the researcher specifies priors that describe their pre-existing state of knowledge. We restrict ourselves to the grossly typical situation in which the researcher has very little prior knowledge about the parameters, and specifies diffuse priors over μ and σ. To formalize this, we adopt a prior in which μ is drawn from a Normal(0,100) distribution and the precision of the data (i.e., $1/\sigma^2$) is drawn from

[1] Code implementing the models can be found online at https://github.com/lauken13/Robust-Intervals

[2] In our initial investigations we also considered two frequentist methods for estimating confidence intervals, one of which was the Student's t interval and the other of which was the 20 % percentile-t trimmed bootstrap. Not surprisingly the overall behavior of the Student's t interval was roughly similar to that of the Bayesian normal model, and the frequentist trimmed bootstrap was somewhat similar to the Bayesian trimmed mean bootstrap. However, given that the behavior of the frequentist methods is well-documented elsewhere in the literature and that our goal in this paper is not to revisit the "Bayesians versus frequentist" debate yet again, we decided to omit the frequentist methods.



(a) Normal

$\mu \sim N(0,100)$
$1/\sigma^2 \sim \Gamma(.0001,.0001)$
$x_i \sim N(\mu, \sigma)$

(b) t-distribution

$\mu \sim N(0,100)$
$\sigma \sim \Gamma(.0001,.0001)$
$v \sim exp(1) +.001$
$x_i \sim t(\mu, v, \sigma)$

(c) Contaminated normal

$\mu \sim N(0,100)$
$1/\sigma_1^2 \sim \Gamma(.0001,.0001)$
$h \sim exp(.1) +.01$
$\gamma \sim Beta(1,9)$
$z_i \sim Bernoulli(\gamma)$
$1/\sigma_2^2 \sim z_i * h * 1/\sigma_1^2$
$x_i \sim N(\mu, \sigma)$

(d) Bayesian bootstrap

$G \sim DP(0, G_0)$
$x_i \sim G$
$\mu \sim E_G[\boldsymbol{x}]$

Figure 4.1: Graphical models for the usual normal model (panel a), the t-distribution model (panel b), the contaminated normal model (panel c), and the statistical model underpinning the Bayesian bootstrap procedure (panel d). Beneath the graphical models we have included prior and relational information about each node. Wherever sensible we have endeavoured to keep the priors for similar parameters similar across the models.

a Gamma(.0001, .0001) distribution, but other variations are possible. Following typical practice in the Bayesian data analysis literature we implemented this model using JAGS, though given the simplicity of the model almost any software package could be used. The normal model appears in various guises in the Bayesian data analysis literature, for example the basic Bayesian t-test (Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009), Bayesian ANOVA (Rouder & Morey, 2012), and many other custom models (Lee & Wagenmakers, 2014).

4.4.2 The t-distribution model

Andrade and O'Hagan (2011) and Kruschke (2013) suggest a convenient alternative to the normal model. These authors take the view that the normal distribution is very sensitive to outliers, and note that this sensitivity can be reduced if the data are assumed to arise from a distribution with heavier tails. With that in mind, they suggest that the t-distribution is a convenient heavy tailed distribution that only adds one additional free parameter (degrees of freedom) when compared to the normal model. Theoretically this distribution is created from a mixture of normals with expected value μ and a precision drawn from $\gamma(df/2, scale^2 * df/2)$. The full model is depicted graphically in Figure 4.1b.

In order to allow for direct comparison across models we chose to use the same prior for μ and precision parameters as the normal model. For the degrees of freedom we use an exponential prior as recommended by Kruschke (2013). In order to avoid increased flexibility due to hyper-parameters, we use exp(1) as the prior, which is relatively conservative in the amount of variation in the standard deviations of the normals in the mixture.

This model construction promises to be robust when faced with outliers. Each datum observed can be thought of as a draw from a normal distribution with the same mean as other data in the sample, but with a precision that is wider or narrower as dictated by its relationship to the rest of the data (i.e. a more extreme point would be drawn from a normal with lower precision and hence higher variance). Whilst theoretically this is true, the data as a whole is fit to a t-distribution with parameters μ, σ and ν (degrees of freedom). This final parameter, ν, provides the only indication to the amount of data that is discounted as outliers or contaminants. The exact relationship, however, is not clear. Certainly it does not give an estimate of the proportion of data that are discounted, nor the degree to which they are discounted. With this in mind, we propose an alternative model that does allow the user some notion of the proportion and extremity of the contamination.

4.4.3 The contaminated normal model

As an alternative to the normal and t-distribution models, we propose the *contaminated normal* model (e.g., Tukey, 1960). Although this model requires an additional number of free parameters, we argue that it is easier overall to interpret. The contaminated normal is a very simple mixture model in which the target distribution is again a normal distribution with unknown mean μ and standard deviation σ_1, as per the usual normal model. However, some unknown proportion of the observations γ are drawn from a contaminant distribution that again has mean μ but has a different (generally larger) standard deviation σ_2. The standard deviation of the contaminants is governed by the scale parameter h, such that $\sigma_2 = h \times \sigma_1$. The contaminated normal model is depicted graphically in Figure 4.1c.

In our applications we assume that the contaminant proportion γ is drawn from a Beta(1,9) distribution, reflecting a somewhat conservative expectation that 10% of the data are contaminants. Similarly, we assume that h is drawn from an Exponential(.1) distribution, implying that on average the contaminant distribution will be 10 times as wide as the target distribution. The model enumerates the probability each datum is a contaminant and the overall proportion of contamination.

Clearly, the contaminated normal model is not intended to serve as a literal model for how contamination processes play out in psychological data. In real life there are very few phenomena of interest to behavioral researchers that are precisely normally distributed, and there is no reason to think that contaminant distributions are any more likely to be normal either. However, in much the same way that researchers treat normality as a sensible first approximation when constructing models for a phenomenon of interest, we propose that normality makes sense as a first approximation to contaminant processes too. Our goal is not to have a realistic generative model for every psychological experiment (which seems impossible) but rather to construct a minimal, sensible account that has the potential to detect the most misleading contaminants and prevent them from having a disproportionately large effect on the inferences about the target.

To our knowledge the contaminated normal model has not previously been considered as a general purpose tool in psychological data analysis, but the principles upon which it is based are fairly standard. For example, the idea of explicitly building a high-variance "contaminant distribution" into the model is a relatively standard approach in cognitive modelling (e.g., Ratcliff & Tuerlinckx, 2002; Zeigenfuse & Lee, 2010). Similarly, the overall effect of adding the contaminant distribution is to create a "heavy tailed" model for the observed data, consistent with standard advice in the Bayesian robust statistics literature (e.g., Berger et al., 1994). Given the simplicity of the idea and its alignment with existing approaches in the literature, the contaminated normal model seems a sensible way to construct a default model for contaminated data.

4.4.4 The Bayesian bootstrap

In the contaminated normal model, the central idea is to model contaminants with the help of an explicit, parameterized contaminant distribution. The model is very much a parametric model, making strong assumptions about the shape of the target distribution and the shape of the contaminant distribution. An alternative approach is to consider nonparametric methods in which the researcher avoids relying on any strong assumption about distributional shape. This approach is very common in frequentist analyses: when presented with unusual data, a typical orthodox solution is to use bootstrapping (Efron & Tibshirani, 1994) to construct confidence intervals (Efron, 1987; Hall, 1988).

Although less commonly used in the psychological literature (e.g., Navarro, Newell, & Schulze, 2016) a Bayesian version of bootstrapping exists and is known as the *Bayesian bootstrap* (Rubin et al., 1981). In much the same way that bootstrapping plays an important role in frequentist robust statistics (Wilcox, 2012), one might imagine that the Bayesian bootstrap approach could serve a similar role for Bayesian data analysts in psychology. The Bayesian bootstrap (henceforth BB) algorithm is described in Figure 4.2, but to understand the logic of the BB procedure it is useful to first discuss Bayesian nonparametric inference in a relatively non-technical way.

From the Bayesian perspective, nonparametric inference proceeds as follows. The observed data are assumed to be generated by some unknown probability distribution G whose properties are largely unknown. In parametric Bayesian inference the data analyst typically places a strong constraint on G, perhaps by assuming that G is a normal distribution or a member of another family of distributions that can be described by a small number of parameters. In nonparametric Bayesian inference no such constraint is imposed: instead, the data analyst specifies a prior $P(G)$ that has broad support across the space of possible probability distributions. After observing the data X the analyst updates their beliefs about the unknown distribution. This yields the posterior $P(G|X)$ that assigns a degree of plausibility to every possible distribution G that could have generated the observed data.

Given these beliefs about the population distribution G it is conceptually straightforward to construct the posterior distribution for *any* characteristic of that distribution. For instance, if we let $E[G]$ denote the expected value (mean) of the distribution G then the posterior density for the mean is given by

$$P(\mu) = \int_{G|E[G]=\mu} dP(G)$$

and a credible interval can be constructed accordingly. The graphical model for this inference problem is illustrated in Figure 4.1d, and highlights the fact that the data analyst "directly" specifies the prior for the distribution G. The population mean μ is simply one of many properties that can be calculated from the distribution and plays no special role in the generative model. As such, the same underlying model can be adapted to compute a credible interval for the population median, trimmed mean, standard deviation, skewness, or any other property that can be computed from a generic probability distribution G. Given our interest in contaminated data, we consider two cases of particular interest: the mean (henceforth denoted BB-mean) and the 20% trimmed mean (denoted BB-trimmed), where the most extreme 20% were trimmed from the distribution before the mean was calculated.

The description above is a fairly general (and imprecise) description of nonparametric Bayesian methods, and like any statistical method the complexity lies in the details. There are many different methods for constructing nonparametric priors $P(G)$, with different strengths and weaknesses. Examples include Dirichlet processes (Ferguson, 1973), Pitman-Yor processes (Pitman & Yor, 1981), Pölya trees (Mauldin, Sudderth, & Williams, 1992), Dirichlet diffusion trees (Neal, 2000) and many others besides. A technical overview of nonparametric methods is given by Ghosh and Ramamoorthi (2003) but there are several tutorials and applications relevant to psychologists in the literature (Bååth, 2016; Gershman & Blei, 2012; Karabatsos, 2017; Navarro et al., 2006). For the purposes of the current paper it suffices to note that the statistical model underpinning the BB is a limiting case of the well-known Dirichlet process in which the statistician expresses the maximum degree of uncertainty about the unknown distribution G.[3]

The end result of this is that the Bayesian Bootstrap in its most common form (Rubin, 1981) considers the observed values to be the only possible values in the population. These values are given weight proportional to that they were observed in the sample. In this way this method is very similar to a traditional frequentest bootstrap.

[3] Formally, the prior in Rubin et al. (1981) is equivalent to the special case of a Dirichlet process that specifies a base measure αG_0 where the base distribution G_0 has broad support and we apply the limiting version of the prior as the concentration parameter $\alpha \to 0$. As has been noted in the statistics literature (Diaconis & Freedman, 1986, 1) the Dirichlet process prior has some limitations when used as a prior over continuous distributions: in the limit the posterior concentrates on discrete distributions with probability 1, and leads to posterior inconsistency when the true distribution is continuous. In such cases the only possible values in G are those that were observed in the sample, leading to a "comb"-like appearance. In addition to the problem noted in the main text the BB inherits the other concerns associated with Dirichlet process models, and as such some caution is warranted.

The Bayesian bootstrap (BB) is based on the nonparametric Bayesian model described in the main text, and describes a method for sampling probability distributions G (referred to as BB replicates) from the posterior belief distribution. Specifically, given data $X = (x_1, \ldots, x_n)$ we can sample from the posterior $P(G|X)$ by the following procedure:

1. Generate $n - 1$ uniform random numbers from the interval $[0, 1]$.
2. Order the numbers such that $u_1 < u_2 < \ldots < u_{n-1}$, and set $u_0 = 0$ and $u_n = 1$.
3. The weight w_i assigned to x_i is the difference $u_i - u_{i-1}$.
4. The sampled distribution G is constructed by setting $P(x = x_i) = w_i$. Values not represented in the original dataset are assigned probability zero.

Given a set of BB replicates generated in this fashion, and a population parameter (e.g., the mean) defined in terms of some function $f(\cdot)$ of the population distribution G, we can construct a simple Monte Carlo approximation to the posterior distribution of the parameter by applying the function to each BB replicate.

Figure 4.2: The Bayesian bootstrap procedure. The distribution G is a weighted set of numbers that were observed in the dataset. Once G has been formed, the population parameter of interest (in this case either the mean, or the mean of a distribution with 20% of the most extreme points removed) is calculated.

From the perspective of dealing with contaminants, the nonparametric Bayesian approach has a good deal to recommend it. First, nonparametric methods do not make strong assumptions about the shape of the population distribution, and as such there is nothing preventing a nonparametric Bayes model from inferring that the data are generated from a complex distribution that includes a long tail of outliers. Second, because the method allows the researcher to construct a posterior distribution for any statistic of interest, it is no more difficult to construct a credible interval for the 20% trimmed mean than it is to do so for the untrimmed mean. This provides a natural Bayesian equivalent of robust frequentist methods that work in a similar fashion by bootstrapping a confidence interval for the trimmed mean (Efron & Tibshirani, 1994). Third, for the BB approach specifically, the computations are straightforward to implement: As Figure 4.2 makes clear, Bayesian bootstrapping is no more complex than its frequentist counterpart.

We considered two Bayesian Bootstrap models. The first focused on estimating the mean as the statistic of interest. The second estimated the 20% trimmed mean. If the Bayesian Bootstrap performs in any way similarly to the frequentest bootstrapping, then the 20% trimmed mean should be robust insofar as the frequentest 20% trimmed mean bootstrap is (Wilcox, 2012).

Set against these advantages, some shortcomings should be noted. In the original paper Rubin et al. (1981) points out that the BB is implicitly reliant on a statistical model that is a somewhat implausible in most practical settings. Most notably, because of the way that the prior $P(G)$ is spread so thinly across the space of possible distributions G, any empirical dataset (no matter how small the sample size) turns out to be large enough to "swamp the prior". The effect is so large that the posterior distribution $P(G|X)$ is entirely restricted to those distributions G that are defined only across values that appeared once in the original sample. Much like the frequentist bootstrap, the BB can only re-use the existing data: it does not consider unobserved values. Nevertheless, in this paper we take a pragmatic perspective. We do not believe that any general purpose statistical models actually represent plausible theories of the data, and in our view the obvious failures of the BB model make it no less plausible than (say) linear regression. The question at hand is whether the BB provides a useful "generic" tool for researchers to express their beliefs about unknown quantities. As Box (1976, p. 792) famously remarked "since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned with mice when there are tigers abroad". With

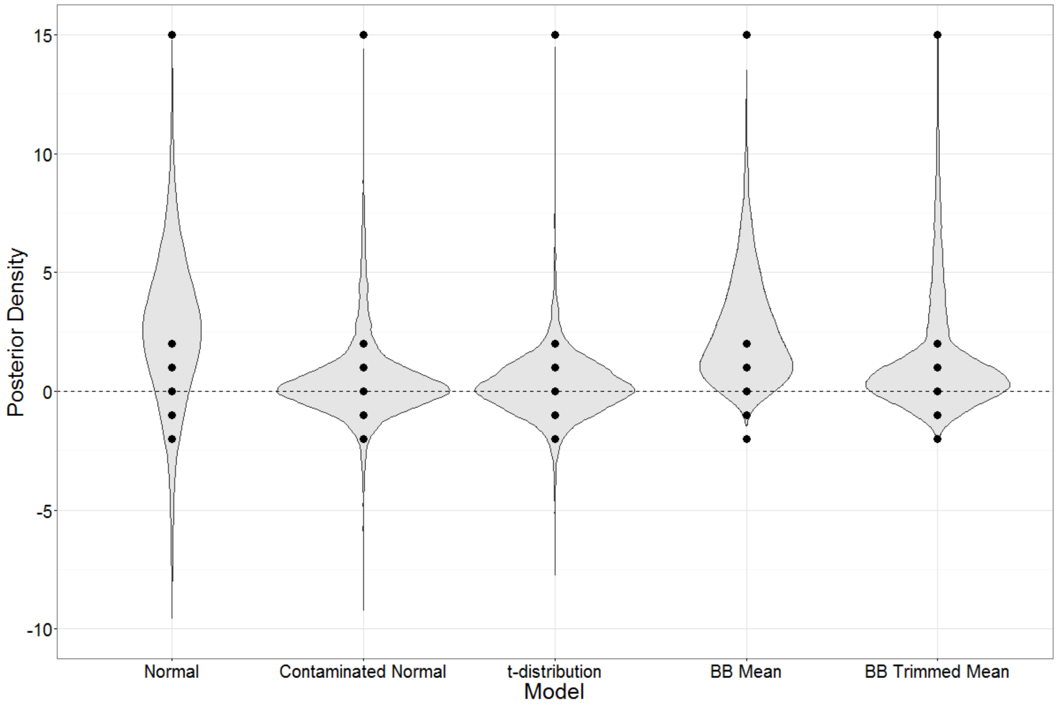


Figure 4.3: Posterior distributions for the population mean (or trimmed mean) when given a small dataset containing one outlier, according to the five different approaches. The dotted line represents the true population mean (without the contaminated point). If the model is estimating the mean well, we would hope the posterior to be peaked near or on the mean. The normal model (leftmost plot) is strongly influenced by the outlier. To a lesser extent, so is the Bayesian bootstrap interval for the mean (fourth plot from the left). The contaminated normal model (second from the left) , the t-distribution model (middle), and the Bayesian bootstrap interval for the trimmed mean (rightmost plot) are all less sensitive to the extreme data point.

that in mind, we turn our attention to the search for tigers.

4.5 When does contamination matter?

Armed with five Bayesian models, we first consider a problem familiar to anyone who has taken an undergraduate research methods class. Suppose a particular experimental condition yields the following six observations:[4]

$$-2, -1, 0, 1, 2, 15$$

and the researcher wishes to construct a 95% credible interval for the mean. It is immediately obvious that the sixth data point is an *outlier*, in the sense that it is quite distant from the other five. What is less than obvious is whether it also counts as a *contaminant*, in the sense of being an observation generated from a distribution entirely unrelated to the phenomenon of interest to the researcher. If it is indeed a contaminant it should have minimal influence on what the researcher believes about the phenomenon of interest. Noting this, it is instructive to consider what the five Bayesian approaches infer from these data.

The violin plots in Figure 4.3 show the relevant posterior distributions for all five models, and it is immediately clear from inspection that the choice of model matters. According to the normal model and the BB-mean approach, the best point estimate (calculated from the mean of the posterior) for the population mean μ is essentially identical to the sample mean of 2.49. The BB estimate of the population 20% trimmed mean discounts the outlier to some extent, producing an estimate of 1.38. The contaminated normal model and the t-distribution model produce the strongest outlier discounting and estimate a population mean of 0.47 and 0.22, respectively, only slightly larger than the mean of the first five observations. Other inferences from these models are equally discrepant.

[4] These numbers were arbitrary, but interestingly enough are also used in the toy example given in Kruschke (2014, Figure 16.5, p. 460).

We next consider the the highest density intervals for the mean.[5] According to the normal model there is a 95% probability that the population mean lies between $[-4.25, 9.08]$ and an 18% probability that the population mean is less than zero. The t-distribution model and contaminated normal model produce much narrower 95% credible intervals of $[-2.23, 2.27]$ and $[-2.49, 4.88]$ respectively, but a much larger probability that the population mean is negative (43% each). A Bayesian bootstrap of the mean gives a qualitatively different interval estimate of $[0.79, 7.03]$, with a lower bound noticeably closer to zero than either of the other two models. Not surprisingly, this model judges that there is only a 6% probability that the true mean is negative. Finally, using the BB to estimate the trimmed mean produces an interval of $[-1.60, 7.12]$ and a 27% probability of a negative mean.

Unique among the models, the contaminated normal model also provides an estimate of which items are contaminants: for instance, the data point 15 is given 85% probability, while the others are identified as contaminants with less than 7% probability. It also estimates the overall proportion of contaminants (γ in Figure 4.1) at 13.1%, and the overall width of the contaminate distribution (h in Figure 4.1) as 11 times the target distribution. As this simple example illustrates, there is a clear difference between the performance of the t-distribution and contaminated normal models on one hand, and the other models on the other. The contaminated normal model has the added benefit of allocating the probability that each datum is an outlier.

[5] Although all of the simulations provided in this manuscript use the highest density interval (HDI), we ran all simulations with equal-tailed intervals between the .025 and .975 quantiles. There was no qualitative difference in the results.

4.5.1 Systematic differences as a function of outlier extremity

Disagreement among models is not necessarily problematic, especially when those models are provided with very few data points. In this case, however, the differences are systematic and reflect a deep incompatibility among the models. To see this, suppose we were to smoothly vary the location of the sixth data point from $x = 0$ to $x = 25$, and observe how the credible intervals change for all five approaches. The results are shown in Figure 4.4, and reveal fundamental differences in how the models behave.

For the normal model and the BB-mean approach, the best estimate of the population mean tracks the sample mean: the more extreme the outlier becomes, the more extreme the estimated population mean comes. In contrast, the t-distribution model, contaminated normal and BB-trimmed methods all discount the outlier but do so in qualitatively different ways. The BB-trimmed method discounts the most extreme observations, but the extent of the discounting does not depend on the magnitude of the extremity: the estimated population mean rises more slowly than the sample mean, but it rises monotonically. Making the outlier more extreme *always* increases the estimated population mean.

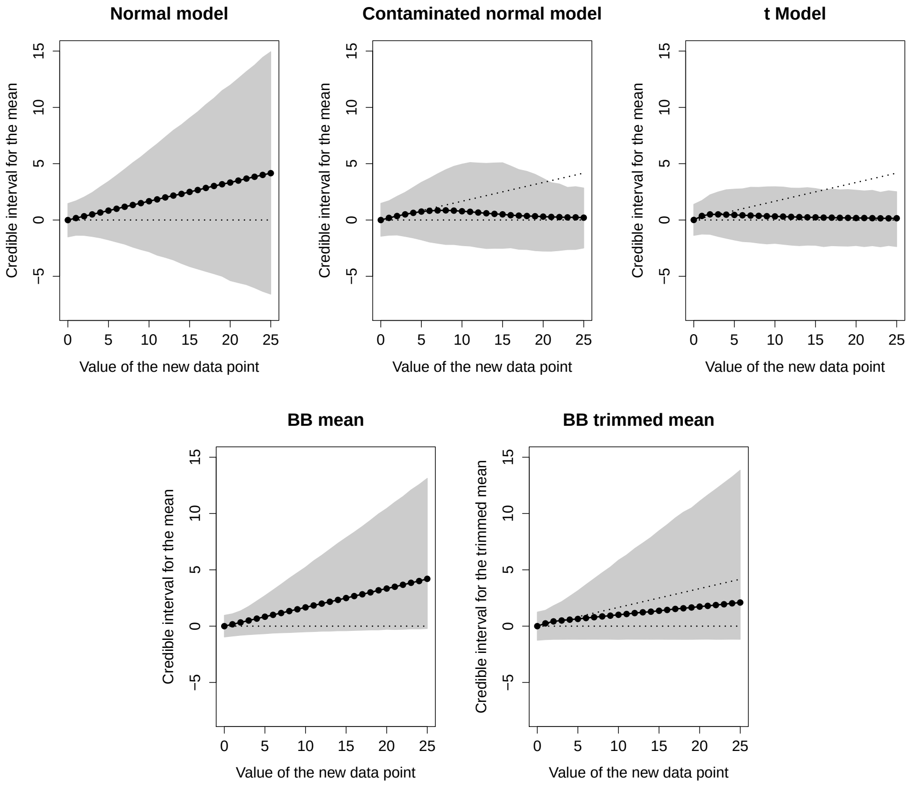
The contaminated normal model shows a fundamentally different pattern. For very modest outliers (up to about $x = 6$) the contaminated model behaves in a similar fashion to the normal model and the BB-mean approach, and the estimated population mean is almost identical to the sample mean. Above this point, the contaminant model begins to discount the outlier, reducing its influence on the es-

Table 4.1: Qualitative differences among the five different Bayesian approaches are summarized here. This table is a summary of the findings demonstrated in Figure 4.4, and summarises the theoretical expectations of how the models incorporate the outlier and thus how the beliefs about the mean vary across models. The third row relates to whether the posterior (and subsequent credible interval) is symmetric around the estimate for the mean. The last row addresses how the outliers are treated by the model.

	Normal	Contaminated	t-dist	BB-mean	BB-trimmed
Should outliers be discounted when estimating means?	No	Yes	Yes	No	Yes
Do beliefs about the mean reverse as the outlier becomes very extreme?	No	Yes	No	No	No
Should uncertainty at the upper end be mirrored in the lower end?	Yes	Maybe	Yes	No	No
Does the model allow explicit outlier labelling?	No	Yes	No	No	No

timated population mean. Eventually a point is reached that the model becomes so certain that the outlier is a true contaminant that any further increases to the outlier act to *reduce* the estimated population mean. By the time that the outlier reaches $x = 25$ the model has discounted it almost completely, and the estimated population mean falls back towards zero (the mean of the other five observations). This pattern is fundamentally at odds with the one produced by the BB-trimmed method.

The t-distribution model performs differently again. Much earlier than the contaminated normal model (i.e. before the outlier has reached a value of 5), it begins to discount this extreme point, producing a credible interval that is relatively consistent regardless of the value of the outlier.


Normal model

Contaminated normal model

t Model

BB mean

BB trimmed mean

Figure 4.4: An illustration of how the credible intervals change as a function of the extremity of an outlier. Each observed sample contains the values -2, -1, 0, 1 and 2 in equal proportions. The y-axis represents the width of the credible interval for these points plus a contaminant. The values on the x-axis represent the extremity of the contaminant point. As these values increase, the distance between this point and the rest of the sample grows. The dark points show the population mean estimated by the model for each value of contamination. The light grey points show the mean of the sample distribution both with and without the extreme point. The normal, BB mean and BB trimmed produce intervals that increase monotonically as the outlier grows more severe. Conversely, the contaminated normal model does not. Once an outlier is sufficiently extreme, the model discounts it and produces a credible interval on the remaining sample. The t-distribution model discounts the moderately small outliers to the same degree as the extremely large outliers.

A similarly dramatic disagreement appears when we consider how the credible intervals change. The normal distribution and t-distribution models are symmetric, and as a consequence any increase of uncertainty at the upper end of the interval must be mirrored by increasing the uncertainty at the lower end as well. The nonparametric model that underpins the BB imposes no such symmetry constraint, and so the upper and lower tails of the credible interval move relatively independently. For both BB models a shift in the outlier produces a change in the upper end of the credible interval that matches the change produced by the normal model, but the lower bound is largely unaffected.

The contaminated model produces a different pattern again. For relatively modest outliers (again, up to about $x = 6$) the upper and lower bounds of the interval look essentially identical to the normal model, and satisfy the same symmetry constraints. However, once the model begins to discount the outlier a different pattern emerges. The lower bound to the credible interval stays roughly constant, not too dissimilar to the nonparametric models. The upper bound, continues to rise for a while but eventually it too begins to fall back down as the model becomes increasingly certain that the outlier is a contaminant.

The various disagreements among models are summarized in Table 4.1, which characterises the differences in terms of three questions about how credible intervals should behave and how beliefs should change as a function of the extremity of an outlier. As the table highlights, the differences are not simple differences of opinion about priors (e.g., how big an effect size is expected); they are fundamental questions about the relationship between data and beliefs. In our view at least, the behavior of the contaminated normal model most closely matches what we think most (but not all) scientists would consider reasonable in most (but not all) psychological experiments. To the extent that this is true, a contaminated normal model makes considerably more sense as a default choice for the analysis of behavioral data.

4.5.2 Increasing the sample size does not erase the differences

A different perspective on the same issue is to consider how the models make different inferences as a function of sample size. In real data analysis problems, adding new data also changes the distribution of those observations. For the purposes of illustrating the "pure" effect of sample size, we create a sequence of datasets by adding multiple copies of the same six values $X = (-2, -1, 0, 1, 2, 15)$, and apply all five methods to the data. The results are shown in Figure 4.5.

Unsurprisingly, as the sample size increases, the normal model and the BB mean approach continue to produce estimates of the population mean that track the sample mean of 2.5, and the credible intervals tighten around that mean. Equally unsurprisingly, as the sample size increases the BB estimate of the 20% trimmed population mean converges to the 20% trimmed mean of the observed data, yielding a value of 0.5 in the large sample limit. The contaminated normal model and t-distribution model produce a different answer again: as the sample size increases the models become increasingly certain that the outliers (the copies of the $x = 15$ observations) are contaminants, and converge to the mean of the other observations, yielding an estimated population mean of 0. The contaminated normal model, whilst producing a wider credible interval for small n, produces a slightly tighter credible interval than the t-distribution model as n grows large.

To a statistician, none of these results might seem especially surprising: models that rely on different assumptions can produce quite different estimates of the same (or similar) quantity. To an empirical scientist, they might be somewhat problematic. The fact that the differences between models can be exaggerated by collecting additional data has non-trivial implications. It is not possible to avoid the problem of contaminants simply by being Bayesian, nor is it possible to sidestep the issue simply by collecting more data. To resolve the problems that these incompatibilities pose, substantive choices must be made by the data analyst.

4.5.3 How robust are the different methods?

The examples in the previous section highlight the importance of taking contamination seriously. In order to provide a better picture of how each of five methods performs, we present a simulation study that investigates the behavior of different Bayesian models when contaminants are introduced to the data. We consider two qualitatively different kinds of contamination, *biased contamination* and *unbiased contamination*, depicted in Figure 4.6. In both instances we assume that there exists a target population distribution (shaded plot) that represents the phenomenon of substantive interest to the researcher, but the dataset is corrupted: some proportion of the data arise from a contaminant distribution (grey plot) that is of little substantive interest.

The unbiased contamination scenario exactly matches the assumptions that under-

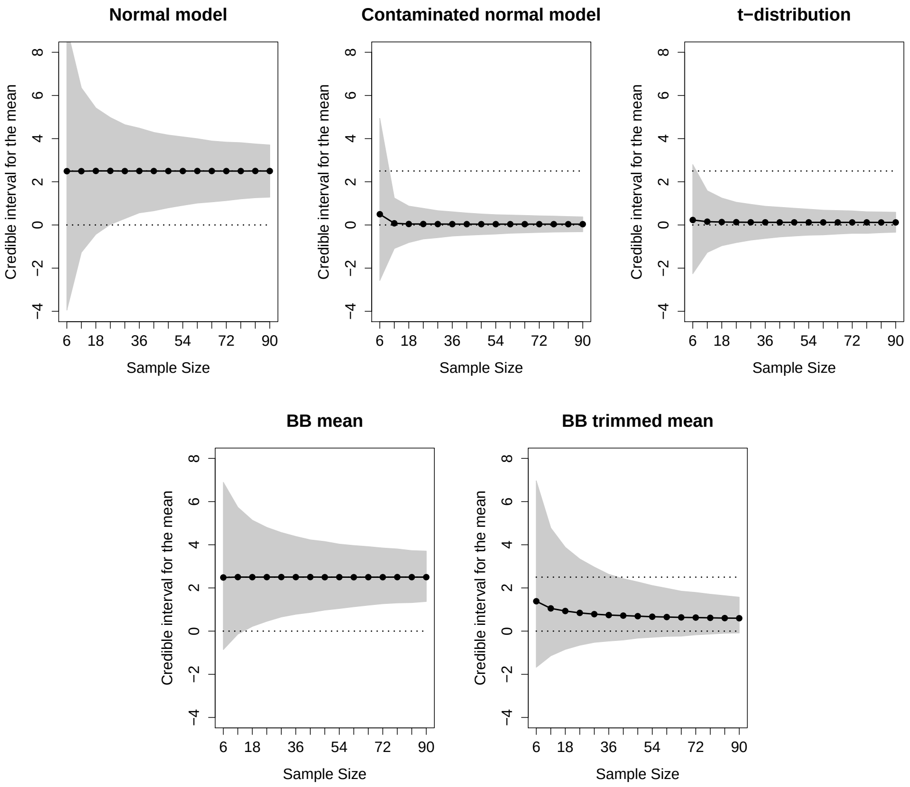


Figure 4.5: An illustration of how the credible intervals change as a function of sample size. To control for the effect of distribution, at every sample size n the observed sample contains the values -2, -1, 0, 1, 2 and 15 in equal proportions. The y-axis represents the width of the credible interval. The x-axis represents the sample size. As before in Figure 4.4, the dark points show the population mean estimated by the model for each value of contamination. The light grey points show the mean of the sample distribution both with and without the extreme point. The convergence rate of the normal model, BB mean and BB trimmed models are roughly similar, whilst the contaminated normal model converges much faster. The normal and BB mean models converge to approximately 2.5, the BB trimmed mean converges close to 0, and the contaminated normal model and t-distribution models converge to approximately 0.

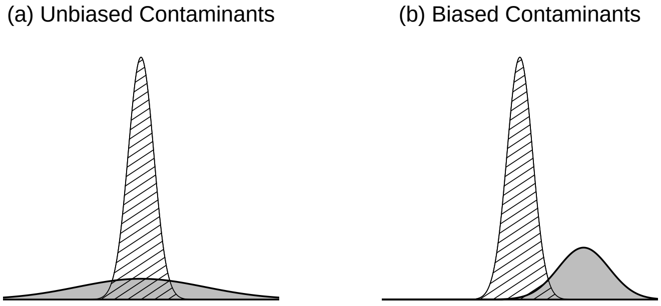


Figure 4.6: Two qualitatively different kinds of contaminant distribution. Each panel plots a target distribution (shaded) mixed together with a contaminant distribution (grey). In the *unbiased contamination* scenario (panel a) the contaminants are generated from a normal distribution that has high variance but is centred on the same location as the target distribution. In the *biased contamination* scenario (panel b) the contaminants have a different mean to the target distribution.

pin the contaminated normal model: the mean of the contaminant distribution is the same as the mean of the target distribution, but it has higher variance. In contrast, a biased contamination scenario is one in which the contaminants arise from a distribution that has a different mean from the target, and is potentially more worrisome.[6] Specifically, in our simulations the target distribution was always assumed to be a normal distribution with mean 0 and standard deviation 1. Unbiased contaminants were sampled from a normal distribution with mean 0 and standard deviation 10, whereas biased contaminants were sampled from a normal distribution with mean 10 and standard deviation 1.

[6] In these simulations we assume that the contaminants are normal. We also ran a version in which the contaminants were generated from a clumpy distribution, which we did by sampling contaminant distributions from a Dirichlet process. However, apart from making the results somewhat more variable everywhere, we did not find that clumpy contamination made much of a difference to our simulation results, and as such we have restricted our discussion to the case when contaminants are normally distributed.

In addition to varying the *kind* of contamination, we varied the *amount* of contamination from 0% to 50%. In practice we would expect that the proportion of contaminants should be well below 50% – indeed, if half of one's data are "contaminants" it does not seem to make sense to consider them contaminants any more. As mentioned earlier in the paper we expect that a value of 10% is more reasonable, but given that our focus is developing statistical methods that are robust even in somewhat extreme situations, it seems sensible to consider the full range of possibilities. Finally, we varied the sample size, considering values with $N = 20$, 50, 100 and 250 in order to cover the range of typical sample sizes in behavioral research. For each scenario we ran 1000 random simulations.

The results are plotted in Figures 4.7 and 4.8. In the top row of Figure 4.7 we see that when the contaminants are unbiased, all four Bayesian methods have reasonable coverage – in the frequentist sense of trapping the true mean with the desired 95% probability[7] – for most choices of sample size and contamination level, suggesting that for the most part these methods are robust when contaminants are unbiased. The one exception to this is the BB mean approach (plotted with black triangles), which performs noticeably worse than the other three approaches when the sample size is small and the contamination level is low. The reason for this is that small

[7] We have on occasion encountered Bayesian data analysts who oppose any evaluation of the frequentist properties of Bayesian methods, arguing that Bayesian methods should be defined on their own terms with no reference to frequentist concepts such as consistency, efficiency, unbiasedness etc. We can understand this impulse, but we take a pragmatic view ourselves. To the extent that these simulations annoy a truly committed Bayesian, we propose that even a committed Bayesian is entitled to treat the entire frequentist apparatus as an elaborate elicitation study: if a model produces long run behavior that the data analyst does not find desirable, it must be the case that this is not a model that the analyst places any (subjective) belief in. In that sense a frequentist evaluation can be very useful to a Bayesian researcher, as a method for uncovering revealed preferences.

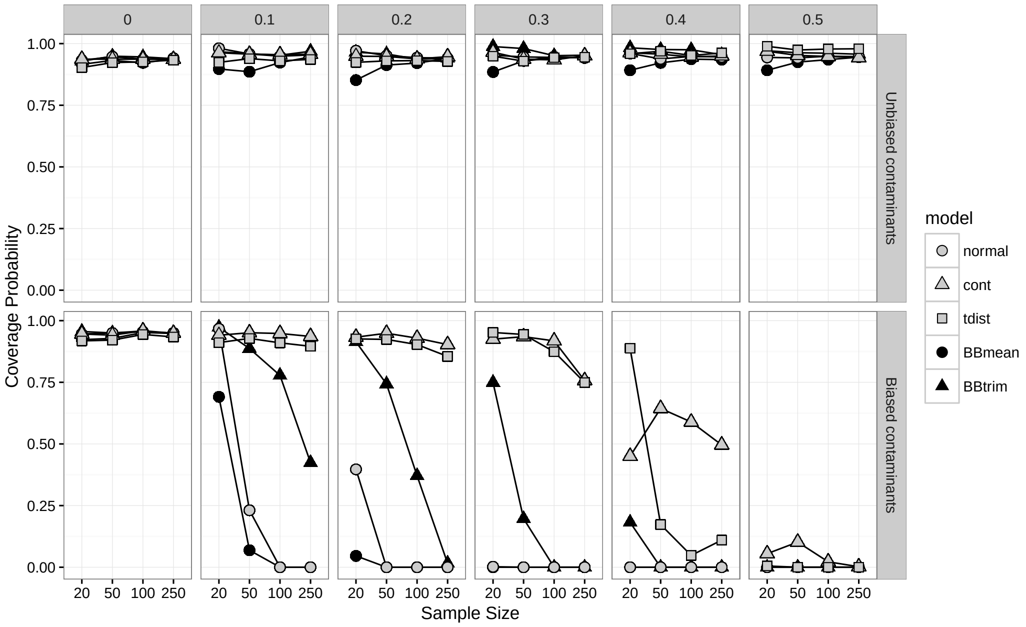


Figure 4.7: Coverage probability (proportion of cases in which the interval contains the true mean) for all five Bayesian methods as a function of sample size (x-axis), contaminant proportion (horizontal panels) and contaminant type (vertical panels). Insofar as each of these represents a 95% credible interval, ideally one hopes to see 95% coverage. Except for the BB mean, all of the models have good coverage probability when the contaminants are not biased to one side of the sample. As expected, when the contaminants are biased, all of the models eventually fail with 50% contamination, even with a very large sample. The contaminated normal model however, is robust at a much higher level of contamination (40%) than even the t-distribution.

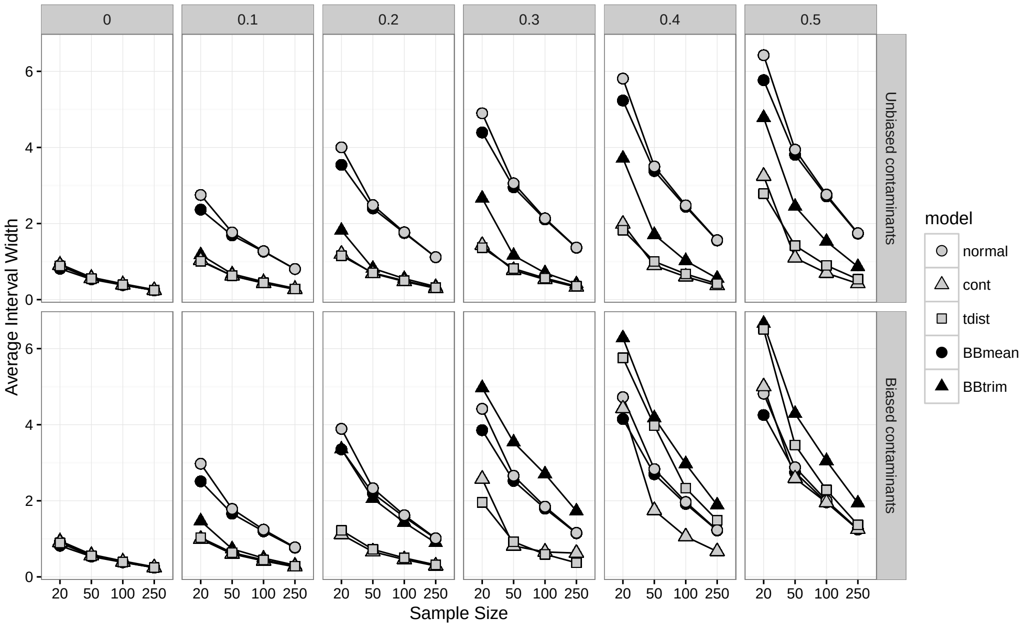


Figure 4.8: Average interval width (y-axis) for all five Bayesian methods as a function of sample size (x-axis), contaminant proportion (horizontal panels) and contaminant type (vertical panels). Although some contaminants produce wider intervals than others, they are quite similar in size. This suggests that the robustness in Figure 4.7 is not an artifact of more conservative (wider) credible intervals.

sample sizes and lower contamination rates is exactly the situation in which it is most likely for all the contaminants to lie on the same side, thereby producing data that is more representative of the biased contaminant situation, and (as we discuss below), the BB mean model does not fare well when faced with biased contamination.

In addition to measuring the coverage probability for the different estimators, we also examined the precision (i.e., average width) of the credible intervals produced be each of the four methods. The top row of Figure 4.8 suggests that the average width of the credible intervals does not differ greatly among all four methods. Taken together with the results in Figure 4.7 the simulations suggest that unbiased contaminants are comparatively harmless, having only modest influence on the accuracy or size of the credible intervals. To the extent that contaminants are unbiased, one might reasonably conclude that Bayesian inference using the default normal model is genuinely robust, and there would be little need to consider alternative approaches.

When we turn our attention to the case when the contaminants are genuinely biased, a different story emerges. As the bottom row of plots in Figure 4.7 the normal model and the BB estimate of the mean are both very sensitive to biased contaminants: even a modest level of contamination (e.g., 10 contaminants in a sample of 100 observations) is enough to make it virtually certain that 95% credible intervals constructed using these methods does not contain the mean of the target distribution. The failure of the BB mean approach is particularly noteworthy because it highlights that, like the frequentest bootstrap, nonparametric approaches are not automatically robust: using the Bayesian bootstrap to estimate the posterior distribution for the parameter mean frees the researcher from the assumption that the population is normal, but it does not provide any mechanism to identify or exclude contaminants. As such, when the contaminants are biased, so too is the the BB estimate of the population mean.

The other three methods perform rather better. The BB estimate of the population 20% trimmed mean is somewhat improved, but – somewhat unsurprisingly – it too becomes unreliable once the proportion of contaminants rises above the trimming level: at 30% contamination it is only marginally better than the other two methods. The contaminated normal model and t-distribution model, on the other hand, are both extremely robust, performing similarly to each other up to 30% contamination level. At a massive 40% contamination, the contaminated normal model produces a credible interval that contains the true mean in more than 50% of cases, beaten by the t-model only in the smallest sample size.

Again we note a slight inconsistency. With higher contamination levels (greater than 40%) the contaminated normal model produces a curious result. Unlike the other models (which start at their most accurate and decrease with increased n), the contaminated normal model is less likely to contain the true mean with 20 samples than 50 samples. This is because the contaminated normal model estimates the contaminant proportion using information from both the prior (which assigns a very small probability to such high levels of contamination) and the data observed. With very small sample sizes, the data provide very little evidence to suggest a high level of contamination. As such the model relies more heavily on the prior and incorrectly identifies contaminated points as members of the target distribution (e.g., only 29% of contaminants are correctly identified at a 40% contaminant level with $N = 20$). However, unlike the other models, increasing the sample size does lead to better estimation of the contaminant proportion (i.e. an increase to 70% correctly identified at 40% contaminant and $N = 50$). This increases the probability of creating a credible interval that contains the true mean.

Finally, we turn our attention away from the coverage probability and consider the precision (as measured by width) of the intervals in the biased contaminant scenario. Consistent with our earlier findings in the unbiased scenario, the robustness of the contaminated normal model does not seem to come at the expense of precision. As the bottom row of Figure 4.8 shows, the credible intervals produced by the contaminated normal model tend to be as narrow as those produced by the other approaches. It is noteworthy that the width of the intervals produced by contaminated normal model decreases with increased sample size in much the same manner as the other models. This suggests that the increased robustness of this model is not an artifact of the model producing more conservative (i.e. wider) intervals.

4.6 Detecting contaminants in the data

In this manuscript we considered five different methods of creating a credible interval for the mean. Overwhelmingly the contaminated normal model and t-distribution model perform best when contamination is present. However, they employ very different methods of accounting for this contamination. The t-distribution model accounts for outliers because its tails can be heavier than the tails of a normal distribution. The contaminated normal model, by contrast, is a mixture model of a target distribution and a (much wider) contaminant distribution. Because it is

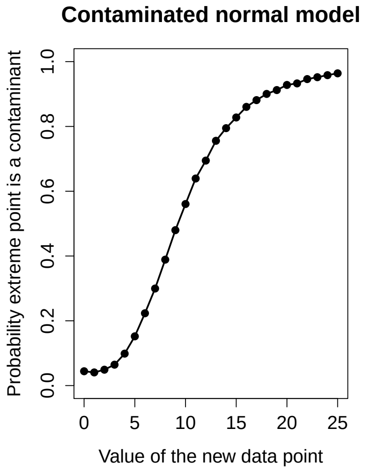


Figure 4.9: An illustration of the probability of the extreme point being labelled as an outlier by the contaminated normal model. The values on the x-axis represent the extremity of the contaminant point. The y-axis represents the probability of this point being classified as a contaminant relative to the other data. As the value grows more extreme, it becomes much more likely to be rejected as a contaminant.

a mixture model, every datum has some probability of being in the contaminant distribution. No other model that we have described in this paper has this feature. We have argued that this feature is rather desirable for a number of reasons. To more fully demonstrate the main one – that it identifies which specific data points are likely to be the outliers – we repeat the simulations from Figure 4.4, but now investigate the probability that the extreme point is identified as an outlier. As Figure 4.9 shows, the contaminated normal does well at outlier identification. Initially the extreme point is included in the target distribution, but as that point grows large relative to the rest of data it is increasingly categorized as part of the contaminant. This mirrors the shape of the credible interval in Figure 4.4.

4.7 Discussion

Statistical inference necessarily relies on simplifying assumptions: the causal mechanisms that produce empirical data are usually complex, and researchers typically do not understand the phenomenon under investigation well enough to properly specify what those mechanisms are (or else why would one be studying it?). The truism that "all models are wrong" is an expression of this necessity (Box, 1976). However, the fact that all models are wrong does not imply that all simplifying assumptions are equally sensible. Given the inherent difficulties associated with behavioral research, we argue that if a data analysis tool is intended to serve as a "default" method for behavioral researchers, it must be robust in the face of contaminated data. In practice is difficult to trust the inferences produced by a statistical model if that model is extremely sensitive to contaminants. This point is acknowleged within the orthodox literature on robust statistics (Wilcox, 2012), but has received comparatively little attention within Bayesian behavioral statistics.[8]

In this paper we restricted ourselves to a very simple statistical problem – the estimation of a credible interval for the population mean of a continuous univariate quantity – and to a set of five Bayesian approaches that are by no means exhaustive, but are arguably representative of the class of models that researchers might consider applying. In some respects our findings would not surprise any statistician: the default normal model is not robust to contamination, and a nonparametric approach is not necessarily better if applied in a naive way. The Bayesian bootstrap approach

[8] We do not suggest that Bayesians are unaware of this as an issue (see Wang and Blei (2015), for example), merely that research has tended to focus on different questions.

is somewhat robust if it is used to construct credible intervals for a quantity (population trimmed mean) that is not easily influenced by the tails of the distribution, but it is not at all robust when it is directly applied to estimate the population mean itself. This finding is mirrored in the parametric models as well: although we find that the default normal model is not robust, the t-distribution model and contaminated normal model are very robust, and appear to remain robust even when the contamination process violates the underlying assumptions of the model. At very high levels of contamination the t-distribution model is not quite as robust as the contaminated normal model and also provides no indication about which items are most likely to be outliers.

In view of the fact that Bayesian methods are not robust to contamination simply by virtue of being Bayesian, to the extent that contamination is commonplace in behavioral research, it seems sensible to prefer those models that are robust to contamination. We suggest that the contaminated normal model, the t-distribution, and the Bayesian bootstrap approach for inference about trimmed means all meet this criterion. In our simulation studies, the contaminated normal model seemed to perform best, but some caution is warranted since the relative robustness of these three methods almost certainly depends on the situation. When the population distribution is skewed, for instance, one might expect the contaminated normal model to perform poorly relative to the BB-trimmed method. More generally, when choosing among different approaches that have some claim to being labelled "robust", we suggest that the qualitative considerations listed in Table 4.1 provide a better guide to researchers than relying too heavily on any one simulation study. Our personal view is that in most situations the behavior of the contaminated model is the closest match to what a human reasoner would consider sensible, but ultimately the choice of model has to depend on the context.

That said, one advantage of Bayesian methods in general is that they make it easy to implement multiple ways to model data with outliers, some but not all of which we evaluated here. For instance, t-distributions can be straightforwardly implemented in JAGS and Stan, along with contaminant models for logistic regression and other hypothesis testing models. The textbook *Doing Bayesian Data Analysis* (Kruschke, 2014) is an accessible resource for many such approaches.

Finally, it is worth commenting on the generality of the problem under consideration. On the one hand, this paper does focus on a very simple problem. On the other hand, the problem is one faced by almost every paper reporting empirical data: in many areas of psychology and other behavioral sciences it is standard to report interval estimates for the population mean in various experimental conditions, and it is exactly this situation that we have considered. That being said, a natural extension of the work is to move beyond estimation problems to hypothesis testing scenarios. The contaminated normal model represents a sensible and simple tool for robust Bayesian estimation, and it is naturally extensible to construct robust Bayesian t-tests, ANOVAs and regression analyses. We are currently working on exactly these extensions.

In addition, although this paper is focused on data analysis, very similar problems arise for human learners in the real world. The data available to people is often contaminated with outliers of one sort or another, and a robust learner would need to deal with them just as statisticians do. The mixture model we employ here may be appropriate as a cognitive model as well, and pursuing this is an interesting alternate line of research. In the meantime, however, it seems clear – even on the basis of this initial investigation – that for a statistician the issue of robustness is just as critical for evaluating Bayesian models as it is for orthodox approaches, and should be given consideration when choosing the appropriate data analysis tool.

4.8 References

Andrade, J. A. A. & O'Hagan, A. (2011). Bayesian robustness modelling of location and scale parameters. *Scandinavian Journal of Statistics*, *38*(4), 691–711.

Bååth, R. (2016). *bayesboot: An Implementation of Rubin's (1981) Bayesian Bootstrap*. Retrieved from https://github.com/rasmusab/bayesboot

Bamnett, V. & Lewis, T. (1994). *Outliers in statistical data*. West Sussex, England: John Wiley & Sons.

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., ... Sivaganesan, S. (1994). An overview of robust Bayesian analysis. *Test*, *3*(1), 5–124.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Diaconis, P. & Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, *14*, 1–26.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185.

Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*(2), 209–230.

Gershman, S. J. & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12.

Ghosh, J. & Ramamoorthi, R. (2003). *Bayesian nonparametrics*. New York, NY: Springer Series in Statistics.

Hall, P. (1988). On symmetric bootstrap confidence intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, *50*(1), 35–45.

Hoaglin, D. C., Mosteller, F., & Turkey, J. (1985). *Exploring data tables, trends, and shapes.* Hoboken, NJ: John Wiley & Sons.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101.

Huber, P. J. & Ronchetti, E. M. (2009). *Robust statistics.* John Wiley & Sons Inc.

JASP Team. (2016). Jasp (version 0.7.5.5)[computer software].

Karabatsos, G. (2017). A menu-driven software package of Bayesian nonparametric (and parametric) mixed models for regression analysis and density estimation. *Behavior Research Methods*, *49*(1), 335–362.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General*, *142*(2), 573–603.

Kruschke, J. K. (2014). *Doing Bayesian data analysis A tutorial with R, JAGS, and Stan.* San Diego, CA: Academic Press.

Kruschke, J. K. & Meredith, M. (2015). *BEST: Bayesian estimation supersedes the t-test.* R package version 0.4.0. Retrieved from https://CRAN.R-project.org/package=BEST

Lee, M. D. & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course.* New York: Cambridge University Press.

Mauldin, R. D., Sudderth, W. D., & Williams, S. (1992). Polya trees and random distributions. *The Annals of Statistics*, *20*(3), 1203–1221.

Morey, R. D. & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes factors for common designs.* R package version 0.9.11-1. Retrieved from http://CRAN.R-project.org/package=BayesFactor

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*(2), 101–122.

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.

Pitman, J. & Yor, M. (1981). Bessel processes and infinitely divisible laws. In W. D. (Ed.), *Stochastic integrals. lecture notes in mathematics* (Vol. 851, pp. 285–370). Berlin, Heidelberg: Springer.

Ratcliff, R. & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481.

Rouder, J. N. & Morey, R. D. (2012). Default Bayes Factors for model selection in regression. *Multivariate Behavioral Research*, *47*(6), 877–903.

Rubin, D. B. et al. (1981). The Bayesian bootstrap. *The Annals of Statistics*, *9*(1), 130–134.

Tukey, J. W. (1960). Contributions to probability and statistics: Essays in honor of Harold Hotelling. In I. Olkin (Ed.), (pp. 448–485). Stanford, CA: Stanford University Press.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wang, C. & Blei, D. M. (2015). A general method for robust Bayesian modeling. *Bayesian Analysis*, *In Press*.

Wetzels, R., Raaijmakers, J. G., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, *16*(4), 752–760.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Amsterdam: Academic Press.

Zeigenfuse, M. D. & Lee, M. D. (2010). A general latent assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*, *54*(4), 352–362.

Chapter 5

When not everyone responds the same way

Empathy is a tool for building people into groups

Neil Gaiman

In the previous chapter I argued for modelling contamination as a process of the way the data were generated. This does not just increase the accuracy of the claims that can be made about the data. It also allows us, as researchers, to make additional claims about the data. In this chapter I extend this idea further with an example that combines the *assumptions* that the researcher makes about the data with the *inference* they would like to make.

We will consider an intervention trial, where the outcome of interest is the difference between pre- and post intervention, focusing on the case where the researcher has some underlying belief that not all participants will be affected by the treatment. I argue that these beliefs should govern both the *structure* of the model and, through this structure, the *parameters* of interest.

Existing models that are commonly used with this sort of data focus on identifying individual change. They do so through a testing process, not through a model that assumes a heterogeneous response. In this chapter I demonstrate that we can use a Bayesian mixture model that directly models this assumption, which allows us to make claims about the data that are richer. With this model we do not need to limit ourselves to making inference about the individual level, but instead can compare and contrast the size and proportion of individuals effected by a given treatment. The previous chapter demonstrated that models based on assumptions of the researcher are more accurate and can make richer claims about the data. This chapter demonstrates that a model based on assumptions about how the data were generated changes the type of outcome variables of interest in a rich and positive way.

Statement of Authorship

Title of Paper	Reliable estimates of individual change with Bayesian latent class models
Publication Status	□ Published □ Accepted for Publication □ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Unpublished

Principal Author

Name of Principal Author (Candidate)	Lauren Ashlee Kennedy		
Contribution to the Paper	Developed initial concept for this study, developed model, ran all simulations, wrote the first draft and edited subsequent drafts.		
Overall percentage (%)	90%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Daniel J. Navarro		
Contribution to the Paper	I contributed to the main idea of the paper, helped to design the simulations, and helped to frame and edit the paper.		
Signature		Date	31/8/2017

Name of Co-Author	Amy Perfors
Contribution to the Paper	I contributed ideas about which simulations to run, how to frame the work, and edited the paper.

Signature		Date	31/08/2017

Name of Co-Author	Andrew T. Hendrickson		
Contribution to the Paper	I contributed to the main idea of the paper, helped to design the simulations, and helped to edit the paper.		
Signature		Date	06/09/2017

Name of Co-Author	Nancy Briggs		
Contribution to the Paper	I contributed ideas about simulation design and relevant literature		
Signature		Date	31/08/2017

Please cut and paste additional co-author panels here as required.

Reliable estimates of individual change with Bayesian latent class models

Lauren A. Kennedy

School of Psychology

University of Adelaide

Daniel J. Navarro

School of Psychology

University of New South Wales

Amy Perfors

School of Psychology

University of Melbourne

Andrew T. Hendrickson

Cognitive Science and Artificial Intelligence Group

Tilburg University

Nancy Briggs

Mark Wainwright Analytical Centre

University of New South Wales

5.3 Abstract

Psychology often relies on making inference about the average individual, which can hide or distort true effects if the effect is not homogeneous. Although there exist traditional modelling attempts to distinguish between individuals who experience significant change and those that do not, the model proposed in this paper uses current state-of-the-art statistical models, which can estimate at both the individual and group-level, as an alternative to current best practice. To achieve this we propose a Bayesian mixture model that allows the researcher to infer at a group-level the proportion of individuals who experience an effect, and the size of the effect if it is experienced, and at an individual level whether individuals have experienced change or not. We demonstrate the efficacy of this model with a four toy datasets, and demonstrate the excellent parameter recovery for three of the four datasets. We then discuss potential extensions to this model class to improve the model further.

Keywords: Clinical trials, statistical analysis, Bayesian methods, Reliable Change Indices

Reliable estimates of individual change with Bayesian latent class models

Psychological science is often the science of averages. As scientists, we care about the average response to given a specific intervention or the average relationship between two different factors. By considering averages we can summarise data from many people using a small set of easily interpreted numbers and use these to predict future events. However, using average effects in this way relies on assuming all individuals have roughly the same response – an assumption that may not necessarily hold, with important consequences. In this paper we present a new application for a class of statistical models that permits researchers to make estimates at both individual and group-levels simultaneously. Our focus is on designs in which a treatment (or intervention) is provided to a group of people, and the question of interest is whether the intervention produces an effect on a continuous-valued outcome (or response). Such designs are commonplace in medical and clinical settings, and the heterogeneity of treatment effects in these situations is well known. For instance, only one to two-thirds of individuals appear to respond to the first prescribed medication for depression (Little, 2009; Souery, Papakostas, & Trivedi, 2006). Heterogeneity also appears present in therapeutic settings as well (Ross et al., 2004).

Despite the existence of individual variability, many of the most common statistical methods focus not on individual-level effects but rather on estimating the mean difference between pre- and post-treatment. Unfortunately, if the treatment is not effective for all participants, then these methods can hide true effects. For example, a very effective treatment that only treats a small sub-section of the population may appear to have a negligible effect across the full population, even though for the people it *does* impact the effect can be very meaningful. Naturally, we are not the first researchers to point this out. Others (e.g., Hageman & Arrindell, 1999; Jacobson & Truax, 1991; Wyrwich, 2004, to name a few) have noted the need to distinguish between individuals who experience an effect and those who do not. They, along with others (see Ferrer & Pardo, 2014, for a review), have proposed various methods for identifying individuals who experience significant[1] change in a pre-post trial.

Such methods typically require a multi-stage approach. In the first stage, a traditional test at the population or group-level is undertaken (e.g., a paired-samples t-test). In the second stage, if there is some suspicion of individual differences in response to the treatment, a method of identifying these individuals is employed (e.g., reliable change indices, or RCI, Jacobson & Truax, 1991). In the final stage, the two groups are compared in terms of the proportion of individuals they impact, sometimes by summing the number of people they identify (e.g., Jacobson, Dobson, Fruzzetti, Schmaling, & Salusky, 1991), other times by using a more sophisticated method (e.g., Hageman & Arrindell, 1999).

This kind of approach has some limitations. As we describe in more detail later, one problem is that typical methods for detecting the individuals who have responded to treatment (such as RCIs) require an independent estimate of scale reliability

[1] This literature makes a key distinction between clinical significance and statistical significance: clinical significance occurs when a change is clinically relevant (e.g., an intervention moves someone into a different diagnosis category) while statistical significance occurs when it reflects true change rather than random measurement error. Our focus in this paper is on statistical significance, since establishing it is a necessary precursor to establishing clinical significance.

which may be difficult to accurately obtain. Another problem is that this kind of multi-stage approach can conceal interesting effects and the issues only compound with more than one treatment condition. For instance, suppose there is a between-condition difference in the proportion of individuals who experience a significant change; however, there is *not* a significant difference between conditions in the *average* change. This might occur due to differing effect sizes of the treatment across condition, but would be hard to detect according to the standard statistical approach.

These considerations suggest that the current state-of-the-art statistical approach lacks the sensitivity to detect certain situations of clinical interest. In this paper we apply Bayesian mixture models which address this shortcoming. This type of model allows the user to make simultaneous estimates at both the *individual level* (i.e., did the individual experience significant change or not) and the *group-level* (i.e., what proportion of people changed? what was the size of the effect for those who did change?). It also allows us to quantify and propagate uncertainty in a precise way.

In the first part of this paper, we consider a single-condition experiment in which the question of interest is whether an intervention leads to change over time. Our model identifies whether each individual has experienced statistically significant change and how large that change is. The next section addresses experiments with two conditions; as before, our model can identify the individuals who experience statistically significant change and quantify the size of that change. It now also quantifies the difference between conditions. We conclude by discussing the benefits of analysing individual change with Bayesian mixture models.

5.4 Single group pretest-posttest designs

As mentioned, current practice typically identifies statistically significant change using a two-stage process in which the overall effect at the group-level is identified first, and individuals are examined only if such an effect is not detected. Having argued that this is inefficient and can miss important cases, we consider here a unified *latent class* model that simultaneously assigns individual participants to responder and non-responder classes, and estimates parameters for the effect size among responders and prevalence of responders in the population.

We apply this approach to the simplest experimental design, one in which all participants are assigned to a single treatment group, and each participant is measured once at pretest and once at posttest, yielding measurements x_{i1} and x_{i2} for the ith participant, and the data are characterised as a single change score $\delta_i = x_{i2} - x_{i1}$ for each participant.[2] For these designs, there are a number of other approaches that could potentially be adopted: for instance, there are four methods discussed in Ferrer and Pardo (2014) that are applicable to a single group design: reliable change indices (RCI: Jacobson & Truax, 1991), the Wyrwich standardized difference (WSD: Wyrwich, 2004), the Gulliksen Lord Novick method (GLN: Hsu, 1996) and the Hageman Arrindell approach (HA: Hageman & Arrindell, 1999). However, all four of these methods rely on a measure of *scale reliability* to estimate the amount of measurement noise inherent to the outcome variable x. Intuitively this makes a good deal of sense, insofar as one aims to detect whether a particular change δ is

[2] It is important to note there are some dangers in analysing differences scores directly rather than modelling the raw measurements. We return to this topic later, but for the moment we begin with the simplest model.

suspiciously large relative to the measurement noise. Unfortunately, it is difficult using these methods to find estimates of scale reliability that yield appropriate Type 1 error rates, making it difficult to apply these methods in practice. Only test-retest reliability has been shown to do so (Ferrer & Pardo, 2014), but some caution is advisable (Martinovich, Saunders, & Howard, 1996). The good performance of test-retest reliability depends on the assumption that there are no changes that occur between test and retest. As the underlying inference problem at hand is precisely to determine whether such changes have occurred, this would seem rather unsatisfactory. More generally, of course, many experiments rely on outcome measures for which no test-retest reliability values are available: from a statistical perspective, it would be desirable to provide scientists with an approach that does not require any a priori estimate of scale reliability in order to provide useful results. The model we present in this section does precisely this.

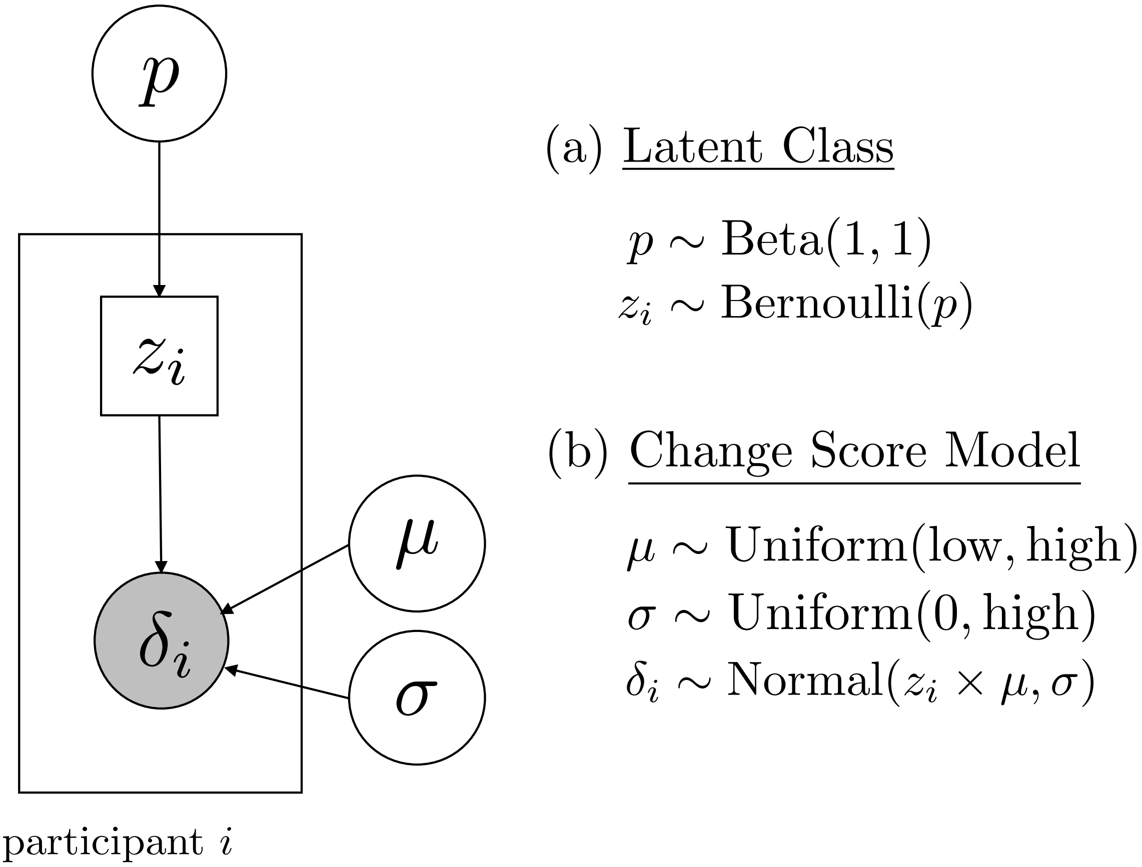
5.4.1 The Bayesian model

Currently the cutting edge technique in cognitive psychology is to utilise Bayesian mixture models to detect contamination (Kennedy, Navarro, Perfors, & Briggs, 2017; Zeigenfuse & Lee, 2010), quantify (Bartlema, Lee, Wetzels, & Vanpaemel, 2014; Danileiko & Lee, 2017; Lee, 2016) and detect individual differences (Dennis, Lee, & Kinnell, 2008). In this manuscript we suggest that this technique provides a beneficial alternative to reliable change models, which to our knowledge has not previously been suggested. Our approach assumes that the observed difference scores δ_i reflect a mixture of two *latent classes* indexed by a class assignment variable z_i.

Each participant is allocated either to the non-responder class ($z_i = 0$), indicating that no change has occurred, or to a responder class ($z_i = 1$) indicating that a change has occurred. The probability of belonging to the responder class p, the average effect size μ among responders, and the amount of variability σ in the data are all treated as unknown variables, and for the purpose of this investigation we place uniform priors (over a fixed range) for all three variables. The uniform prior was chosen to span the entire range of possible values for difference between times, and was set using information from the data (the maximum upper and lower change possible)[3]. While each of the particular modelling choices (e.g., normally distributed measurement error, uniform priors etc) could be questioned, they will be sufficient for our purpose, which is to highlight the value of using Bayesian latent class models for the detection of individual change.

The structure of this model is illustrated schematically in Figure 5.1, and – in essence it treats the observed difference scores δ as arising from a mixture of two normal distributions. One distribution corresponding to the non-responders and has fixed mean 0, the other correspond to responders and has unknown mean μ, and the relative weight assigned to these two components reflects the proportion p of people in the population that are responsive to treatment. However, as Figure 5.1 illustrates, we build the model by explicitly assuming that each participant is sampled from only one of these two distributions, using the class assignment variable z_i as an indicator variable. By formulating this mixture distribution as a hierarchical Bayesian model, we can use MCMC methods to estimate the joint posterior distribution over all

[3] Scales with definite upper and lower bounds could also be used to provide this information



(a) Latent Class

$$p \sim \text{Beta}(1,1)$$
$$z_i \sim \text{Bernoulli}(p)$$

(b) Change Score Model

$$\mu \sim \text{Uniform}(\text{low}, \text{high})$$
$$\sigma \sim \text{Uniform}(0, \text{high})$$
$$\delta_i \sim \text{Normal}(z_i \times \mu, \sigma)$$

Figure 5.1: A simple Bayesian mixture model for a single-group pretest-posttest design. Participants are assumed to fall in one of two latent classes (a), consisting of non-responders who show no systematic change ($z_i = 0$) and responders who show a systematic effect of the treatment ($z_i = 1$). The population rate of responding is an unknown variable denoted p. In (b), the observed difference score $\delta_i = x_{i2} - x_{i1}$ between the second and the first measurement is sampled from a normal distribution with standard deviation σ, but has mean 0 for non-responders and mean μ for responsive participants.

model parameters. Thus, given the observed data $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)$ for n subjects we are able to estimate not only the posterior distribution over the population level parameters p, μ and σ, but also the individual subject parameters z_i. Accordingly, for each person we obtain the Bayesian estimate $\phi_i = P(z_i = 1|\boldsymbol{\delta})$ corresponding to the posterior probability that the i-th person is responsive to the treatment.

The latent class model described above has a number of desirable properties: it is simple, it estimates population level parameters at the same time as it estimates individual subject parameters, and unlike standard methods in the field it does not require the researcher to have access to a "ground truth" measure of scale reliability. However, these properties are of little value if the method is unable to reliably detect individual change. In the next section we present a simulation study illustrating that our Bayesian latent class model is surprisingly good at estimating the *proportion* of people who experience the effect, categorising *individuals* as responders or non-responders, and estimating the *size of the effect* experienced by responders.

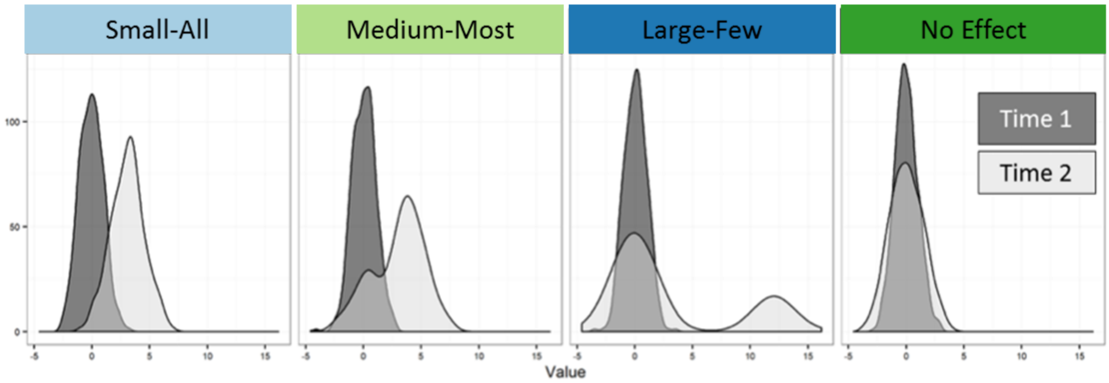
5.4.2 The data

In our simulations we consider four qualitatively distinct scenarios, illustrated schematically in Figure 5.2. In all instances we assume a single group pretest posttest design in which every participant is measured once prior to the treatment and once afterwards. The four scenarios are as follows. The first (Small-All) describes the situation where there is a small effect of the intervention, but it effects everyone. The second (Medium-Most) applies to a situation in which the intervention has a medium-sized effect but on only 75% of the participants. The third (Large-Few) corresponds to a situation where the effect is large but only for a minority (25%) of people. Finally, we also consider the case where the intervention has no effect on anyone (No Effect). To make the scenarios comparable, the datasets are designed to ensure that the average difference score is the same ($\bar{\delta} = 3$) in all scenarios except the No Effect dataset, which had $\bar{\delta} = 0$. To that end, in the Small-All case we set $p = 1$ and $\mu = 3$, the Medium-Most case sets $p = 0.75$ and $\mu = 4$, and finally Large-Few sets $p = 0.25$ and $\mu = 12$. Except where otherwise noted we report the performance of the model across 500 simulated datasets, implementing the model in JAGS, and using 1200 thinned samples (thinned every 100^{th} sample) from the posterior to estimate posterior quantities of interest, after a burn-in period of 3000.

5.4.3 Results: Individual subject parameters

In the first instance we consider cases where a moderately large number of observations are available (400 participants), and our first question is how well the model identifies the individuals who experienced a statistically significant change. For each person the model produces an estimate ϕ_i, corresponding to the proportion of posterior samples in which that the i-th participant was assigned to the responder group, $\phi_i = P(z_i = 1|\boldsymbol{\delta})$. To score the performance of the model we used a hard classification rule: if $\phi > .5$ the participant is labelled a responder, if $\phi \leq .5$ they were labelled a non-responder. This is convenient in order to compute a simple measure of model performance (i.e., percentage of correct classifications) but in many applications it might be most appropriate to simply report the ϕ_i values.

As illustrated in Figure 5.3, the model was highly successful in assigning participants



p	100%	75%	25%	0%
μ	3	4	12	0
σ	1	1	1	1

Figure 5.2: Four sets of simulated data given to the model. Each of these datasets is intended to capture a common situation confronting a researcher who studies the effect of an intervention on participants. The initial measurement x_{i1} is drawn from a normal distribution and shown with dark shading. The intervention occurs and then another set of measurements x_{i2} is taken at time 2 (light shading), yielding a set of difference scores δ_i. While the simulated datasets Small-All and No Effect are unimodal (indicating that all participants experienced either a small effect or no effect, respectively), datasets Medium-Most and Large-Few represent data where not all participants experience the effect. Below each figure the parameters used to simulate the data are shown: proportion of individuals who experience the effect (p), size of the effect (μ) and standard deviation of the effect (σ).

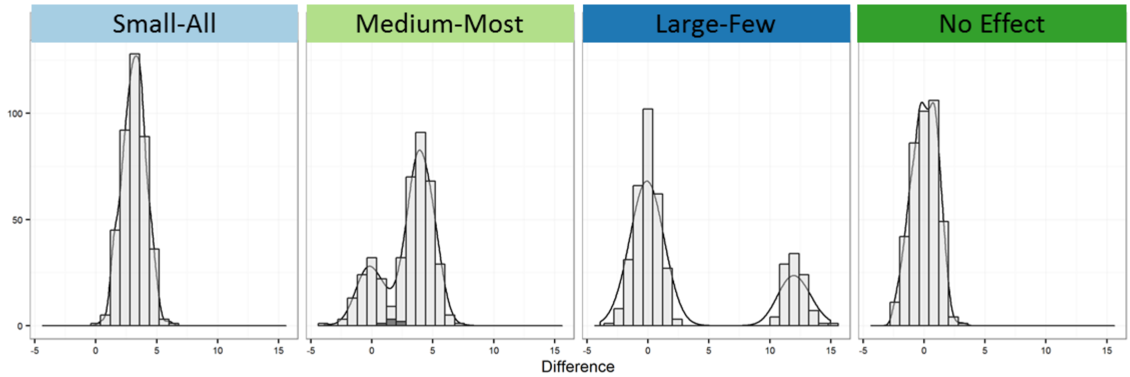


Figure 5.3: Model identification of individuals who experience change. This shows the same dataset as Figure 5.2, but here the x-axis represents the difference score δ between the time 2 and time 1 raw scores. The light shading represents the individuals who were correctly classified by the model. The dark shading represents the individuals who were incorrectly classified. All participants were correctly classified in all of the datasets except Medium-Most, which had difficulty for the 7% of participants whose small effect size was compatible with both RESPONDER and NON-RESPONDER latent classes.

to the correct latent class (RESPONDER or NON-RESPONDER) in most cases. For the Large-Few, Small-All and No Effect datasets, each participant was correctly identified. For the Medium-Most dataset, there was some overlap between participants at the margins, leading to some misclassification. In total, 7% of the participants were misclassified by the model; these were the ones who experienced such a small amount of change that they could plausibly be classified as either RESPONDER or NO-RESPONDER.

So far we have shown that a Bayesian latent class model correctly classifies most of the individuals in these simulated datasets. However, this simulation study corresponds to a relatively easy inference problem because the effect size is fairly substantial and the sample size is relatively large. With this in mind we repeated the previous simulation using sample sizes ranging from 24 to 400 to see how the

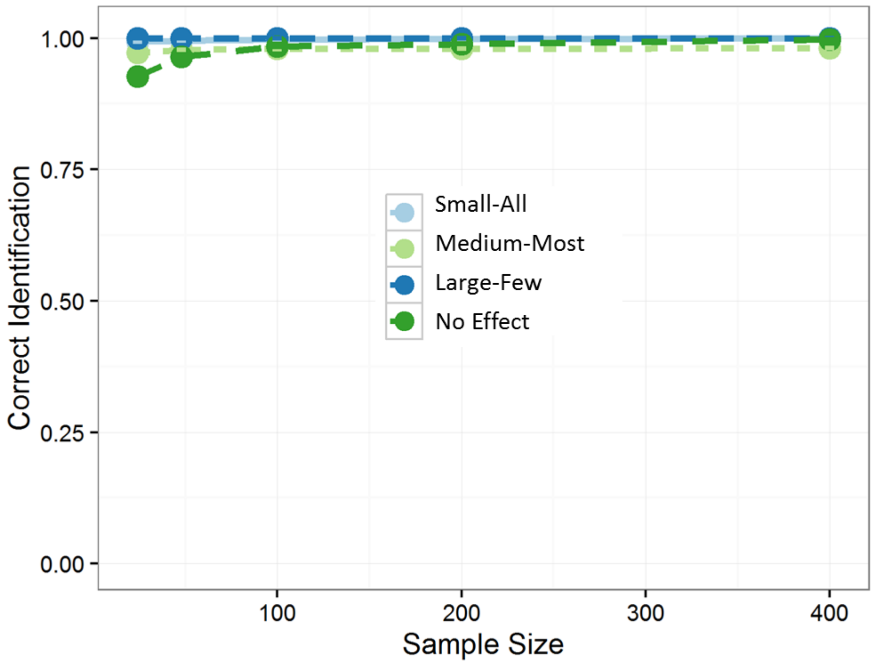


Figure 5.4: Effect of sample size on the ability to correctly classify individual participants as RESPONDERS ($z_i = 1$) and NON-RESPONDERS ($z_i = 0$). The y-axis shows the proportion of individuals who are correctly classified, while the x-axis indicates the sample size for the simulated dataset. Model performance on all four scenarios (indicated by the four colours) is very good, with the only departures from 100% accuracy being for the Medium-Most and No Effect datasets.

model performs when there are fewer observations available. As shown in Figure 5.4, the model performs well in the Small-All and Large-Few scenarios. It is somewhat less accurate in the Medium-Most, due to the overlap between the CHANGE and NO CHANGE groups. To some extent this is inevitable, as the population distribution over δ for responders and non-responders overlaps in the Medium-Most scenario, and as such perfect classification is impossible. The model is least accurate in the No Effect scenario where the model is misspecified.

5.4.4 Results: Population level parameters

As the previous section illustrates, a Bayesian latent class model is very successful at identifying individual responders and non-responders. Having verified this, we consider how effectively it estimates the two population level parameters of interest, namely the true probability of experiencing change (p) and the effect size among responders (μ). Figure 5.5 plots an example of the 95% highest density credible intervals (e.g., Kruschke, 2013) estimated for the population probability p against the sample size n,[4] and Figure 5.6 does likewise for the effect size μ. As is clear from inspection, when there is a true effect that is experienced by a non-negligible proportion of participants the credible intervals include the parameter value from which the data were simulated. As expected, with more data these intervals tighten, and importantly they do so while still containing this true parameter value. This indicates that in samples where there are differences in response, it is possible to estimate it with this model. This cannot be observed where there is No Effect. Although these intervals contain the true parameter value, they *do not* tighten with additional data. This occurs as a feature of the miss-specification between the model and the data. When no effect exists, the joint posterior distribution over p and μ concentrates around those values where $p \times \mu$ is small, which occurs whenever either of these parameters is small. Consequently, the model has some uncertainty about whether the proportion of responders p is near zero, the effect size μ is near zero, or both. The large credible intervals in Figure 5.5 and 5.6 are a reflection of this inherent ambiguity.

[4] Full simulations results can be found in Appendix 5.9

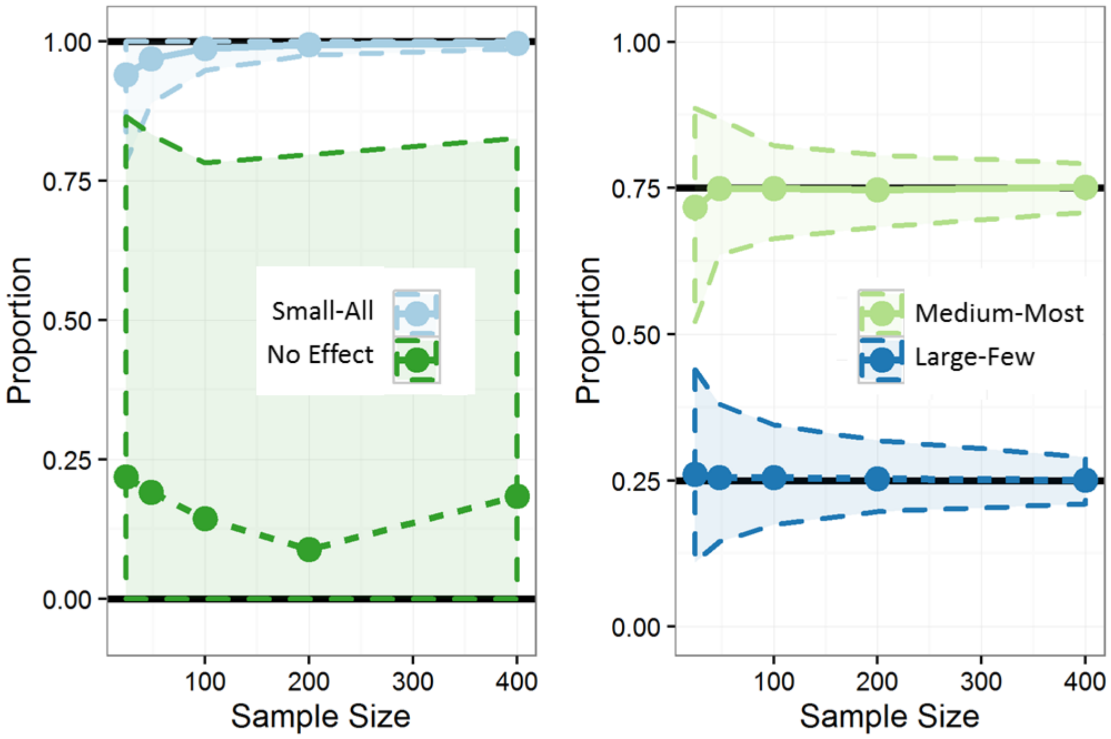


Figure 5.5: Effect of sample size on the estimated proportion p of people belonging to the RESPONDER class at the population level. The y-axis shows the model's estimate of p, while the x-axis reflects the sample size in the simulated dataset. The four different colours represent the four different datasets, split across two panels for ease of viewing. Black lines mark the true value. The left panel shows the datasets Small-All and No Effect, in which the effect on all participants was drawn from the same distribution. The model does very well at recovering the proportion of change in the Small-All case but has very large credible intervals when there is No Effect. As Figure 5.6 suggests, this is probably because the model also has high uncertainty around the effect size (which is zero). In essence, the model cannot decide between having a relatively high proportion of people in a CHANGE group with effect size of zero or having almost nobody in a CHANGE group with undetermined effect size. The panel on the right shows the datasets in which only a proportion of the individuals experienced an effect; the model accurately estimates that proportion both when it is 75% (Medium-Most) and 25% (Large-Few).

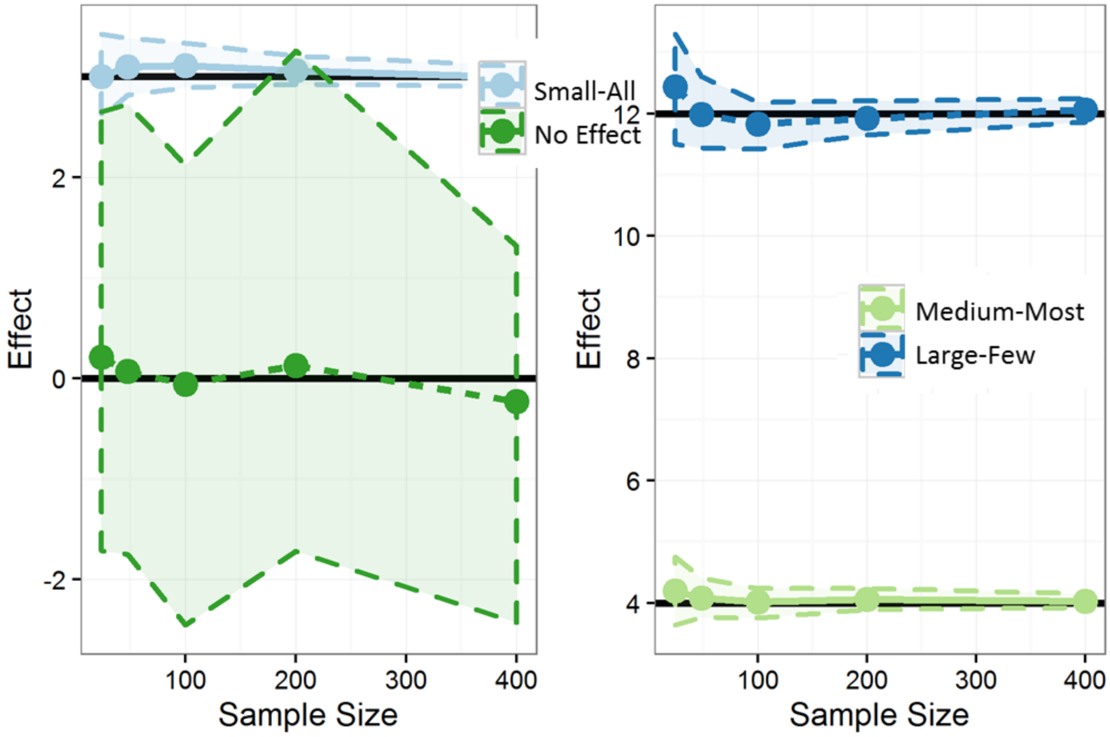


Figure 5.6: Effect of sample size on the estimation of the effect size μ among RESPONDERS. The y-axis shows the model's estimate of μ, while the x-axis reflects the sample size in the simulated dataset. The four different colours represent the four different datasets, which we separated over two figures for ease of viewing. Black lines mark the true value. The left panel shows the datasets Small-All and No Effect, in which the effect on all participants was drawn from the same distribution. The model does very well at recovering the effect size in the Small-All case but has very large credible intervals when there is No Effect, consistent with the results shown in Figure 5.5. The panel on the right shows the datasets in which only a proportion of the individuals experienced an effect; the model accurately estimates the effect size in both cases.

At first glance, this ambiguity between p and μ and the corresponding width of the credible intervals may seem undesirable. However, there is a sense in which it correctly reflects the actual knowledge that the researcher has obtained. To illustrate this, Figure 5.7 plots the (marginal) posterior distributions over the population level parameters p and μ and the individual subject assignments z_i for all four scenarios, for a small dataset with $n = 24$ participants. Of particular note are the results for the No Effect scenario: for every individual participant the model has accrued modest evidence that they belong to the NON-RESPONDER class (solid blue bars are all below the baseline of 0.5), and the posterior distributions over p (pink) and μ (yellow) have shifted towards 0 in both cases. However, the model "remains open" to the possibility that there is a very weak effect, or an effect that is experienced by only a very small proportion of people (which would have been undetectable in this sample), and so the population level estimates in particular are quite uncertain. In short, this is not a failure of the model so much as an inherent ambiguity in the inference problem.

5.4.5 Summary

Single group pretest-posttest designs are very prevalent in the literature, typically analysed at the aggregate level with paired samples t-tests. The significance of individual subject changes are rarely assessed, and when they are it is typically using methods such as RCI that require the researcher to have access to a ground truth measure of scale reliability (measurement error). However, in this initial investigation at least, it is clear that in a number of situations a Bayesian latent class

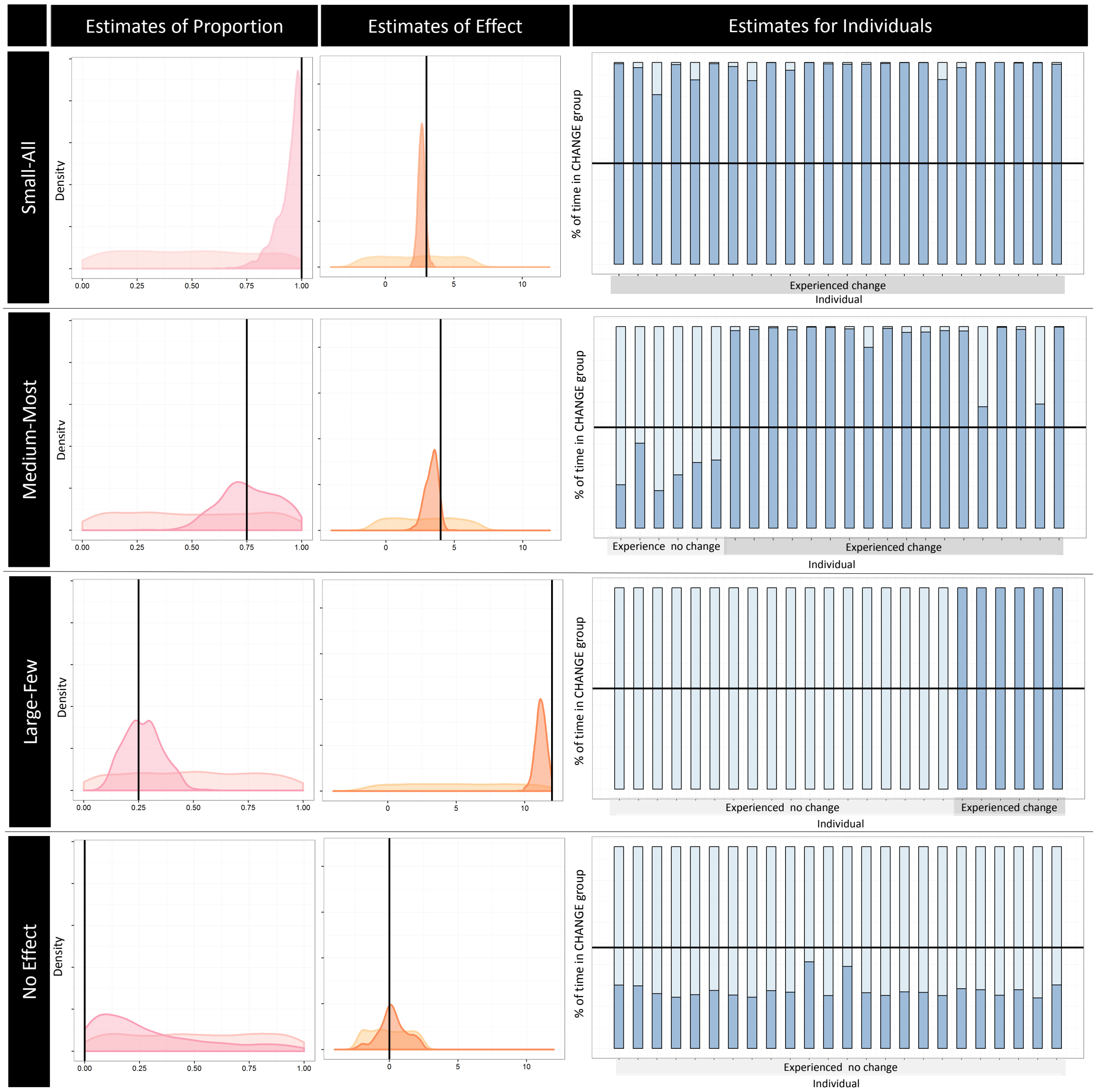


Figure 5.7: Group and individual estimates. This figure demonstrates model estimates for the group-level (namely, proportion and size of effect) and individual level. Datasets are represented along the rows, while group proportion(*pink tones*), group effect(*yellow tones*) and individual estimates(*blue tones*) form the columns. On the left hand side the plots compare the prior (*lighter tones*) to the posterior posterior(*darker tones*). On the right hand side, the figure shows the number of times participant was coded as changed (*dark blue*), or not changed (*dark blue*). This means the y-axis can be interpreted as the proportion of times coded as changed. The dark line represents 50%, our decision rule. On the x-axis of each figure the grey shading indicates the true group. While the model does not return a strong posterior for the No Effect dataset, it is still accurate at classifying at an individual level.

analysis is able to correctly estimate population level effects *and* correctly identify which participants are responsive to the treatment and which are not. Of course, the single group design is the simplest possible repeated measures design. Oftentimes experiments will compare two (or more) groups against each other over time, and in the next section we examine this situation. We develop a minor extension to the Bayesian latent class models that is appropriate to a multiple group situation, showing that it can be used to not just estimate individuals who have experienced change, but also to compare the different conditions in terms of the proportion and size of effect observed.

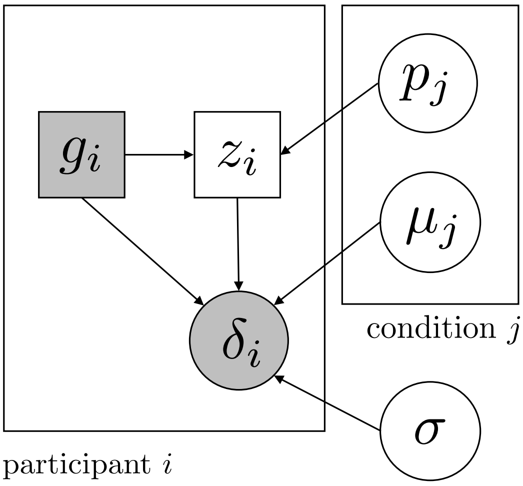
5.5 Multiple group pretest-posttest designs

The extension of the Bayesian latent class model to multiple group pretest-posttest designs is relatively straightforward. In this section we restrict ourselves to considering the situation where there are only two possible treatment conditions, and hence two groups, for the sake of simplicity. Once again, there are a number of existing methods that are applicable to these designs. However, two of the main methods, discussed in Ferrer and Pardo (2014), identify change by comparing it to a norm or control group. They are known as Standardised Individual Difference or SID (Payne & Jones, 1957) and Prediction Intervals or PI (Crawford & Howell, 1998). In these approaches, it is assumed that the control group is one in which there is no true change. This can be somewhat problematic in practice due to placebo effects, Hawthorne effects (McCarney et al., 2007) and so on. To the extent that changes within a control group need not be homogenous (e.g., if not everyone experiences the placebo effect) it is not safe to assume that the control group is in fact the "no change" group. In some cases a change would be expected to occur in the control group (e.g., practice effects), and in some instances a suppression of such effects (e.g., if practice effects vanish under some circumstances) might be the phenomenon of interest (e.g., Hinton-Bayre et al., 1999). With this in mind, we adopt a more general approach that does not require any one group to be declared a priori to be a no-change "reference group".

5.5.1 The Bayesian model

The structure of the Bayesian latent class model is illustrated in Figure 5.8, and is essentially identical to the original model in Figure 5.1. The only addition to the model is an observed group assignment variable g_i, indicating the treatment group to which the i-th participant is assigned. Each treatment is assumed to have its own population level parameters, p_j and μ_j, corresponding to the probability of responding to the treatment and the effect size among responders respectively. For simplicity we place independent priors over the parameters for each group.[5] As in the previous section, we are interested in exploring the model estimates at the population level (p_j and μ_j parameters) and the individual subject level (the z_i class assignment parameters).

[5] It would not be difficult to extend the model to add an additional layer to the hierarchical model, assuming that the population parameters are themselves drawn from a higher level distribution over possible effects, but we do not explore that extension in this paper.



(a) Latent Class

$$p_j \sim \text{Beta}(1,1)$$
$$z_i|g_i = j \sim \text{Bernoulli}(p_j)$$

(b) Change Score Model

$$\mu_j \sim \text{Uniform}(\text{low}, \text{high})$$
$$\sigma \sim \text{Uniform}(0, \text{high})$$
$$\delta_i|g_i = j \sim \text{Normal}(z_i \times \mu_j, \sigma)$$

Figure 5.8: Graphical model of the Bayesian latent class model for a two-group design. The subscript i denotes the participant, and j refers to the condition. Participants are allocated to one of two conditions by the researcher, so the group assignment g_i parameter is observed. The model contains population level parameters for the proportion of people p_j who are RESPONSIVE to the j-th treatment and the size of the effect μ_j they experience. For each participant, the model assigns $z_i = 1$ if they were responsive and $z_i = 0$ if they were not.

Table 5.1: Set of situations presented to the model. Each of the two conditions could have four distinct responses to the intervention (Small-All, Medium-Most, Large-Few, or No effect), producing the 10 distinct combinations with markers. The four conditions marked with ✗ correspond to situations where there is no difference between treatments, and both groups have the same population parameters. Scenarios marked with ✓ correspond to cases where one treatment is effective and the other is not. Finally, scenarios marked with $\sim$ correspond to cases where both treatments are effective (and the average change score is the same in both cases) but the pattern of change is different in the two groups.

		Condition 2			
		Small-All	Medium-Most	Large-Few	No Effect
Condition 1	Small-All	✗	$\sim$	$\sim$	✓
	Medium-Most		✗	$\sim$	✓
	Large-Few			✗	✓
	No Effect				✗

5.5.2 The data

In the previous section we considered four different scenarios that researchers might encounter: Small-All, where all the individuals improve a small amount, Medium-Most, where most individuals improve a moderate amount, Large-Few, where some individuals improve a large amount, and No Effect, where no individuals improve. We again consider these four different situations, but this time applied to both conditions, leading to the ten distinguishable scenarios listed in Table 5.1. Unless stated otherwise, the simulations and model performance measures in this section mirror those in the previous section: in particular, we retain the property that the average change score δ is identical in the Small-All, Medium-Most and Large-Few scenarios.

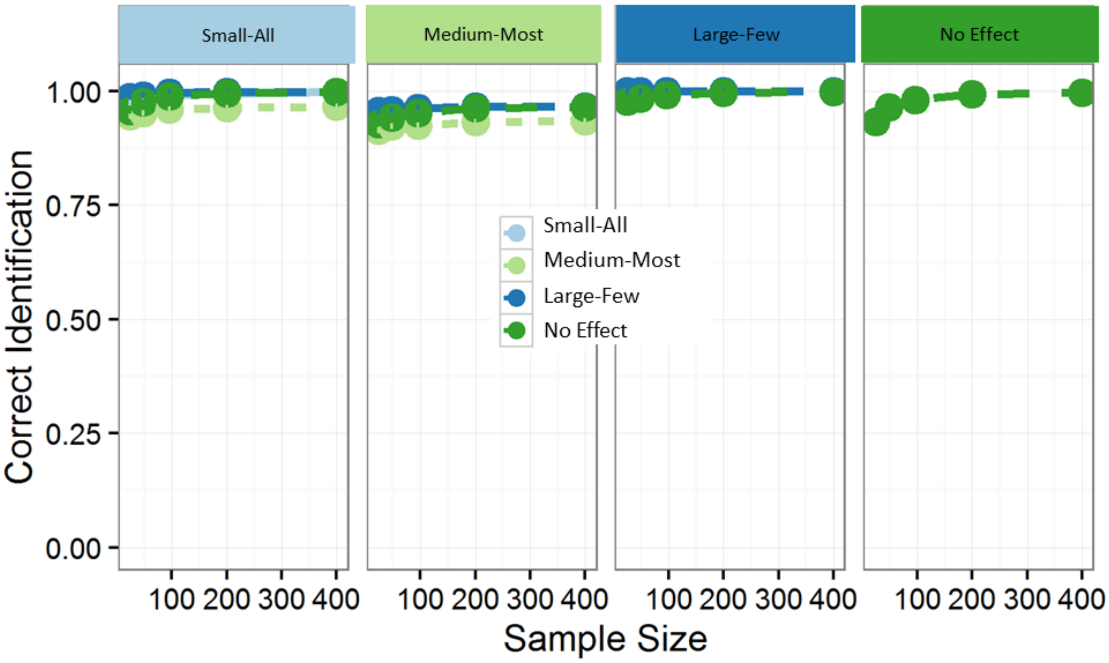


Figure 5.9: Performance of the Bayesian latent class model in assigning individual subjects to RESPONSIVE ($z_i = 1$) and NON-RESPONSIVE classes when there are two treatment conditions, plotted as a function of sample size. Panels represent the four qualitative patterns (e.g., Small-All, Large-Few) that might be encountered in group 1, and lines represent the pattern of change in group 2. In all situations, the model makes good classifications.

5.5.3 Results: Individual subject parameters

Perhaps unsurprisingly, the pattern of performance at the individual subject level is exactly as one might expect from the previous section. In our simulated datasets, individual subjects are correctly classified RESPONDERS or NON-RESPONDERS over 80% of the time and in the vast majority of cases performance is over 90%. Consistent with the previous simulation, the misclassifications arise most often when one or more groups has a Medium-Most structure, due to the overlap between the two classes in this scenario. Noting this, we turn to a consideration of the inferences that can be made at the population level.

5.5.4 Results: Population level parameters

To illustrate the value of the Bayesian latent class model when estimating population level effects in a multiple group design, consider how a standard independent samples t-test applied to the difference scores δ would behave when confronted with the 10 scenarios listed in Table 5.1. Assuming that sufficient data were available, a t-test has the capability to correctly detect population level differences when one of the two groups shows No effect (marked with ✓) and to retain the null when the two treatments produce identical effects (marked with ✗), but it cannot draw meaningful distinctions between treatments that produce qualitatively different *patterns* of change (those marked with $\sim$). Because the "two-stage" approach tends to rely on t-tests at the population level and separate tests (e.g., RCI) at the individual subject level, it is very difficult for this approach to handle these situations. A model that draws a distinction between the magnitude μ of an effect and the proportion p of participants that experience it is the only way to reconcile this pattern. [6]

To investigate the performance of the latent class model, we examine the credible intervals over population level differences in effect size $\mu_2 - \mu_1$ and proportion of people showing the effect $p_2 - p_1$. If the 95% credible interval contains 0 then we can conclude that it is likely that the proportion of individuals who experience change is not different between the two conditions. If it does not, then it is likely that there was a difference between conditions.

Here we chose to present credible intervals of the difference between the proportion and effect size in the two treatments. This approach allowed us to visually see the

[6] More generally, a statistical tool needs to be sensitive to the distributional shape and not merely the location parameter of a distribution.

effect of increasing the sample size, and also allows us to gain a greater understanding of how volatile the credible interval bounds are (do they smoothly decrease as the sample size increases or are they more jagged?). Here when the credible interval contains zero, we loosely interpret this as providing some evidence that there is no difference between the two treatments. This is not a formally correct way to compare a 'no difference' model against one where there is a difference. The credible interval is conditioned on the alternative hypothesis (much like how the frequentist p-value is conditioned on the null), and so we cannot find evidence for the null (as the frequentist cannot find evidence for the alternative) (Wagenmakers, Lee, Rouder, & Morey, 2015).

What alternatives are there for formal model comparison? The most formally correct method would be to use a Bayes Factor to compare the null model to the alternative model where there is a difference. In this case the two models are nested, and so we could use the Savage-Dickey method (Dickey, 1971) to compare the two. Unfortunately the bounds of the uniform prior (the prior for the effect size in our model) impact the magnitude of the Bayes Factor, making it hard to calculate in this instance. We used credible intervals here because they allowed a qualitative understanding of model behaviour, but future work should compare how to formally test a null hypothesis with this model.

We begin by considering those cases where no true difference exists between the two treatments, an example of which is plotted in Figure 5.10.[7] As is clear from inspection, the model tends to produce larger credible intervals when the sample size

[7] As before, full simulations results can be found in Appendix 5.9, at the conclusion of this chapter.

is small (as one would expect), but in all cases the interval includes zero, indicating that the model does not imply differences between the groups when there is none.

How does the model perform in situations where there is a genuine difference between the two treatments? As highlighted in Table 5.1, it is useful to distinguish between cases where one group shows an effect (be it Small-All, Medium-Most or Large-Few) and the other group shows No Effect, and cases where both groups show effects but those effects are of different kinds. In Figure 5.11 we again plot credible intervals for the two parameter differences of interest, namely $p_2 - p_1$ and $\mu_2 - \mu_1$. As is clear from inspection, the posterior distributions over these variables tend to be highly diagnostic for detecting those situations where both groups show effects, but of different kinds. In all such cases (left and middle panels), the credible intervals are narrow enough to exclude zero even at modest sample size. However, when one treatment produces No Effect (right panels), these quantities are largely unhelpful, with credible intervals including zero in many instances.

5.5.5 Summary

In extending the latent class model to multiple group pretest-posttest designs, the results are again encouraging. Even with a more complex experimental design, the model correctly identifies almost all of the individuals, in both conditions, who CHANGE with the intervention. The model also provides a means to compare the conditions based on the proportion p of individuals that change as well as the size of the effect if the change occurred. It is important to note that three of the four datasets would be indistinguishable using more traditional methods, since all showed

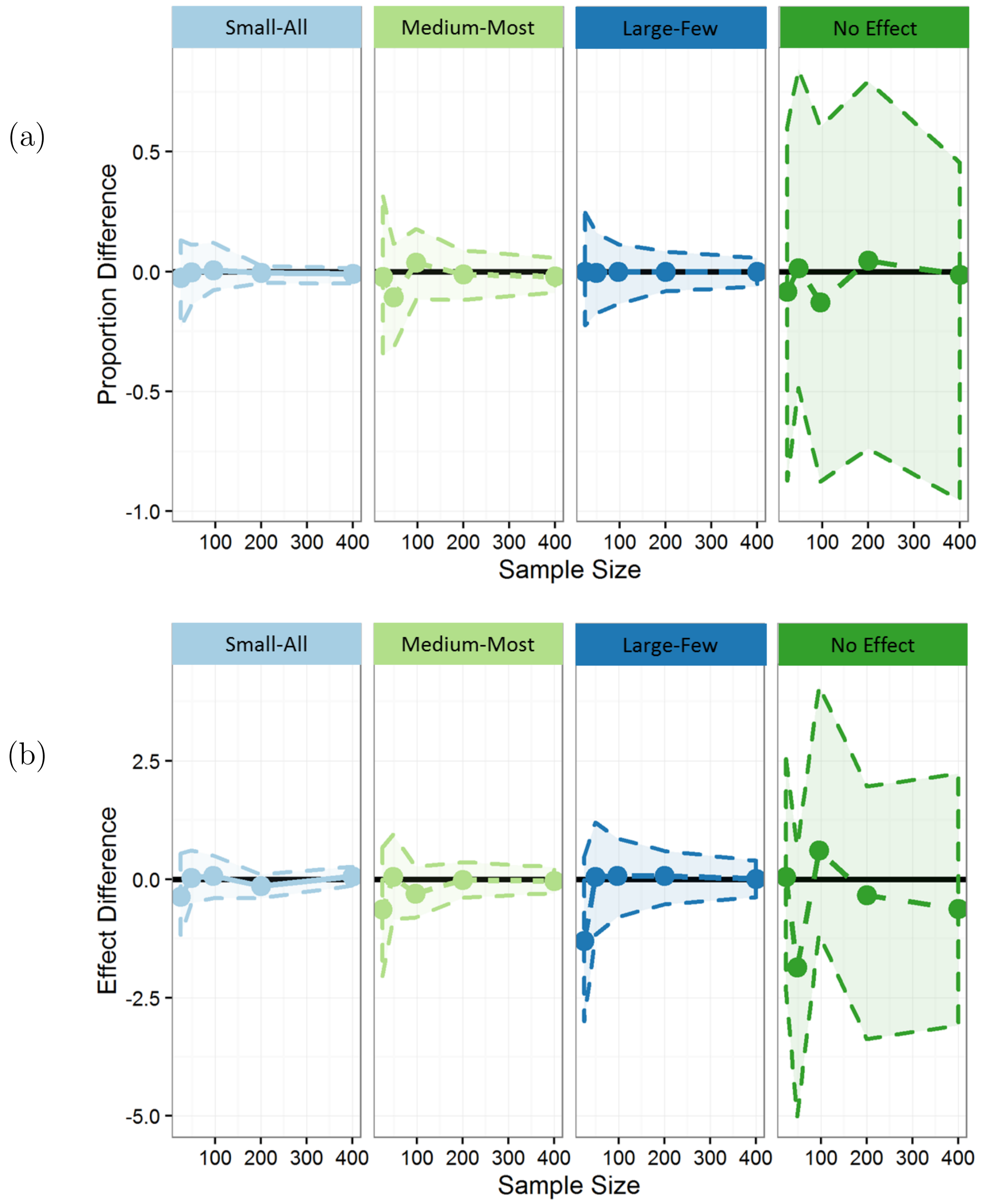


Figure 5.10: Performance of the Bayesian latent class model at the population level, when both treatment groups experience the same change (i.e., null effect of treatment, marked with ✗ in Table 5.1). The top row (a) plots an example of 95% credible intervals for $p_2 - p_1$, the difference between conditions in terms of the population probability of experiencing change. The bottom row (b) shows credible intervals for $\mu_2 - \mu_1$, the difference in the effect size. In each panel, the results are plotted as a function of sample size. The shaded area represents the credible interval area, with the dotted boundary representing the upper and lower bounds of this interval. In all cases the credible intervals contain zero.

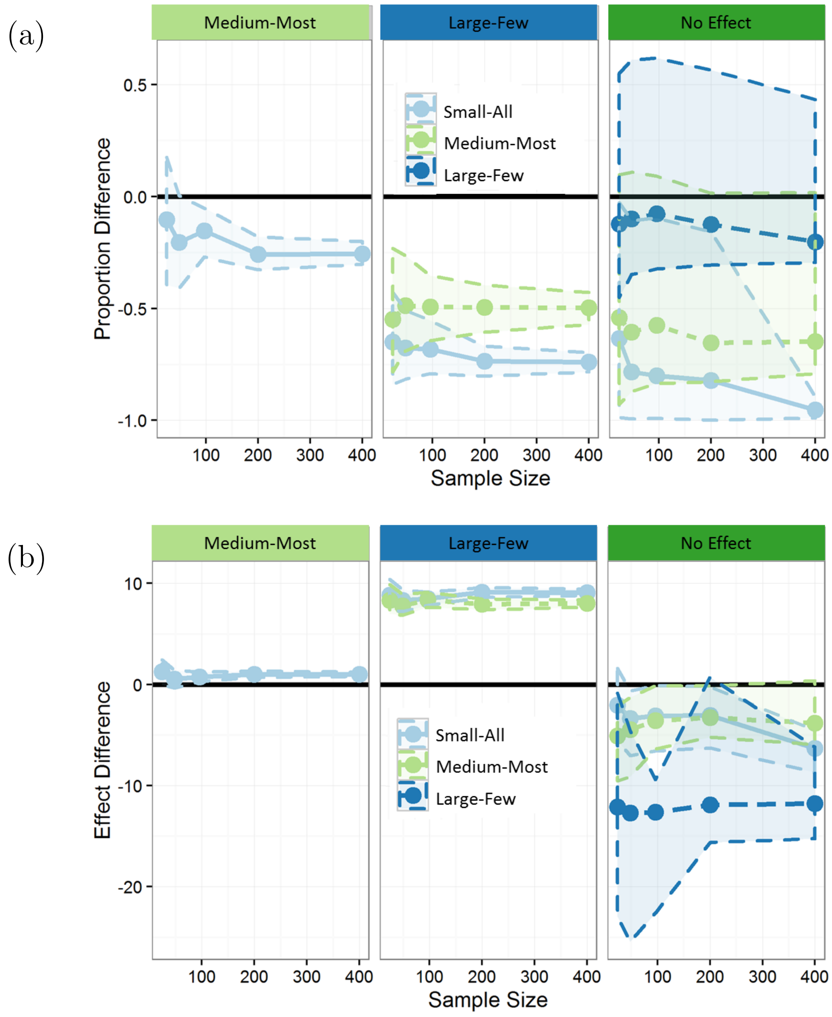


Figure 5.11: Performance of the Bayesian latent class model at the population level, when the pattern of change in the population is different (a true effect of treatment exists). The top row (a) an plots an example of the 95% credible intervals for $p_2 - p_1$, the difference between conditions in terms of the population probability of experiencing change. The bottom row (b) shows an example of the credible intervals for $\mu_2 - \mu_1$, the difference in the effect size. In each panel, the results are plotted as a function of sample size. The shaded area represents the credible interval area, with the dotted boundary representing the upper and lower bounds of this interval. In every case where both groups show an effect of a different kind (i.e., those marked with $\sim$ in Table 5.1; left and middle panels) the relevant credible intervals exclude zero, indicating that the model correctly describes this feature of the data. However, when one group shows no effect and the other does not (i.e., those marked with ✓ in Table 5.1; right panel) the credible intervals are quite large and often include zero.

the same mean change and only differed in the distributions. Overall, our model demonstrated low Type 1 and Type 2 errors, indicating that it could identify effects where they occurred but did not yield too many false negatives. The model was the least powerful when one of the conditions had no effect at all, which (as in Part I) reflects the difficulty in estimating two variables that are highly confounded with one another (the proportion of people experiencing a change and the effect size of the change).

5.6 Comparison with existing methods

Up to this point we have focused on illustrating the desirable qualitative features of adopting a Bayesian latent class approach, showing that it can reliably detect individual changes, correctly estimate population level effects, and draws a meaningful distinction between situations where all participants show a small effect and situations where a few people show large effects. Moreover, it correctly identifies individual changes without relying on an external ground truth measure of scale reliability, and as such may be of considerable use in many research contexts.

Arguably, the principal virtue of this approach is the fact that it can be applied in situations that other approaches cannot. Nevertheless, as our Bayesian approach is also applicable in situations to which other approaches can be applied, it seems sensible to ask how well it performs in relation to those cases. In this section we address this question, which means we must focus on datasets where we can treat one condition as a reference group (i.e., where at least one of the conditions had no effect). Note that this procedure is somewhat advantageous to existing approaches because it provides them strong information about the presence of a lack of effect. Although we *could* give the same information to the Bayesian latent class model, we choose not to for two reasons. Firstly, as we have seen, the model already does well at classifying individuals as RESPONDERS and NON-RESPONDERS without making this strong assumption. Secondly, as we have discussed, there are many scenarios in which we do not know whether such an assumption holds true. One of the main benefits of the Bayesian mixture model is that while it *can* use a strong assumption like this, it does not *need* to.[8]

What models should we compare our Bayesian mixture model to? There are two classes of model in the literature: those that require an estimate of scale measurement error (best estimated as a function of test-retest reliability), and those that require a reference group. Instead of including all of the alternative models, which is quite a list, we choose to focus on the most popular model within each group. The most commonly used model that requires test-retest reliability is the Reliable Change Indices, or RCI (Jacobson & Truax, 1991), while the most commonly used model that requires a reference group is the Prediction Intervals, or PI (Crawford & Howell, 1998). We provide details of these two models below.

[8] Of course, it would be trivial to modify the model to incorporate such an assumption if the researcher wanted by modifying the priors.

5.6.1 Existing methods

Reliable change indices (RCI)

The RCI method is one of the most popular ways of determining whether an individual has experienced statistically significant change, and has been shown to have a good Type 1 error rate (Jacobson & Truax, 1991). It estimates for each participant in the treatment group a Reliable Change Index which is calculated based upon the standard error measurement (SEM) of the scale as well as the degree of change that individual has experienced. If the absolute value of the RCI for that individual is greater (or less) than a threshold (most commonly ±1.96), then that person is considered to have experienced statistically significant change. The original RCI was later amended in order to estimate the SEM (Christensen & Mendoza, 1986; Ferrer & Pardo, 2014) based on the standard deviation of both the pre- and post-scores ($\mathrm{SD}_{\mathrm{pre}}$ and $\mathrm{SD}_{\mathrm{post}}$) as well as the reliability R of the scale. Equation 5.1 shows the full implementation that we used.

$$\mathrm{RCI}_{\mathrm{i}} = \frac{D_i}{\sqrt{(\mathrm{SD}_{\mathrm{pre}}\sqrt{1-R})^2 + (\mathrm{SD}_{\mathrm{post}}\sqrt{1-R})^2}} \tag{5.1}$$

We estimate the denominator of this equation using data from the No Effect dataset (which we always used for condition 1). The difference D between time 1 and time 2 scores is calculated in the treatment condition (condition 2), and the absolute value of the RCI for each participant in condition 2 is compared to the 1.96 criterion.

Prediction intervals

The PI method is also referred to as "confidence intervals for the individual predicted value" (Ferrer & Pardo, 2014) and unlike RCI it relies on a reference sample (in our case, we use the No Effect group in condition 1 as the reference sample). The prediction interval is derived by first plotting the time 1 scores against the time 2 scores for the reference sample. A linear regression model is then fit to these values, and the resulting regression line is used to create a 95% prediction interval (Crawford & Howell, 1998), shown in Equation 5.2 below. Prediction intervals are, by definition, wider than a confidence interval; whilst a 95% confidence interval contains the true parameter with 95% confidence, a 95% prediction interval includes a *new observation* with 95% confidence. The prediction interval is then used to identify the individuals in the treatment condition who experienced change: they are the ones whose scores fall outside of the prediction interval. The equation to calculate this interval for each new observation is included below.

$$\mathrm{PI_i} = \pm \frac{(Y_i - \hat{Y}_i)}{\mathrm{MSE}\sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}}} \tag{5.2}$$

5.6.2 Model evaluation

Both the RCI and PI models are trained on data from the reference group (condition 1), which consists of 400 pre-post samples from the No Effect dataset. They are then evaluated on how well they categorise 400 individuals drawn from the Small-All, Medium-Most, Large-Some and No Effect datasets. By contrast, our Bayesian mixture model is given both sets of data and estimates parameters in the same way as in previous sections, and we focus our analysis on the predictions it makes about individuals in the treatment condition (condition 2). This process is biased in favour of RCI and PI because it effectively gives them more information (i.e., the fact that there is no effect in condition 1) whereas our model is expected to infer that from the data.

Figure 5.12 compares all three models – RCI, PI, and ours – in terms of their performance identifying which individuals in the treatment condition were identified as being in the CHANGE group. The Bayesian mixture model is more accurate than the other two models for all datasets considered. Of course, accuracy in classification of individuals is only one possible measure of model performance. The Bayesian mixture model, unlike the RCI and PI, can also quantify the probability that any given individual has experienced significant change (rather than using a binary decision rule), the ability to estimate the proportion of individuals changed, and the size of the change for those individuals. In the next section we discuss several additional benefits of this approach.

5.7 Discussion

Moving beyond group-level change is a growing trend in research. With the understanding that not all individuals respond to a treatment in the same way, methods for identifying those individuals who have experienced change have been proposed. These methods all require (either directly or indirectly) a reference group where no change has occurred to quantify the inherent measurement error in the data. We suggest that in practice, a pure reference group is difficult if not impossible to ob-

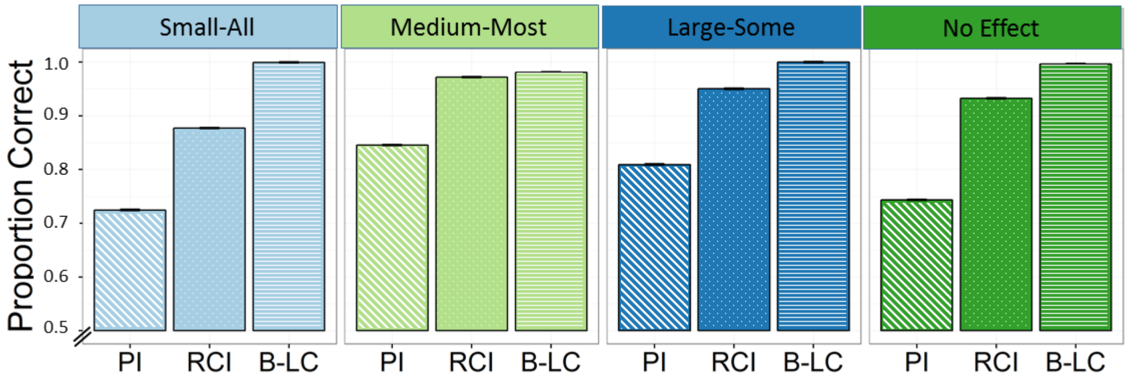


Figure 5.12: Comparison of three models at identifying individuals who experienced change. Models were compared in their accuracy in classifying the individuals in the treatment condition for four different datasets (panels); the control condition was always presumed to show No Effect. The three models were the Reliable Change Indices (*RCI*), Prediction Intervals (*PI*), and our Bayesian mixture model (*B-LC*). The B-LC was the most accurate in all cases.

tain. To the extent that this situation applies quite broadly within applied research, it is important to have statistical tools – such as our Bayesian latent class model – that can make sensible inferences about individual subject changes even in the face of real world concerns about the quality of reference data.

Based on simulated datasets reflecting four qualitatively distinct situations that a research might find themselves in, we showed that this model can accurately classify individuals as RESPONDERS who are sensitive to the treatment and NON-RESPONDERS who show no change from baseline. The model provides good group-level estimates of proportion p and effect size μ of those who do change, and enables researchers to compare conditions based on these parameters. Moreover, model performance was robust to variations in sample size and had relatively low Type 1 and Type 2 error rates. In this section we discuss two additional benefits of this modelling approach. Firstly, a number of benefits emerge because it is a *Bayesian* latent class model. Secondly, we show that it is relatively easy to adapt the model to handle different data analysis situations, and provide a few examples highlighting how this can be achieved.

One of the major difference between the previously described models and the one we implement here is that this model takes the form of a Bayesian latent class model. The benefits of the Bayesian framework of statistics has been well documented (see Kruschke, 2014; Wagenmakers, Morey, & Lee, 2016, for examples). In this section we wish to highlight some of the Bayesian benefits that are especially useful in the context of individual change. One particularly useful benefit is the fact that although the model expresses the class assignment as a binary variable z_i (reflecting the idea that people either do or do not respond to treatment), the posterior distribution over z_i expresses the full range of possible beliefs about each person. By defining $\phi_i = P(z_i|\boldsymbol{\delta})$ to correspond to the posterior probability that the i-th participant belongs to the RESPONDER group, the model is able to make graded assessments about each person. This is particularly evident in the individual change analysis presented in Figure 5.7, which highlights the extent of uncertainty that the model has about each individual.

A second benefit to the Bayesian latent class approach is that it does not require us to rely on a reference sample for which the researcher can safely assume that no change has occurred. By embedding the individual subject estimates within a hierarchical Bayesian framework that explicitly estimates the underlying change parameters p and μ we can draw inferences about the inherent noise in the measurement σ. Given this knowledge, the model is capable of constructing estimates of individual subject class memberships z_i that are as accurate as those produced by PI or RCI.

The final benefit to this Bayesian approach is that the inference method for the model is entirely generic, relying on standard MCMC routines implemented in packages such as WinBUGS, JAGS and Stan. All that the researcher is required to do is specify the structure of the model. This makes the framework highly extensible. In the current application, we have restricted ourselves to considering simple pretest posttest designs (and considered only the simplest situation in which the raw data are the difference scores $\delta_i = x_{i2} - x_{i1}$), but there are a much larger variety of repeated measures designs to which the approach could be extended. With very little work we can add more interventions and outcome measures by adding additional variables to the model. With slightly more work to change the structure of the model we can change the model to allow for wait-list or stepped-wedge designs. By making some assumptions about how the interventions will effect individuals over time (e.g linearly, exponentially etc.), we can modify the model to allow for multiple time points. In some contexts one might wish to replace the normally distributed pre- and post-scores with alternative distributions such as a t-distribution (for increased robustness; Kruschke, 2013), negative-Binomial (e.g., for count data; Holsclaw, Hallgren, Steyvers, Smyth, & Atkins, 2015), or even a logistic link function (for binary outcome data).

These Bayesian latent class models were inspired by Reliable Change Indices (Jacobson & Truax, 1991) and the notion that not all individuals change in a pretest-posttest design. Others have built upon this the initial model proposed by Jacobson and Truax (1991) to account for regression to the mean (Hageman & Arrindell, 1999). Regression to the mean is a statistical phenomenon where an individual is more likely to be closer to the population mean at the posttest than they were at the pretest (Healy & Goldstein, 1978). The model described in this manuscript *does not* account for regression to the mean, which we can see reflected in the categorization of extreme points in the No Effect data as changed. Changes to the model that reflect the structure of a repeated measures model, namely including independent measurement error at each time point in conjunction with either a fixed or random effect, are currently being discussed.

Of course, as with all methods of identifying individual change, care is needed in interpreting the results of these analyses. As individuals are clustered together into classes — those who experience change and those who did not – it may be tempting to conduct unplanned, post-hoc analyses to investigate whether there are any differences between the two groups. Some care is necessarily required in these situations. Even though the researcher might have an a priori hypothesis that people who experience change are different from those who do not (with respect to some covariate), the assignments of individuals to the RESPONDER and NON-RESPONDER classes is necessarily post-hoc even if the choice of covariates to compare is constrained by an a priori hypothesis.

These caveats and potential extensions notwithstanding, we suggest that Bayesian latent class models provide a valuable addition to the statistical toolbox when analysing the results of pretest-posttest designs. They are more broadly applicable than most alternative methods, and support a broader range of inferences at the individual subject and population level. If our goal is to draw inferences about changes experienced by people as well as by populations, models of this kind are extremely useful.

5.8 References

Bartlema, A., Lee, M. D., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150.

Christensen, L. & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy*, *17*(3), 305–308.

Crawford, J. R. & Howell, D. C. (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of Clinical and Experimental Neuropsychology*, *20*(5), 755–762.

Danileiko, I. & Lee, M. D. (2017). A model-based approach to the wisdom of the crowd in category learning. *Cognitive Science*, *Accepted*.

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language*, *59*(3), 361–376.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*(1), 204–223.

Ferrer, R. & Pardo, A. (2014). Clinically meaningful change: False positives in the estimation of individual change. *Psychological Assessment*, *26*(2), 370.

Hageman, W. & Arrindell, W. A. (1999). Establishing clinically significant change: Increment of precision and the distinction between individual and group level of analysis. *Behaviour Research and Therapy*, *37*(12), 1169–93.

Healy, M. & Goldstein, H. (1978). Regression to the mean. *Annals of Human Biology*, *5*(3), 277–280.

Hinton-Bayre, A. D., Geffen, G. M., Geffen, L. B., McFarland, K. A., & Frijs, P. (1999). Concussion in contact sports: Reliable change indices of impairment and recovery. *Journal of Clinical and Experimental Neuropsychology*, *21*(1), 70–86.

Holsclaw, T., Hallgren, K. A., Steyvers, M., Smyth, P., & Atkins, D. C. (2015). Measurement error and outcome distributions: Methodological issues in regression analyses of behavioral coding data. *Psychology of Addictive Behaviors*, *29*(4), 1031.

Hsu, L. M. (1996). On the identification of clinically significant client changes: Reinterpretation of Jacobson's cut scores. *Journal of Psychopathology and Behavioral Assessment*, *18*(4), 371–385.

Jacobson, N. S., Dobson, K., Fruzzetti, A. E., Schmaling, K. B., & Salusky, S. (1991). Marital therapy as a treatment for depression. *Journal of Consulting and Clinical Psychology*, *59*(4), 547.

Jacobson, N. S. & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19.

Kennedy, L. A., Navarro, D. J., Perfors, A., & Briggs, N. (2017). Not every credible interval is credible: Evaluating robustness in the presence of contamination in Bayesian data analysis. *Behavior Research Methods*, *49*(6), 2219–2234.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General*, *142*(2), 573–603.

Kruschke, J. K. (2014). *Doing Bayesian data analysis A tutorial with R, JAGS, and Stan*. San Diego, CA: Academic Press.

Lee, M. D. (2016). Bayesian outcome-based strategy classification. *Behavior Research Methods*, *48*(1), 29–41.

Little, A. (2009). Treatment-resistant depression. *American Family Physician*, *80*(2), 167–72.

Martinovich, Z., Saunders, S., & Howard, K. (1996). Some comments on "Assessing clinical significance". *Psychotherapy Research*, *6*(2), 124–132.

McCarney, R., Warner, J., Iliffe, S., Van Haselen, R., Griffin, M., & Fisher, P. (2007). The Hawthorne effect: A randomised, controlled trial. *BMC Medical Research Methodology*, *7*(1), 30–38.

Payne, R. & Jones, H. G. (1957). Statistics for the investigation of individual cases. *Journal of Clinical Psychology*, *13*(2), 115–121.

Ross, J. S., Fletcher, J. A., Bloom, K. J., Linette, G. P., Stec, J., Symmans, W. F., . . . Hortobagyi, G. N. (2004). Targeted therapy in breast cancer the HER-2/neu gene and protein. *Molecular & Cellular Proteomics*, *3*(4), 379–398.

Souery, D., Papakostas, G. I., & Trivedi, M. H. (2006). Treatment-resistant depression. *Journal of Clinical Psychiatry*, *67*, 16–22.

Wagenmakers, E.-J., Lee, M. D., Rouder, J., & Morey, R. (2015). Another statistical paradox. *Manuscript submitted for publication.*

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169–176.

Wyrwich, K. W. (2004). Minimal important difference thresholds and the standard error of measurement: Is there a connection? *Journal of Biopharmaceutical Statistics*, *14*(1), 97–110.

Zeigenfuse, M. D. & Lee, M. D. (2010). A general latent assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*, *54*(4), 352–362.

5.9 Appendix

In the text we presented examples of credible intervals for different sample sizes and different datasets. Here I extend this with the results of two simulation studies; one investigating informativeness and accuracy of the credible interval for population level parameters in the single group scenario; and the second to investigate estimation of the difference between groups for the two-group scenario.

Figure 5.13 demonstrates the average accuracy and precision (as interval width) of the credible intervals produced by the B-LC model for the single group scenario. As we saw with the example credible intervals in the text, the model is very likely to produce a credible interval of the proportion and effect size that contains the true parameter value. The model struggled most with the No Effect scenarios, producing much wider credible intervals. As we discussed in the main text, this is to be expected.

With this in mind, we turn our attention to the two-group simulation. As before, although the text reported example credible intervals, we used the B-LC model with 500 iterations of these comparisons. Figure 5.14 shows the average accuracy and width for credible intervals around the difference in proportion, while Figure 5.15 shows the same for the difference in effect size. We find greater evidence for our two main findings from the main text for difference in proportion estimates; namely low power and wider credible intervals for comparisons against the No Effect scenario. We also see evidence for the findings in the main text that suggest estimating the difference in effect size had higher power than estimating the difference in proportion.

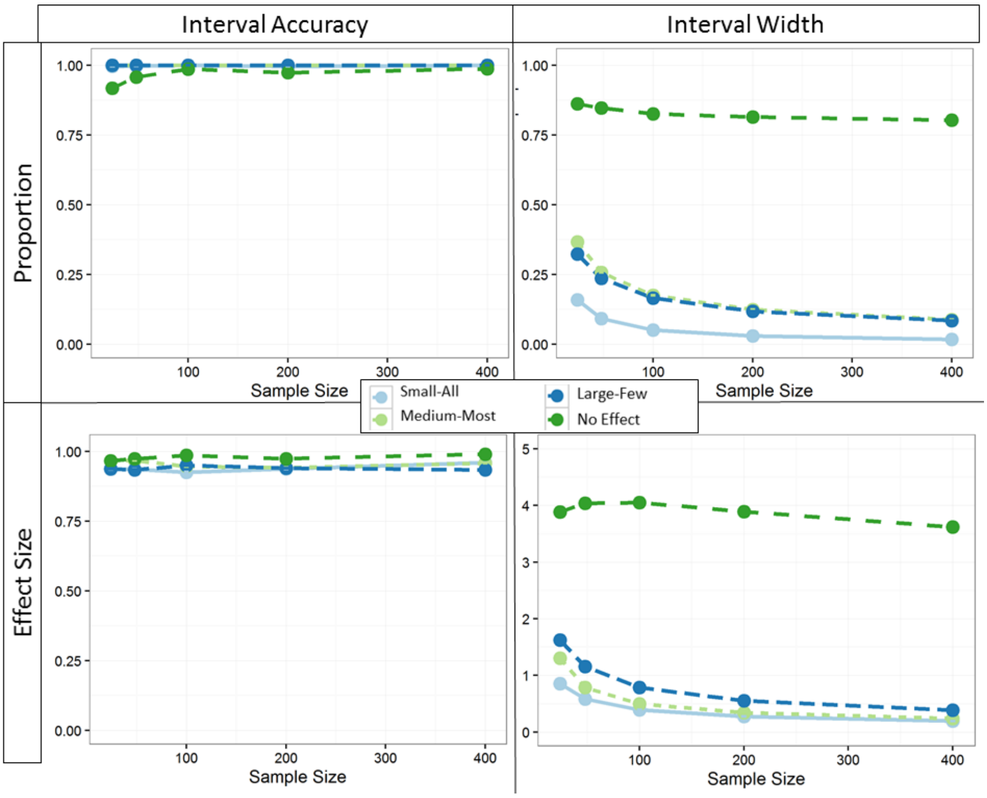


Figure 5.13: Ability of the model to correctly estimate the true parameter in a single group scenario. Five-hundred of the four example datasets were generated with 5 different sample sizes. For each the B-LC model was used to estimate an interval for which we are 95% confident contained the true proportion of individuals effected (*first row*) and true effect size for those who were effected (*second row*). Each interval was recorded in terms of its accuracy (*first column*) – whether it contained the true parameter or not – and its width (second column). Even though the B-LC model produced wider intervals for the No Effect dataset, all intervals were very likely to contain the true parameter.

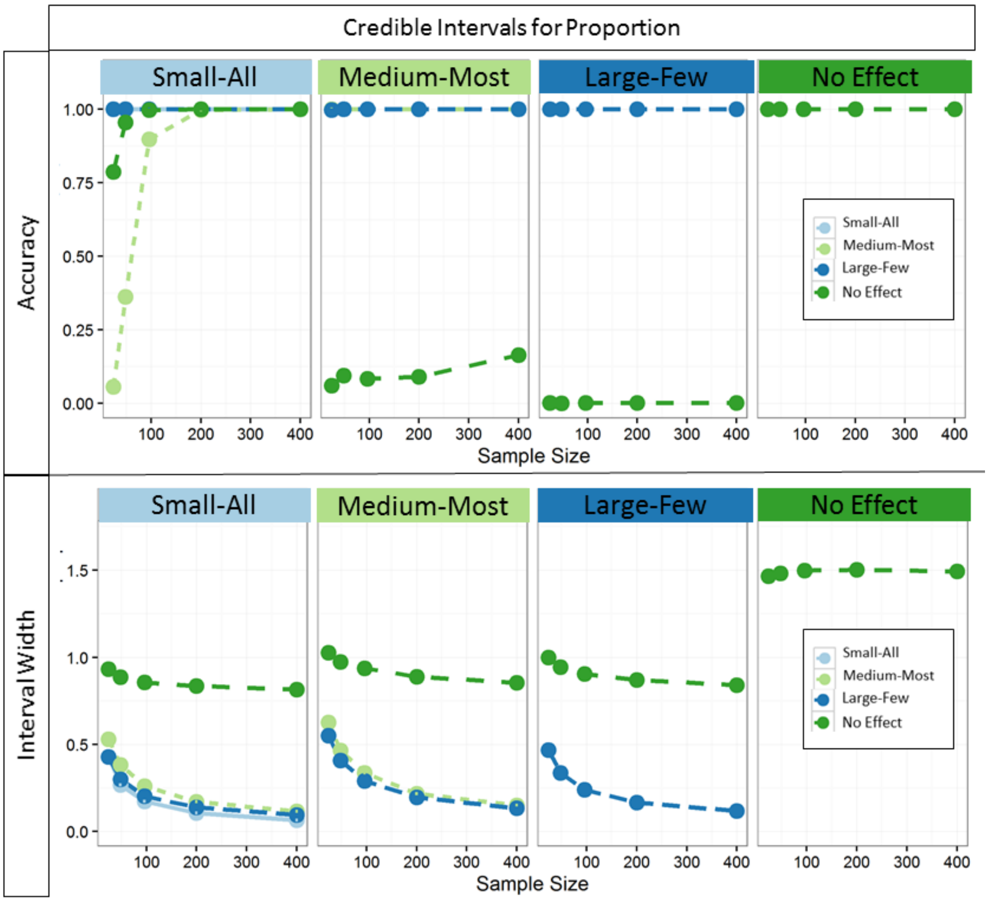


Figure 5.14: Ability of the model to correctly estimate a significant difference between proportions in a two group scenario. Five-hundred of the dataset comparisons were generated with different sample sizes. For each the B-LC model was used to estimate an interval for which we are 95% confident contained the true difference between proportion of individuals effected in group one and the proportion in group 2. Each interval was recorded in terms of its accuracy (*first row*) – whether it contained the true difference or not – and its width. Overall the B-LC model produced intervals that were very accurate for all pairings except those comparing the No Effect to the other datasets.

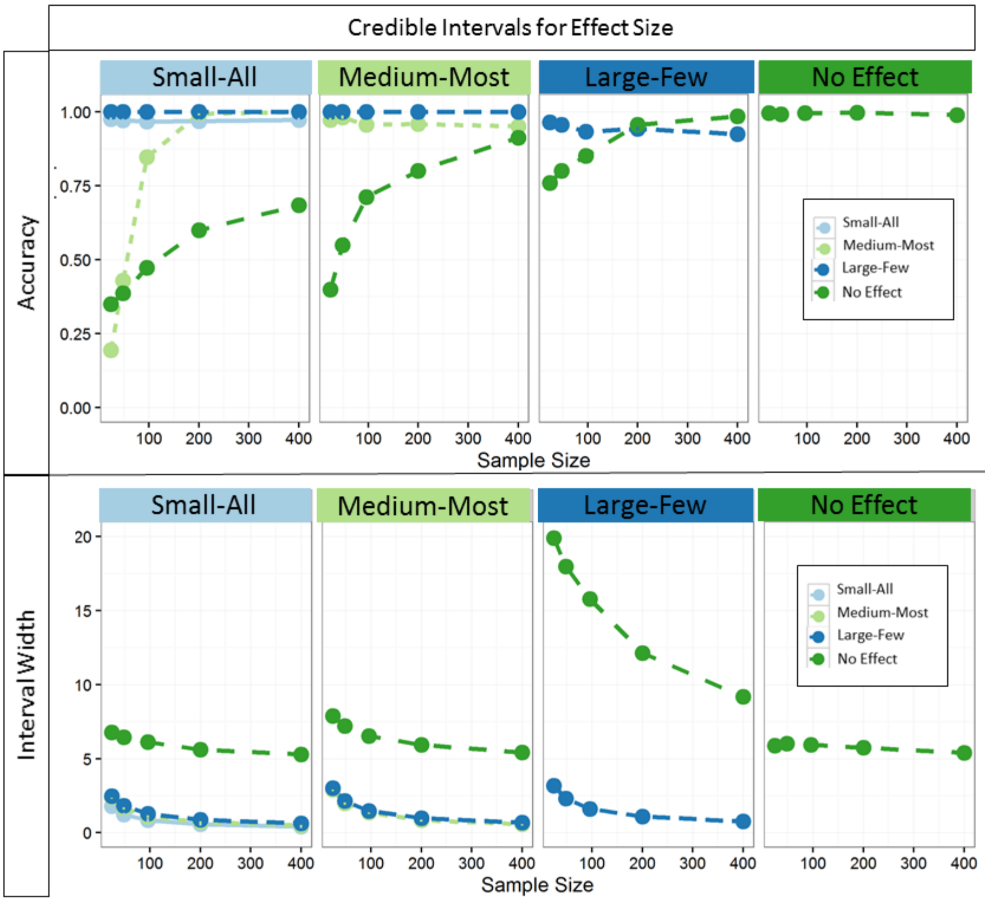


Figure 5.15: Ability of the model to correctly estimate a significant difference between effect size in a two group scenario. Five-hundred of the dataset comparisons were generated with different sample sizes. For each the B-LC model was used to estimate an interval for which we are 95% confident contained the true difference between the effect size in group one and the effect size in group 2. Each interval was recorded in terms of its accuracy (*first row*) – whether it contained the true difference or not – and its width. Overall the B-LC model produced intervals that grew in accuracy with increased sample size, and relatively narrow intervals for all but the Large-Few and No Effect combination.

Chapter 6

When there is missing data

> It is possible at times to develop a mathematical analysis based on a more complex set of assumptions,... (t)his is more troublesome in many ways than the analysis based on simple assumptions, but where feasible it is to be preferred.
>
> *Lee J. Cronbach*

In the previous chapters of this thesis I moved from a model that made few assumptions about the data (chapter three) to models that incorporated assumptions or knowledge the researcher believes to be true about the data (chapters four and five). However, there is substantial evidence from the cognitive sciences to suggest that people are susceptible to making a number of incorrect decisions based upon the data they have observed (e.g., Ellsberg, 1961; Epley & Gilovich, 2006; Landy et al., 2013). If this is true, then perhaps Cronbach is incorrect: instead of moving towards analysis based on complex assumptions about the data, we should move *away* from complex models to more simple models.

In this chapter I consider the well-known effect of ambiguity aversion (Ellsberg, 1961). Whilst this effect is often used to demonstrate that people have an aversion to ambiguous information, in practice it requires that the participants estimate information that is missing. As I have already discussed in the introduction, statisticians categorise missing data into three causes; missing completely at random, missing at random and missing with some underlying process. Participants in ambiguity aversion tasks, however, have been shown to be relatively resistant to various manipulations that suggest different causes of missing data (e.g., Garcia-Retamero, Müller, Catena, & Maldonado, 2009), behaving with similar disdain for ambiguous options regardless of the source of ambiguity. In this work I explore whether this is caused by aversion to ambiguous information or incorrect judgements about the underlying missing data. Our results suggest that ambiguity aversion remains robust against different levels of ambiguity and many of the prior manipulations. Despite this, participants were remarkably sensitive to manipulations of the *source* and *distribution.*

This has interesting implications for both ambiguity aversions as well as for the use of researcher assumptions in statistical data analysis. Based only on the judgements of participants, we should not trust researchers to make accurate assumptions about their data. However, when we consider the estimates participants make about the underlying data, they are quite sensitive to the underlying distributions.

Statement of Authorship

Title of Paper	Priors, informative cues and ambiguity aversion	
Publication Status	□ Published	□ Accepted for Publication
	□ Submitted for Publication	☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Unpublished	

Principal Author

Name of Principal Author (Candidate)	Lauren Ashlee Kennedy		
Contribution to the Paper	Contributed to the original core concept of this experiment (in conjunction with AP), as well as the more specific experimental design (with DN and AP). Coded the website to collect the data; analysed and produced the figures in the manuscript with discussion with AP and DN. Wrote first draft of manuscript and contributed to revisions.		
Overall percentage (%)	75%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Amy Perfors		
Contribution to the Paper	I contributed to the main idea of the paper, helped to design the experiments, and helped to frame and edit the paper.		
Signature		Date	31/08/2017

Name of Co-Author	Daniel J. Navarro
Contribution to the Paper	I contributed to the main idea of the paper, helped to design the experiments, and helped to frame and edit the paper.

Signature		Date	31/08/2017

Please cut and paste additional co-author panels here as required.

Priors, informative cues and ambiguity aversion

Lauren A. Kennedy

School of Psychology

University of Adelaide

Amy Perfors

School of Psychology

University of Adelaide

Daniel J. Navarro

School of Psychology

University of New South Wales

6.2 Abstract

Ambiguity aversion, or the preference for options with known rather than unknown probabilities, is a robust phenomenon within the decision making literature (e.g., Camerer & Weber, 1992). There are some suggestions that this aversion is due to people inferring differences in the prior distribution for the ambiguous option (Güney & Newell, 2015). In this study we investigated the relationship between prior distributions and experienced information cues on people's decision making and their judgments about the underlying distribution in the ambiguous item. We used three different prior cues; POSITIVE (suggesting a positive underlying distributional cue), NEUTRAL (no distributional cue) and NEGATIVE (suggesting a negative underlying distributional cue) and five different information cues, varying both the bias of the information and the degree of ambiguity. While we found that both prior and information manipulations had the expected impact for participants' judgements of underlying distributions, they only impacted the decisions participants made some of the time.

Keywords: ambiguity; uncertainty; priors; information

6.3 Introduction

> *When dealing with real data, the practicing statistician should explicitly consider the process that causes missing data far more often than he [sic] does.*
>
> – Donald Rubin, 1976

A classic finding in the human decision making literature is that people are *ambiguity averse*: when presented with a choice between options that ostensibly offer the same expected reward, one of which is specified in more precise terms than the other, people typically prefer to select the less ambiguous option (Camerer & Weber, 1992; Ellsberg, 1961). For example, people will typically prefer to bet on the flip of a fair coin – where the probability of heads is known to be 50% – rather than bet on the outcome of a weighted coin, where the bias θ on the coin is unknown. However, following Laplace's "principle of insufficient reason", a Bayesian reasoner should specify a symmetric prior $P(\theta)$ over the coin bias, leading to the conclusion that in the absence of any other information, the two bets are equivalent for all practical purposes. Human decision makers rarely show this indifference, with the majority of people preferring to avoid the ambiguity inherent in the second example (Liu & Colman, 2009).

Why does this occur? One possibility is that people tend to be pessimistic about their prospects when ambiguity is present. On the surface, the ambiguity in the biased coin scenario seems innocuous, but it need not be so. If a professional stage magician were offering you the bet – or a social psychologist for that matter – you might have reason to be suspicious. Perhaps they know something about the situation that you do not. Ambiguous scenarios maximize the potential for malfeasance, and a cautious decision maker might be wise to avoid them. This tension can be seen in the "tennis match" scenarios discussed by Gardenfors and Sahlin (1982): if you and I both know that players A and B are matched in skill and have played each other many times, you have far fewer opportunities to take advantage of me by virtue of superior knowledge than if the two contestants have never played each other. When competing against others, missing data matters and ambiguity aversion seems reasonable, because the things you do not know can be used against you.

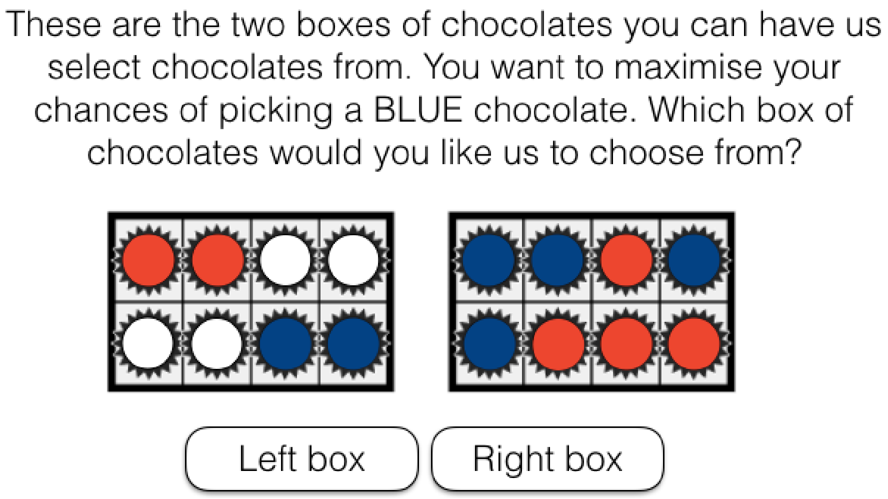
There is some evidence to support the idea that people evaluate ambiguous options pessimistically. Keren and Gerritsen (1999) asked people to predict which of two decision makers – one who chose a precise option, the other an ambiguous option – were most likely to succeed in their bets. Participants rated the decision-maker who chose the precise option as more likely to win. Viewed in Bayesian terms, this makes sense if the prior $P(\theta)$ that people use to evaluate an ambiguous option is pessimistic, and to assume that the omitted information is biased against them. Interestingly, in experimental scenarios that allow people to verify that the "ambiguous" option is constrained by a simple stochastic process that is not biased against them (i.e., ambiguity is reduced to "mere" second order probability), people do not display ambiguity aversion to the same extent (Güney & Newell, 2015).

In this paper we extend this idea, introducing a manipulation that aims to shape the PRIOR that people use to evaluate the ambiguous option by providing a causal story for why some information is missing. At the same time, we manipulate the amount of information available to people as well as the apparent favorability of the gambles (the INFORMATION CUE). There is some evidence that these two factors may interact (Garcia-Retamero et al., 2009). For instance, Tversky and Kahneman (1980) suggest that observed base rates are only taken into account when they appear to be causal (i.e. directly related to the outcome), Garcia-Retamero et al. (2009) find that people also display a confirmation bias, attending to informational cues only to the degree that they support the initial base rate. Here we manipulate each of these factors more systematically.

6.4 Method

6.4.1 Participants

79 University of Adelaide first-year psychology students and 364 Amazon Mechanical Turk (AMT) workers participated on their own computer or laptop. Of the initial participants, 73 were randomly allocated to a control condition with no ambiguity, just to make sure the instructions were clear. As expected, they showed no preference between options. They are thus are excluded from all subsequent analyses, leaving 370 in the full dataset. Of these, 167 were female, 201 were male, and 2 marked 'other' for gender; ages ranged from 17 to 73 (mean = 32.2). Student participants were given course credit, while AMT workers were paid $1 for the five minute task. Initial analysis indicated no qualitative difference in performance between the two subject pools, so all have been combined for all analyses.



These are the two boxes of chocolates you can have us select chocolates from. You want to maximise your chances of picking a BLUE chocolate. Which box of chocolates would you like us to choose from?

Figure 6.1: **The basic scenario.** Participants were asked to select a box from which the experimenter would randomly select one chocolate: blue ones were worth 100 points and red ones were worth nothing. In all conditions, people had to choose from one box in which the distribution of red (which looks medium grey in greyscale) and blue (dark grey) items was fully apparent; the other box was fully or partially ambiguous, represented by chocolates in white wrappers that could be either red or blue underneath.

6.4.2 Materials & procedure

Regardless of condition, people were told they were taking part in a study about making choices. According to the cover story, they were part of a promotion being held by a fictitious chocolate factory, which produced boxes containing eight chocolates. As a part of the promotion, the researcher would select one chocolate randomly from the box. If it was blue, the participant would be awarded 100 points, but if it was red, they would not be awarded anything. The job of each participant was to choose which of two boxes they would like the researcher to choose a chocolate from.

Importantly, as shown in Figure 6.1, participants always had to choose between one box in which the distribution of chocolates was fully known, and another in which it was fully or partially ambiguous. The side with the ambiguous box was randomized. All participants were told the same cover story explaining the ambiguity: the machine that wrapped the chocolates had malfunctioned and randomly re-wrapped some chocolates with a white wrapper. This did not impact the wrapping underneath, so if a white wrapper was selected people would be given points based on whether the wrapper underneath was blue or red.

This particular cover story was chosen for two main reasons. With the first part we could manipulate the underlying prior distribution for the ambiguous tokens (either expecting more red, more blue or a neutral condition). With the second part we introduced a reason for ambiguity that was, to the best of our powers, neutral and separate to the underlying distribution of red or blue wrappers in a given box.

The study manipulated two main factors. The first was the PRIOR expectation participants had about how many chocolates of each kind might be found in each box, which we accomplished through a cover story that relied on their social reasoning. The second was the amount of INFORMATION provided about the more ambiguous box. The experiment was fully between-subjects, with each participant randomly allocated to one of the 15 conditions (3 PRIOR x 5 INFORMATION level). Each participant was thus asked to make one choice and one estimate about the number blue chocolates among those that were hidden.

6.4.3 Manipulating the prior expectations

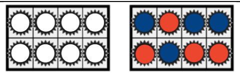
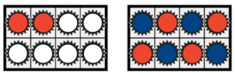
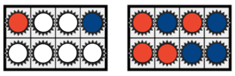
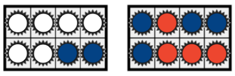
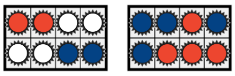
One of our main questions was how people form prior expectations of what an ambiguous situation might indicate, and how those would affect their degree of ambiguity aversion. The manipulation of prior assumptions therefore depended on a social cover story about a worker at the factory with different kinds of motivations. In the POSITIVE condition, the worker was trying to help people win points by putting more blue chocolates in the boxes; in the NEGATIVE condition they were trying to make them more likely to lose by putting more red ones in; and in the NEUTRAL condition they were rearranging chocolates with no bias one way or another. We were careful to note in all cases that the employee didn't affect all of the chocolates, so that there would be doubt about the exact nature of the distribution the ambiguous candies. The exact instructions are:

- POSITIVE: "One of our employees wants to increase the odds that people will get prize money so he randomly puts more blue chocolates in some of the boxes, though he didn't get to all of them."
- NEUTRAL: "One of our employees enjoys rearranging the order of the chocolates in the boxes, though he doesn't change which ones go where and he didn't get to all of them."
- NEGATIVE: "One of our employees who thinks the company is losing money on this promotion randomly puts more red chocolates in some of the boxes, though he didn't get to all of them."

6.4.4 Manipulating the distribution of observations

Because our primary question was how ambiguity aversion relates to the combination of prior beliefs about the distribution as well as the information they did have, we also manipulated the information provided about the ambiguous box. There were

Table 6.1: The five INFORMATION conditions, which varied the degree of ambiguity and the distribution implied by the visible pieces in the ambiguous box. The other box always contained half red and half blue pieces, all visible. In the actual experiment the side that the ambiguous box appeared on was randomized.

Condition	Stimuli
100%	
Negative 75%	
Neutral 75%	
Positive 75%	
50%	

two components to this. The first is the total amount of ambiguity: what percentage of the wrappers in the "ambiguous" box were ambiguous, i.e., white? The second was the true distribution suggested by the revealed wrappers if there were any.

This resulted in five experimental conditions, as shown in Table 6.1. Ambiguity ranged from 100% (in which all of the wrappers were white) to 50% in which half were white. To manipulate information content, we split the 75% condition into three versions. In the 75% NEGATIVE the two non-white wrappers were red, in the 75% POSITIVE they were both blue, and in the 75% NEUTRAL one was red and one was blue. These conditions are interesting because they are mostly ambiguous but provide a small amount of information about the true underlying distribution, and thus are ideal for teasing apart the roles of the amount of ambiguity and the nature of the observed information. In order to test the success of this manipulation, after making their choice we asked each participant how many of the hidden chocolates they thought were red, how many were blue, and how many might have been another color.

6.5 Results and Discussion

6.5.1 The effect of prior knowledge

Our first question was whether manipulating the prior assumptions affected people's levels of ambiguity aversion. To analyze this, we evaluated whether the proportion of people choosing the ambiguous option differed between PRIOR conditions, collapsing across the amount of ambiguity but excluding the cases where there were additional cues (i.e. the 75% negative and positive conditions).[1] There was a significant effect, which can be seen by considering the overall difference between the colored bars in Figure 6.2 ($\chi^2(2) = 25.98, p < 0.001$, $BF_{10} = 39523$).[2]

Indeed, those in the POSITIVE condition were much more likely to choose the ambiguous box. In that condition the aversion to ambiguity was eliminated ($\chi^2(1) = 14.75, p < 0.001$, Holm correction for multiple tests used, Bayesian 95% credible interval [.62,.85]) since people actually preferred the ambiguous box: an eminently sensible choice if they thought it was more likely to contain more blue chocolates. People in both the NEUTRAL ($\chi^2(1) = 4.55, p < 0.05$, Holm corrected; Bayesian 95% CI [0.29,0.49]) and NEGATIVE ($\chi^2(1) = 7.58, p < 0.05$, Holm corrected, Bayesian

[1] The code used for all analyses presented in this manuscript are available at LK's GitHub page https://github.com/lauken13/Ambiguity

[2] For this analysis we collapsed across the neutral ambiguity conditions, but excluded the conditions that gave distribution cues. Including these two conditions made no difference to the outcome.

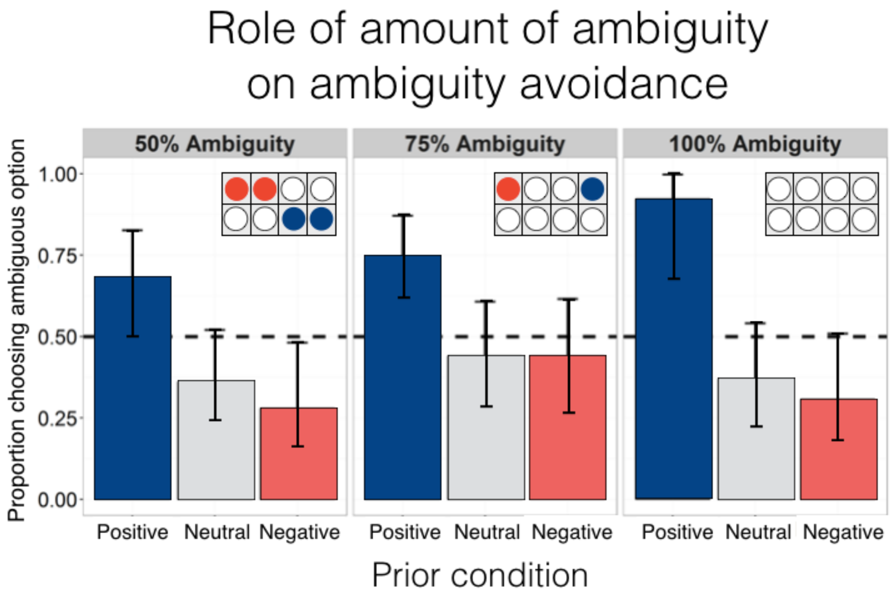


Figure 6.2: **Proportion of participants choosing the ambiguous option as a function of degree of ambiguity.** Intervals show the 95% highest density credible interval. The x axis reflects the prior manipulation affecting people's expectations about what the ambiguous box held, with people in the POSITIVE condition expecting that it held more good (blue) chocolates and the NEGATIVE condition expecting it held more red. Each of the three panels depicts different degrees of ambiguity with example stimuli shown in the insets. Those in both the NEUTRAL and NEGATIVE conditions showed the classic ambiguity aversion affect. Interestingly, it did not seem to change with the degree of ambiguity.

95% CI [.25,.45]) conditions did show the standard ambiguity aversion, preferring to avoid the ambiguous box.

Several aspects of these results are interesting. First, people in the NEUTRAL condition were just as ambiguity-averse as those in the NEGATIVE condition ($\chi^2(1) = 0.18, p = .67$, BF_{10}=0.33). Second, the degree of ambiguity aversion in both the NEUTRAL and NEGATIVE conditions also did not appear to change based on the amount of ambiguity: people were just as averse to selecting chocolates out of a box with four hidden chocolates ones as with eight hidden, despite the fact that the former situation contains more information about the true distribution of chocolates

(NEUTRAL: $\chi^2(2) = 0.43, p = .81$, $BF_{10} = 0.21$), NEGATIVE: $\chi^2(2) = 1.63, p = .44$, $BF_{10} = 0.41$). In the POSITIVE condition, there appeared to be a trend where people *did* seem to be sensitive to the amount of ambiguity: as it increased, they were more likely to choose the ambiguous box, but this was not statistically significant (Bayes Factor "more likely" vs "all different" = 1.68, "more likely" vs "all the same" = 0.64).[3]

Why might people be doing this? One possibility is that people in the POSITIVE condition believed that it contained more blue chocolates, while those in the NEUTRAL and NEGATIVE conditions believed it to be mostly red, thus inducing ambiguity aversion. To test this possibility, we analyzed participants' estimates of the number of blue chocolates in the ambiguous box. As Figure 6.3 shows, those in the POSITIVE condition inferred that more than half of the chocolates were blue ($\bar{X} = 4.72, t(56) = 5.45, p < 0.001$, $BF_{10} = 13974$), those in the NEUTRAL condition did not have strong opinions either way ($\bar{X} = 3.99, t(87) = -0.13, p = .90$, $BF_{10} = 0.12$), and those in the NEGATIVE condition felt that half or less were blue ($\bar{X} = 3.68, t(75) = -2.76, p < 0.01$, $BF_{10} = 4.26$).

These results suggest three things. First, they indicate that ambiguity aversion can be overcome if there is reason to believe that the ambiguous item is more appealing than the unambiguous item (in this case, that it contains more than 50% of blue chocolates). Second, they suggest that ambiguity aversion is not solely caused by the prior: people showed the ambiguity aversion in the NEUTRAL condition while

[3] We used Bayes Factors for this analysis because of the ease with which it allowed us to test an ordinal hypothesis (i.e., that the groups were increasing) for binomial data, which a χ-square does not.

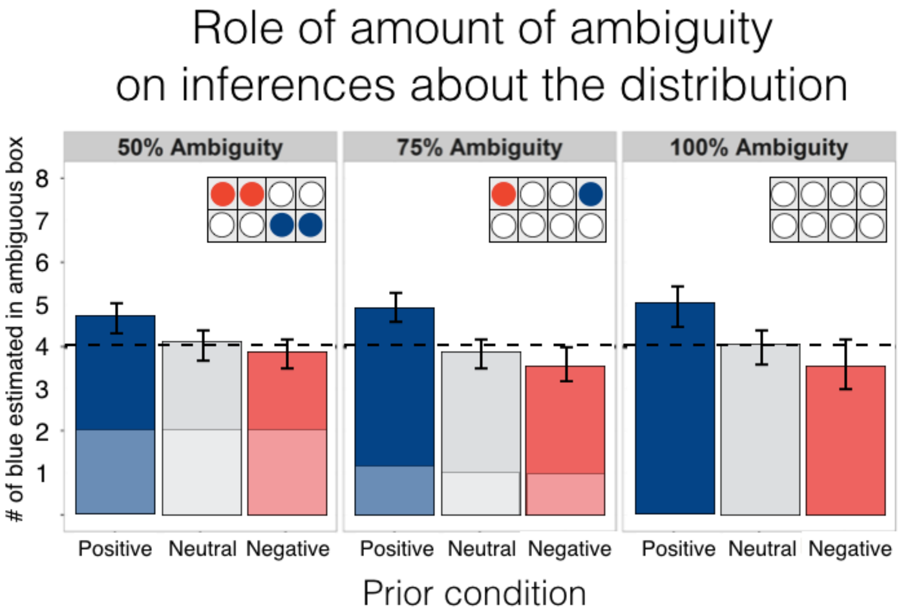


Figure 6.3: **Estimated number of blue chocolates in the ambiguous box as a function of degree of ambiguity.** The y axis shows the number of chocolates that people estimated were in the ambiguous box. Only the people in the POSITIVE condition assumed that the ambiguous box contained more than 50% blue chocolates, those in the NEUTRAL condition assumed that about half were blue, and those in the NEGATIVE condition thought half or fewer were blue. For reference, the lightly shaded boxes at the bottom of the bars indicate how many of the estimated chocolates were directly visible in that condition.

simultaneously estimating that the ambiguous box contained the same number of blue chocolates as the unambiguous one (i.e., four).

Overall, these results suggest that people's prior assumptions play a strong but not dispositive role in ambiguity aversion: when people are given a reason to believe that an ambiguous choice is better they are entirely willing to choose it. Moreover, people are sensitive to a cover story about how the observations were generated, and are willing to adjust their decision making on the basis of their beliefs about this generative process. Nevertheless, it is noteworthy that prior beliefs are not the entire story: people are ambiguity averse even when they believe the ambiguous option is statistically identical to the unambiguous one.

6.5.2 The effect of distributional information

As the previous section illustrates, people's prior beliefs about an ambiguous situation are sensitive to the causal explanation for why the missing information is missing. We now consider the question of how sensitive people are to the distribution of observed evidence. To that end we consider the three 75% ambiguous conditions, all of which are matched on ambiguity, but present people with evidence that is positive (two blue chocolates are observed), negative (two reds), or neutral (one blue and one red).

The results, shown in Figure 6.4, reveal that people were largely insensitive to the distribution of observed chocolates, at least for the POSITIVE and NEGATIVE cover stories. People given the POSITIVE cover story tended to prefer the ambiguous box regardless of what the two visible chocolates looked like ($\chi^2(2) = 0.06, p = .97$, $BF_{10} = 0.17$), while those given the NEGATIVE cover story avoided it regardless ($\chi^2(2) = 1.87, p = .39$, $BF_{10} = 0.47$). In these two conditions, people's prior beliefs about the source of the ambiguity (i.e., the cover story) overwhelmed any effect that the observed chocolates might have had. Only in the NEUTRAL condition – where no clear reason for the ambiguity was provided – did people change their behavior consistently, picking the ambiguous box more often when the two visible chocolates were blue (BF "increasing" vs "all the same" = 3.30, BF "increasing" vs "all different" = 2.62). While it seems plausible that people would be willing to adjust their behavior in the POSITIVE and NEGATIVE conditions if the quantity

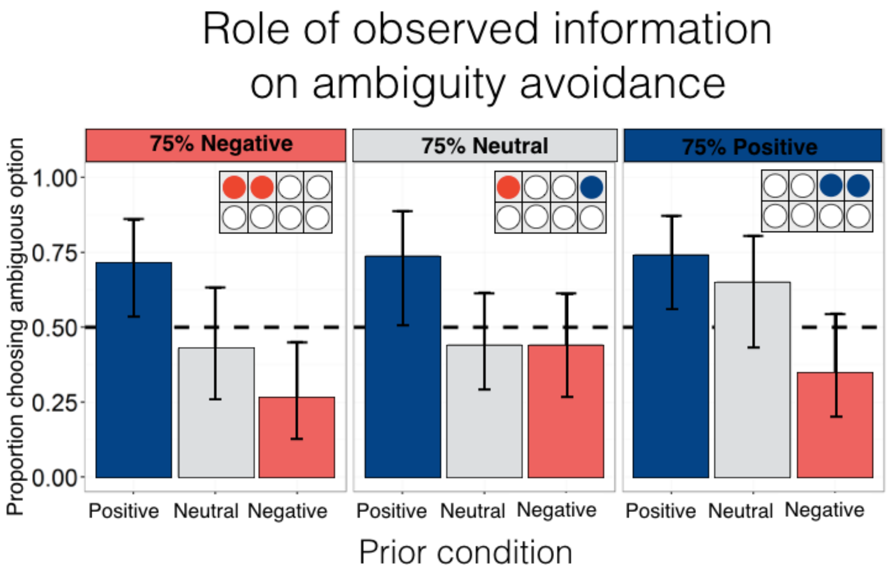


Figure 6.4: **Proportion of participants choosing the ambiguous option as a function of observed information.** Each of the three panels show boxes with 75% ambiguity (i.e., two known chocolates and six unknown) but differ in the distributional makeup of the known chocolates, as shown in the inset diagrams: in the 75% NEGATIVE condition both chocolates were red, in the 75% NEUTRAL condition there was one of each, and in the 75% POSITIVE condition both were blue. It is clear that informational content did not affect ambiguity aversion except in the 75% NEUTRAL condition, suggesting that only then were people's priors weak enough to be overcome by data.

of observations were larger (e.g., a box with 99 blue chocolates observed and 1 unknown is much better than one with 50 blue chocolates and 50 red, regardless of how maliciously the 1 ambiguous chocolate was chosen!), the fact that no observed differences were found in our data suggests that the biases imposed by the cover story are quite strong.

This explanation is supported when we consider the estimates of the number of blue chocolates in these three conditions. As Figure 6.5 makes clear, people take the observed information into account in all conditions: regardless of whether the cover story was POSITIVE ($R^2 = .20, F(2,71) = 8.71, p < 0.001,\ BF_{10} = 79.81$), NEUTRAL ($R^2 = .29, F(2,63) = 12.91, p < 0.001,\ BF_{10} = 971.65$), or NEGATIVE

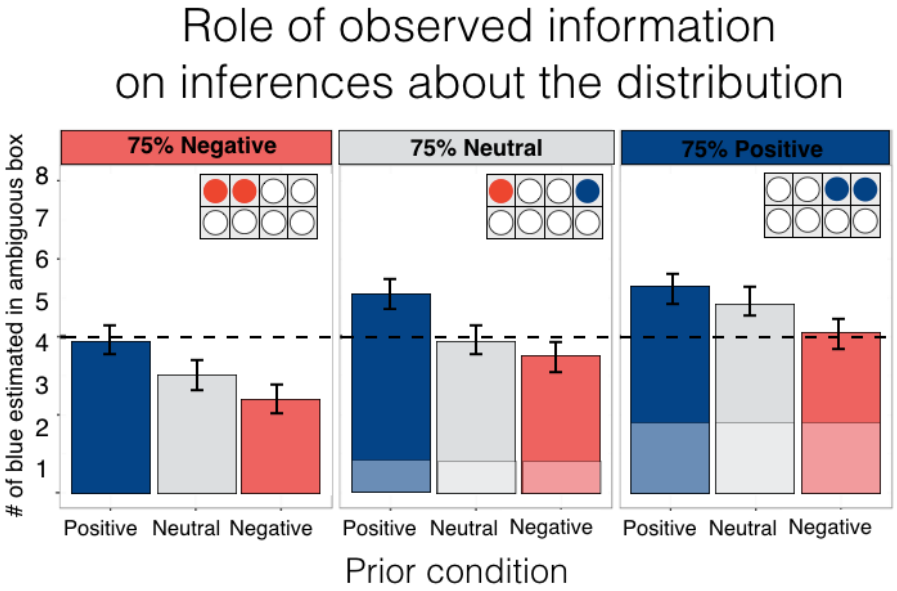


Figure 6.5: **Estimated number of blue chocolates in the ambigous box as a function of amount of information.** The y axis shows the number of chocolates that people estimated were in the ambiguous box. People were sensitive to both the prior (cover story) and data (visible chocolates). Priors played a role, with people (regardless of the pattern of visible chocolates) estimating the most blue chocolates when they were given the POSITIVE cover story and the least when given the NEGATIVE cover story. The data played a role too, with people (regardless of cover story) inferring the fewest blue chocolates when two red ones were visible, and the most when two blue ones were.

($R^2 = .39, F(2, 75) = 23.52, p < 0.001,\ BF_{10} = 997272.9$), people thought the box with two visible blue chocolates probably contained more blue chocolates, while the box with two visible red chocolates contained fewest. There is an effect of the prior cover story as well, with people who received the POSITIVE cover story consistently estimated more blue chocolates than those who received the NEUTRAL one ($t(126.82) = 2.95, p < 0.01,\ BF_{10} = 8.14$), and those more than the NEGATIVE one ($t(145.18) = 5.90, p < 0.001,\ BF_{10} = 503803$). This is consistent with the suggestion that people are sensitive to the quantity observed evidence, but consider the cover story to be the more important factor when driving decisions in situations like these where comparatively little information is available.

6.5.3 Summary

The experiment suggests that participants are very sensitive to the reason given for why the ambiguity exists: it appears to exert a substantial influence on the priors people bring to the statistical inference problem that ambiguity presents. This effect is seen regardless of whether we look at which box people chose, or whether we look at their estimates of the number of chocolates in the ambiguous box. That said, ambiguity aversion is not fully explained by the combination of priors, the observed data, and estimates about the true nature of the underlying distribution of chocolates in the ambiguous box. This is evident for several reasons. First, people tended to avoid ambiguity to a similar degree regardless of whether the cover story was NEUTRAL or NEGATIVE (Figure 6.2), and differences in the amount of ambiguity when it was POSITIVE resulted in different levels of ambiguity aversion (Figure 6.2) but not differences in the estimated number of blue chocolates (Figure 6.3). Moreover, while differences in the *distribution* of the observed chocolates affected ambiguity aversion when the cover story was NEUTRAL, they did not affect it when the cover story was POSITIVE or NEGATIVE (Figure 6.4), though this may be a consequence of strong priors. As shown in Figure 6.5, people do use the observed data to guide their estimates in these condition, but the data never outweigh the prior.

6.6 General Discussion

It is perhaps unsurprising that ambiguity aversion is malleable when people are presented with partial data or given reasons to view ambiguity in a positive light. After all, real world decision making always involves some degree of ambiguity: it would be very strange if people could not embrace it under the right circumstances. The critical finding from our experiment, however, is that people seem to assess ambiguity in a fairly rational way: when given reasons to prefer ambiguity people do so, and when given observational data that makes the ambiguous option more favorable, people adjust their beliefs about the desirability of that option. This is consistent with findings by Güney and Newell (2011, 2015) that ambiguity aversion is related to the prior beliefs that people use to interpret the ambiguity, and is reminiscent of results showing that other decision-making effects can be manipulated by shifting the quality and quantity of information available to people (e.g., Welsh & Navarro, 2012).

Of particular interest to us is the fact that ambiguity aversion seems to be the default behavior in our experiment: despite our best efforts to create a "neutral" prior condition, participants were disposed to treat that scenario with some suspicion, and adopt a relatively negative strategy that more closely resembles the NEGATIVE cover story than the POSITIVE one. Moreover, they did so in spite of the fact that their estimates of the proportion of blue chocolates were roughly 50% on average. Does this suggest that our participants take additional information into account when making decisions about ambiguity? If so, what other cues could they be using?

One possibility is the social aspect of ambiguity aversion that has been demonstrated by several researchers. Charness, Karni, and Levin (2013) demonstrate that allowing individuals to control aspects of the experiment – reducing suspicion about the ambiguous option – reduces ambiguity aversion, whist allowing participant collusion results in a trend towards ambiguity neutrality. This effect of collusion was also demonstrated by Keck, Diecidue, and Budescu (2012), suggesting that a fruitful avenue for future work would be to examine the effect of social co-operation and competition on ambiguity preferences.

Another possibility is that participants are relying on a different decision process, one that doesn't aim to maximize expected utility. For instance, Gilboa and Schmeidler (1989) propose the Maxmin Expected Utility theory, which suggests that participants have in mind a prior range for the number of desired tokens in the ambiguous option. They will choose this ambiguous option if and only if the expected utility of the ambiguous option for the minimum of the range is less than the expected utility of the unambiguous option. If people are using this rule to make decisions, but still give estimates about the expected value of blue chocolates, this theory might explain the dissonance between decisions and inferred distribution. Future work could elicit this directly by asking participants to give an upper and lower bound for the number of blue chocolates that they expect, but would also need to explain why this is true for some but not all conditions.

More generally, a natural direction to extend this work is to consider a broader range of scenarios and mechanisms that might give rise to ambiguous data. In real life, this can happen in many different ways. The interpretation of surveys with missing data can be very different depending on whether the data are missing at random or have been systematically censored. Ambiguous phrasing in an opinion piece might be innocuous, or it might reflect an attempt to mislead with half truths. Even the ambiguity in something as simple as Ellsberg's (1961) original urn task might carry a different interpretation depending on whether participants believe the experimenter is asking a trick question. In short, when trying to construct a general theory of why people tend towards ambiguity aversion, we suggest it makes sense to ask where ambiguity arises in the real world. Is ambiguity usually associated with positive outcomes, or does the absence of key information tend to suggest something more nefarious? Our current results are ambiguous, but future work will investigate additional informative cues.

6.7 References

Camerer, C. & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, *5*(4), 325–370.

Charness, G., Karni, E., & Levin, D. (2013). Ambiguity attitudes and social interactions: An experimental investigation. *Journal of Risk and Uncertainty*, *46*(1), 1–25.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, *75*(4), 643–669.

Garcia-Retamero, R., Müller, S. M., Catena, A., & Maldonado, A. (2009). The power of causal beliefs and conflicting evidence on causal judgments and decision making. *Learning and Motivation*, *40*(3), 284–297.

Gardenfors, P. & Sahlin, N. E. (1982). Unreliable probabilities, risk taking, and decison making. *Synthese*, *53*, 361–386.

Gilboa, I. & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, *18*(2), 141–153.

Güney, Ş. & Newell, B. R. (2011). The ellsberg problem and implicit assumptions under ambiguity. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 2323–2328). Cognitive Science Society Austin, TX.

Güney, Ş. & Newell, B. R. (2015). Overcoming ambiguity aversion through experience. *Journal of Behavioral Decision Making*, *28*(2), 188–199.

Keck, S., Diecidue, E., & Budescu, D. (2012). Group decisions under ambiguity: Convergence to neturality. *Working paper.*

Keren, G. & Gerritsen, L. E. (1999). On the robustness and possible accounts of ambiguity aversion. *Acta Psychologica*, *103*(1), 149–172.

Liu, H.-H. & Colman, A. M. (2009). Ambiguity aversion in the long run: Repeated decisions under risk and uncertainty. *Journal of Economic Psychology*, *30*(3), 277–284.

Tversky, A. & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. *Progress in Social Psychology*, *1*, 49–72.

Welsh, M. & Navarro, D. J. (2012). Seeing is believing: Priors, trust, and base rate neglect. *Organizational Behavior and Human Decision Processes*, *119*, 1–14.

Chapter 7

Discussion

"This isn't magic, is it?"

"I don't think so," said Johhny.

"It's probably just very, very, very strange science."

"Oh, good," said Yo-Less. "Er... What's the difference?"

Terry Pratchett

This thesis investigates four areas where researcher and statistical assumptions conflict. In discussing each area, I argued that we *can* and *should* create statistical models that incorporate researcher assumptions. To conclude I overview the findings of this thesis, identify some caveats and future directions and then make my final conclusions.

7.1 Overview

Chapter three considered the nature of the data obtained using clinically aimed scales with a non-clinical population. This combination creates a strong floor effect and produces skewed data. I proposed ranked alternatives to increase the accuracy of conclusions made from this data, and a cdf quantile method to increase the richness of possible conclusions (Smithson & Shou, 2017). Overall, this chapter demonstrates that while *not* violating the assumptions of the statistical method results in more accurate results, increased complexity with some underlying theory leads to richer claims.

Chapter four extended this by investigating the most appropriate model when the researcher has some belief that their sample might be contaminated. In this chapter I proposed a Bayesian mixture model that corresponds to a process of contamination. This model was as or more accurate than the current best practice while also allowing for additional claims to be made about the data. The chapter provides evidence that incorporating researcher assumptions improves accuracy and changes the claims that can be made from the data.

The next chapter built upon this. Here we considered a situation where the researcher hypothesised that not all individuals will respond in the same way. I utilised a family of models that incorporates these assumptions. These models not only allowed the researcher to make the claims about individual change, but also about group level differences of proportion and effect size. Other models that do not incorporate these assumptions do not have this property. This model, by incorporating researcher assumptions, allowed the researcher to make inference about the parameters they are *truly interested* in.

Lastly, chapter six took an alternate approach by experimentally investigating the relationship of ambiguity aversion to underlying distribution assumptions. We show that despite replicating an aversion to ambiguity, participants were sensitive to different conditions with different underlying information cues. I include this chapter as an example that despite a robust finding of rejection of ambiguity, people are still able to make the *sensible assumptions* about the underlying data. This supports other chapters in this thesis that suggest researchers should incorporate their assumptions about data into their statistical models.

With these four main findings laid out, in the next sections we will go on to consider four different considerations raised by the findings of this thesis.

7.2 Researcher Assumptions

I began this thesis by arguing that statistical methods rely on the researcher's *assumptions*, *intentions* and *interpretation.* To demonstrate this, I outlined a number of different examples where these three factors changed the chosen method and the interpretation of its findings. However, the bulk of this thesis explored four examples where these assumptions do not currently match the model most often used.

I am not the first to propose that researcher knowledge could be built into Bayesian models for data analysis. Hemmer, Tauber, and Steyvers (2015) propose incorporating researcher driven priors into the analysis of Bayesian models of cognition. While their work considers models of cognition, ours extends this concept to discuss the inclusion of models of the *data generation* process in general.

I argue through simulations studies, modelling and a study of decision making, that the researcher's *assumptions*, *intent* and *interpretation* should be incorporated into the statistical modelling process. To conclude I turn to discuss the strengths and weaknesses of my simulation-based approach, problems associated with model complexity and the impact of cognitive biases on researcher assumptions.

7.3 Simulation Studies

Chapters three through five used a series of simulation studies to evaluate relative model efficacy. Simulation studies were chosen as they allowed us to compare and contrast different models given data sampled with different assumptions. They afforded control that elucidated three benefits for the work presented in this thesis. Firstly, it allowed us to explore the efficacy of the model in question with different data scenarios. Secondly, because the data were simulated, we could investigate where parameter recovery was possible and where it was not. This allows us to gain an estimate of the accuracy of the model. Lastly, using simulated data allows us to compare different models over identical data types. Any difference between the efficacy of the models is due to differences in the model and not due to differences in the sample.

Simulation studies are to real world datasets as basic science is to applied science. Both are important, but while simulation studies and basic science (Calder, Phillips, & Tybout, 1982) allow for control within the study, real world data and applied science (Steckler & McLeroy, 2008) increases the external validity of the method. A lack of real-world data is one of the largest limitations of this thesis. However, comparing models with purely real-world data does not disambiguate which model makes the correct claims.

Future work could compare and contrast the claims made by the different models with real datasets (in a method similar to Hinton-Bayre, 2012). This type of study would be beneficial as it indicates how often the choice of model changes the conclusions made in real-world data analysis. This style of investigation alone does not demonstrate which model is the correct model, or which conclusion is most trustworthy, which is the benefit of pure simulation studies. An alternative would be to conduct pseudo-simulations: calculating the effect of sample size using subsamples of a real world dataset.

The procedure for this is as follows. Given a large dataset of two variables $\{Y, X\}$, and a model that estimates a desired parameter θ.

1. For sample size n_1, draw n_1 random cases from dataset $\{Y, X\}$, which we denote $\{Y_{n1,1}, X_{n1,1}\}$.

2. Using this sub sample, fit the model and obtain an estimate for θ, $\hat{\theta}_{n1,1}$, and compare this to $\hat{\theta}_{pop}$ as estimated with the full dataset. Calculate an interval around $\hat{\theta}_{n1,1}$ that reflects the uncertainty of this estimate.

3. Repeat steps one and two for j samples of size n_1 to obtain an estimate of accuracy and precision for every respective $\hat{\theta}_{n1,j}$. Calculate an average expected accuracy and precision for that sample size.

4. Repeat steps one through three for sample sizes in the set of interest $N = \{n_1, n_2, ..., n_B\}$, constrained so that $n_B \leqslant N$.

One last critique with simulation studies is that it is possible to create a dataset that perfectly matches the model assumptions. In such a case, it would be hardly surprisingly that this model outperforms any other models with different assumptions, as they are miss-specified for the data. This issue is particularly important in chapters four and five. In chapter four we addressed it by being careful to use datasets that violate the assumptions of *all* the models. In chapter five, this is less of an issue as the message of this chapter is that it is possible to create a model that incorporates assumptions underlying repeated-measures data.

7.4 Complex models and overfitting

Many of the assumptions I have identified that are very common in psychological research were introduced to simply reduce the number of parameters and the overall complexity of the statistical model. The importance of using the simplest model possible has long been noted (e.g., minimum description length (Rissanen, 1978) and minimum message length (Wallace & Boulton, 1968)). Minimising the number of parameters allowed to freely move prevents the model from overfitting to the observed data at the expense of accuracy in predicting future data. In this section I discuss a) why minimal model methods sometimes are insufficient for the data they fit, b) how more complex models are useful in many cases, but must be used with caution, and c) the methods to differentiate between the two.

7.4.1 Defining model simplicity

The search for a model that best describes the data at hand is often required to conform to Occam's razor, the simplest model that explains the data ought to be preferred. Model simplicity can be thought of two different ways.

The first way is to consider the number of free parameters. Free parameters are variables contained within the model that are allowed to freely vary to best fit to the data. Models with fewer free parameters are generally considered to be simpler.

The second way to consider the complexity of the model is to consider the shape of possible relationships. Given a model where some outcome variable Y is predicted by X, would a model with a quadratic term (e.g., $Y \sim aX^2 + c$), a model without a quadratic term (e.g., $Y \sim aX + c$) or a model with a power term (e.g., $Y \sim X^a + c$) be a more complicated model? All three have just two free parameters, a and c, but differ greatly in the overall shape they can describe.

If model complexity is too ambiguous, then by what metric should we choose one model over the other? If an analytic solution is desired, it will be faster but practical considerations mean that some distributions (notably the normal distribution) will be over-emphasised. However, analytic tractability is not usually a problem nowadays thanks to the presence of computational sampling methods.

The models proposed in this thesis are not simply more complex alternatives to the standard practice. They are also not chosen to be the most complicated description of the data possible. Instead the models, especially those chosen in chapters four and five use additional complexity to allow the researcher to make claims about aspects of the data they are truly interested in. This is accomplished by incorporating their beliefs about the underlying data generation process. Of course, it can be difficult to define model complexity. When it is difficult, the choice between two models depends on the context and beliefs of the researcher.

7.4.2 Comparing models

Complexity is often seen as an undesirable trait when modelling as it increases the probability of *overfitting*. A model that is more complex can describe a wider variety of data and relationships than a less complex model. This might seem desirable, but if complex models are used unnecessarily, they might account for small random fluctuations in the data as part of the true relationship. If this occurs, predictions made by the model may be less accurate than simpler models.

With decreased prediction accuracy as a potential side effect of complexity, when is it justified to use a more complex model instead of a simpler one? If the more complicated model is closer to “true” underlying model than the simpler model, then it will both fit the observed data *and* predict future data as well as or better than the simpler model. There are a number of different techniques to identify the model that best describes the data whilst penalising for model complexity. In a review, Myung and Pitt (2016) categorise these methods into penalised likelihood methods, Bayesian methods and direct estimation methods.

Penalised likelihood models, such as the AIC (Akaike, 1981) and the BIC (Schwarz et al., 1978) seek to compare models on the likelihood of the model given the data whilst penalising model complexity by a function of the number of free parameters and sample size. While this method does account for complexity, it only accounts for the complexity that can be attributed to the number of free parameters the model has, not the flexibility of relationships possible.

The Bayesian alternative takes one of two approaches. The first is a Bayes Factor approach, or more generally a Bayesian Model Selection (BMS) approach (Myung & Pitt, 2016). The BMS is the negative log of the marginal likelihood of the model given the data. The difference between the BMS for two models is the log Bayes Factor (Kass & Raftery, 1995), where the log Bayes Factor is the ratio of the marginal likelihood of model 1 over the marginal likelihood of model 2. Because the BMS is a measurement across the parameter space, it accounts for model complexity without the need for additional parameters. When the BMS is difficult to calculate, an alternative method is to use either a DIC (Spiegelhalter et al., 2002), which uses the marginal likelihood estimated at the posterior mean, or the WAIC (Watanabe, 2010), which uses predictive density. As the WAIC and DIC doesn't average over the whole posterior, they still need to penalise for the number of parameters to adjust for complexity.

Lastly there are what Myung and Pitt (2016) call direct estimation methods. These are methods that use part of the sample to fit the model, and then test how well the model can predict the remaining sample. The main differences between these methods is how the sample is split, spanning from cross-validation (Geisser, 1993), that splits the sample into equal halves, to leave one out (LOO) (Vehtari, Gelman, & Gabry, 2017), which predicts each datum in turn from the rest of the sample.

I have argued that statistical models should be based on the assumptions of the researcher, suggesting that statistical models are the outcome of cognitive processes. While the aforementioned model comparison techniques consider prediction, estimation, variance accounted for and complexity in a statistical sense, they *don't* account for the aims and beliefs of the researcher. In cognitive science it has been argued that the choice of priors needs to reflect the beliefs and knowledge of the researcher (Vanpaemel, 2010), especially in terms of the underlying beliefs about the process represented by the data (Lee & Wagenmakers, 2014). Even if the model fits exceptionally well, if the researcher wants to make claims about specific aspects of the process and the model that does not contain parameters that represent this, it is not useful.

Cognitive modelling has an additional set of considerations by which models ought to be evaluated (see Myung & Pitt, 2016, for a full discussion). These recommendations might potentially be of use to statistical models that incorporate researcher beliefs. Of specific interest here are the model's plausibility and interpretability. Namely, model assumptions and structure should be based on established theory and interpretable in the context of the problem.

7.5 Cognitive biases and assumptions

Throughout this thesis I have argued that researcher's *assumptions*, *intent* and *interpretations* have and should be included in the statistical modelling process. In chapters four and five, I demonstrated that this approach can be beneficial by increasing the overall model accuracy and increasing the richness of the claims that can be made. However, can the researcher be trusted to make the correct inference about the data generation process? Decades of research in the cognitive sciences have detailed a number of cognitive biases that suggest that people are not always rational at making judgements with numbers (e.g., Ellsberg, 1961; Epley & Gilovich, 2006; Landy et al., 2013).

The work in chapter six bears on this question. We investigated whether ambiguity bias reflected a distortion in the underlying belief of missing data. We discovered that despite showing an overall bias to avoid ambiguity, when asked about the hidden tokens, participants were sensitive to different environmental cues.

This is not the first work to suggest that cognitive biases are not necessarily evidence of irrational thinking. Others have suggested that Bayesian reasoning accounts of human behaviour make good predictions, even when rational accounts do not. One example of this is evidence that suggests thinking previously identified as irrational can be better explained as evidence of a heuristic (Gigerenzer, 1991; Hahn & Oaksford, 2007).

An interesting direction for future work could involve continuing to investigate the impact on human biases on inference made in statistical analyses. Is it possible that confirmation bias (Nickerson, 1998) and the desire for researchers to find statistical evidence for their hypotheses[1] result from the same underlying cognitive mechanism? If it is, would research that aims to reduce confirmation bias (e.g., Schwind & Buder, 2012) provide clues to reduce some of the issues identified by Simmons et al. (2011)?

The work in this thesis analysed both accuracy of a binary conclusion (i.e., were the two groups different or not) and also uncertainty about these estimates. Ap-

[1] One concern with the researcher choosing statistical techniques after seeing the data (Simmons et al., 2011)

proaches that emphasise the uncertainty of measurement are strongly advocated for throughout Bayesian and frequentist literature (e.g., Cumming, 2008; Kruschke, 2014; Wagenmakers, 2007). In practice however, many scientists are reluctant to use these methods. This has led to solutions that simply decrease the probability of making a Type 1 error (Benjamin et al., 2018). Why, in practice, is there relatively little focus on estimation rather than binary assignment? Is it that the estimation makes obvious the inherent uncertainty of data analysis, and uncertainty is not preferred? This is an intriguing open question inspired by the findings of chapter six.

7.6 Final Conclusions

In this thesis I argued that researcher assumptions, intent and interpretation form an integral part of statistical analysis. In chapter one I argued that this was not a novel concept, but one that has underpinned the design and assumptions of our most common statistical methods. In chapter two I outlined methods that I had chosen to use throughout this thesis, where I begin my argument that the 'best' model is one that is both accurate and precise. Chapter three demonstrated that significant improvements can be made simply by using a model that is more appropriate for the data at hand, while chapter four demonstrated that employing a model that relies on researcher assumptions resulted in additional benefits. In chapter five I demonstrate that not only do models that incorporate researcher assumptions have significant improvements in accuracy, they also allow the researcher the ability to make claims about the parameters that are truly of interest. In chapter six I consider whether the researcher can be trusted to be sensitive to different environments that produce data. I demonstrate for a simple case, ambiguity aversion, that they are.

In a broader sense this thesis argues against the idea that researchers are inherently biased, and statistics should be undertaken by an impartial observer. Instead I argue that the researcher, an expert in their field, has a large amount of domain-specific knowledge about what they expect the process of obtaining the data entails. I argue that by incorporating this into the model itself, it is possible to make these currently implicit assumptions explicit. This increases transparency, allowing the method to be debated explicitly in the field. The ultimate goal is to lead to models built with better assumptions and to give the researcher the tools to make cleaner and more interesting descriptions of their data. In the end, transparency, reproducibility, and scholarly debate is where the *real* magic of science is.

Chapter 8

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, *16*(1), 3–14.

Andrade, J. A. A. & O'Hagan, A. (2011). Bayesian robustness modelling of location and scale parameters. *Scandinavian Journal of Statistics*, *38*(4), 691–711.

Bååth, R. (2016). *bayesboot: An Implementation of Rubin's (1981) Bayesian Bootstrap*. Retrieved from https://github.com/rasmusab/bayesboot

Bamnett, V. & Lewis, T. (1994). *Outliers in statistical data*. West Sussex, England: John Wiley & Sons.

Bartlema, A., Lee, M. D., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150.

Bayarri, M. J. & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, 58–80.

Beck, A. T., Rial, W. Y., & Rickels, K. (1974). Short form of depression inventory: Cross-validation. *Psychological Reports*, *34*, 1184–1186.

Beck, A. T., Steer, R. A., Brown, G. K., et al. (1996). Beck Depression Inventory-II. *San Antonio, TX: Psychological Corporation*, 78204–249.

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*(1), 77–100.

Beekman, A., Copeland, J., & Prince, M. J. (1999). Review of community prevalence of depression in later life. *The British Journal of Psychiatry*, *174*(4), 307–311.

Benjamin, D. J., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Camerer, C., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10.

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., . . . Sivaganesan, S. (1994). An overview of robust Bayesian analysis. *Test*, *3*(1), 5–124.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1982). The concept of external validity. *Journal of Consumer Research*, *9*(3), 240–244.

Camerer, C. & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, *5*(4), 325–370.

Casella, G. & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167–174.

Charness, G., Karni, E., & Levin, D. (2013). Ambiguity attitudes and social interactions: An experimental investigation. *Journal of Risk and Uncertainty*, *46*(1), 1–25.

Chib, S. & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, *49*(4), 327–335.

Christensen, L. & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy*, *17*(3), 305–308.

Coro, G. (2013). A lightweight guide on Gibbs sampling and JAGS. *ISTI-CNR, Technical Report.*

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104.

Crawford, J. R. & Howell, D. C. (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of Clinical and Experimental Neuropsychology*, *20*(5), 755–762.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286–300.

Cundill, B. & Alexander, N. D. (2015). Sample size calculations for skewed distributions. *BMC Medical Research Methodology*, *15*(1), 1–9.

Danileiko, I. & Lee, M. D. (2017). A model-based approach to the wisdom of the crowd in category learning. *Cognitive Science*, *Accepted.*

Davies, L. & Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association*, *88*(423), 782–792.

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language*, *59*(3), 361–376.

Diaconis, P. & Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, *14*, 1–26.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*(1), 204–223.

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test–retest reliability and practice effects of expanded Halstead–Reitan neuropsychological test battery. *Journal of the International Neuropsychological Society*, *5*(4), 346–356.

Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, *40*(1), 1–5.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185.

Efron, B. (1992). Bootstrap methods: Another look at the jackknife. In K. S. & J. N.L. (Eds.), *Breakthroughs in statistics. Springer Series in Statistics (Perspectives in Statistics)* (pp. 569–593). New York, NY: Springer New York.

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *The Annals of Applied Statistics*, *6*(4), 1971–1997.

Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap.* CRC press.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, *75*(4), 643–669.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341.

Epley, N. & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, *17*(4), 311–318.

Fang, D. Z., Young, C. B., Golshan, S., Moutier, C., & Zisook, S. (2012). Burnout in premedical undergraduate students. *Academic Psychiatry*, *36*(1), 11–16.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*(2), 209–230.

Ferrer, R. & Pardo, A. (2014). Clinically meaningful change: False positives in the estimation of individual change. *Psychological Assessment*, *26*(2), 370.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*(5), 700–725.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

Galton, F. (1894). *Natural inheritance.* New York, NY: Macmillan and Co.

Garcia-Retamero, R., Müller, S. M., Catena, A., & Maldonado, A. (2009). The power of causal beliefs and conflicting evidence on causal judgments and decision making. *Learning and Motivation*, *40*(3), 284–297.

Gardenfors, P. & Sahlin, N. E. (1982). Unreliable probabilities, risk taking, and decison making. *Synthese*, *53*, 361–386.

Geisser, S. (1993). *Predictive inference.* New York, NY: Chapman and Hall.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, *85*(412), 972–985.

Gelfand, A. E. & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Gelman, A. & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University.*

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(6), 721–741.

Gershman, S. J. & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12.

Ghassemzadeh, H., Mojtabai, R., Karamghadiri, N., & Ebrahimkhani, N. (2005). Psychometric properties of a Persian-language version of the Beck Depression

Inventory-Second edition: BDI-II-PERSIAN. *Depression and Anxiety*, *21*(4), 185–192.

Ghosh, J. & Ramamoorthi, R. (2003). *Bayesian nonparametrics*. New York, NY: Springer Series in Statistics.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology*, *2*(1), 83–115.

Gilboa, I. & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, *18*(2), 141–153.

Glass, D. C., McKnight, J. D., & Valdimarsdottir, H. (1993). Depression, burnout, and perceptions of control in hospital nurses. *Journal of Consulting and Clinical Psychology*, *61*(1), 147–155.

Godfrey, L. (2006). Tests for regression models with heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, *50*(10), 2715–2733.

Goodman, N. D. & Stuhlmüller, A. (2014). The Design and Implementation of Probabilistic Programming Languages. http://dippl.org. Accessed: 2017-8-20.

Gulliksen, H. (2013). *Theory of mental tests*. New York, NY: Routledge.

Güney, Ş. & Newell, B. R. (2011). The ellsberg problem and implicit assumptions under ambiguity. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 2323–2328). Cognitive Science Society Austin, TX.

Güney, Ş. & Newell, B. R. (2015). Overcoming ambiguity aversion through experience. *Journal of Behavioral Decision Making*, *28*(2), 188–199.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282.

Hageman, W. & Arrindell, W. A. (1999). Establishing clinically significant change: Increment of precision and the distinction between individual and group level of analysis. *Behaviour Research and Therapy*, *37*(12), 1169–93.

Hahn, U. & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704.

Hall, P. (1988). On symmetric bootstrap confidence intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, *50*(1), 35–45.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Healy, M. & Goldstein, H. (1978). Regression to the mean. *Annals of Human Biology*, *5*(3), 277–280.

Hemmer, P., Tauber, S., & Steyvers, M. (2015). Moving beyond qualitative evaluations of Bayesian models of cognition. *Psychonomic Bulletin & Review*, *22*(3), 614–628.

Hewitt, P. L. & Flett, G. L. (1993). Dimensions of perfectionism, daily stress, and depression: A test of the specific vulnerability hypothesis. *Journal of Abnormal Psychology*, *102*(1), 58–65.

Hinton-Bayre, A. D. (2012). Choice of reliable change model can alter decisions regarding neuropsychological impairment after sports-related concussion. *Clinical Journal of Sport Medicine*, *22*(2), 105–108.

Hinton-Bayre, A. D., Geffen, G. M., Geffen, L. B., McFarland, K. A., & Frijs, P. (1999). Concussion in contact sports: Reliable change indices of impairment and recovery. *Journal of Clinical and Experimental Neuropsychology*, *21*(1), 70–86.

Hoaglin, D. C., Mosteller, F., & Turkey, J. (1985). *Exploring data tables, trends, and shapes*. Hoboken, NJ: John Wiley & Sons.

Holsclaw, T., Hallgren, K. A., Steyvers, M., Smyth, P., & Atkins, D. C. (2015). Measurement error and outcome distributions: Methodological issues in regression analyses of behavioral coding data. *Psychology of Addictive Behaviors*, *29*(4), 1031.

Hornik, K., Leisch, F., & Zeileis, A. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of DSC* (Vol. 2, pp. 1–1).

Hsu, L. M. (1996). On the identification of clinically significant client changes: Reinterpretation of Jacobson's cut scores. *Journal of Psychopathology and Behavioral Assessment*, *18*(4), 371–385.

Hubbard, R. & Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (alpha's) in classical statistical testing. *The American Statistician*, *57*(3), 171–178.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101.

Huber, P. J. & Ronchetti, E. M. (2009). *Robust statistics*. John Wiley & Sons Inc.

Hubert, M. & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, *52*(12), 5186–5201.

Humphreys, M., Sanchez de la Sierra, R., & Van der Windt, P. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, *21*(1), 1–20.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, *50*(2), 120–126.

Jacobson, N. S., Dobson, K., Fruzzetti, A. E., Schmaling, K. B., & Salusky, S. (1991). Marital therapy as a treatment for depression. *Journal of Consulting and Clinical Psychology*, *59*(4), 547.

Jacobson, N. S. & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19.

JASP Team. (2016). Jasp (version 0.7.5.5)[computer software].

Jordan, M. I. (1998). *Learning in graphical models.* Norwell, MA: Springer Science & Business Media.

Karabatsos, G. (2017). A menu-driven software package of Bayesian nonparametric (and parametric) mixed models for regression analysis and density estimation. *Behavior Research Methods*, *49*(1), 335–362.

Kashani, J. H., Carlson, G. A., Beck, N. C., Hoeper, E. W., Corcoran, C. M., McAllister, J. A., ... Reid, J. C. (1987). Depression, depressive symptoms, and depressed mood among a community sample of adolescents. *American Journal of Psychiatry*, *144*(7), 931–934.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Keck, S., Diecidue, E., & Budescu, D. (2012). Group decisions under ambiguity: Convergence to neturality. *Working paper.*

Kennedy, L. A., Navarro, D. J., Perfors, A., & Briggs, N. (2017). Not every credible interval is credible: Evaluating robustness in the presence of contamination in Bayesian data analysis. *Behavior Research Methods*, *49*(6), 2219–2234.

Keren, G. & Gerritsen, L. E. (1999). On the robustness and possible accounts of ambiguity aversion. *Acta Psychologica*, *103*(1), 149–172.

Kloke, J. D. & McKean, J. W. (2012). Rfit: Rank-based estimation for linear models. *The R Journal*, *4*(2), 57–64.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658–676.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General*, *142*(2), 573–603.

Kruschke, J. K. (2014). *Doing Bayesian data analysis A tutorial with R, JAGS, and Stan*. San Diego, CA: Academic Press.

Kruschke, J. K. & Meredith, M. (2015). *BEST: Bayesian estimation supersedes the t-test*. R package version 0.4.0. Retrieved from https://CRAN.R-project.org/package=BEST

Landy, D., Silbert, N., & Goldin, A. (2013). Estimating large numbers. *Cognitive Science*, *37*(5), 775–799.

Lee, M. D. (2016). Bayesian outcome-based strategy classification. *Behavior Research Methods*, *48*(1), 29–41.

Lee, M. D. & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. New York: Cambridge University Press.

Little, A. (2009). Treatment-resistant depression. *American Family Physician*, *80*(2), 167–72.

Liu, H.-H. & Colman, A. M. (2009). Ambiguity aversion in the long run: Repeated decisions under risk and uncertainty. *Journal of Economic Psychology*, *30*(3), 277–284.

Lovibond, P. F. & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335–343.

Lovibond, S. H. & Lovibond, P. F. (1993). *Manual for the Depression Anxiety Stress Scales (DASS)*. Sydney, NSW: Psychology Foundation of Australia.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York, NY: Springer Science & Business Media.

Mackenzie, S., Wiegel, J. R., Mundt, M., Brown, D., Saewyc, E., Heiligenstein, E., . . . Fleming, M. (2011). Depression and suicide ideation among students accessing campus health care. *American Journal of Orthopsychiatry*, *81*(1), 101–107.

Martinovich, Z., Saunders, S., & Howard, K. (1996). Some comments on “Assessing clinical significance”. *Psychotherapy Research*, *6*(2), 124–132.

Mauldin, R. D., Sudderth, W. D., & Williams, S. (1992). Polya trees and random distributions. *The Annals of Statistics*, *20*(3), 1203–1221.

McCarney, R., Warner, J., Iliffe, S., Van Haselen, R., Griffin, M., & Fisher, P. (2007). The Hawthorne effect: A randomised, controlled trial. *BMC Medical Research Methodology*, *7*(1), 30–38.

McClelland, D. (1980). Motive dispositions: The merits of operant and respondent measures. *Review of Personality and Social Psychology*, *1*, 10–41.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28–50.

McDonald, R. P. (2011). *Test theory: A unified treatment.* New York, NY: Taylor & Francis.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Morey, R. D. & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes factors for common designs.* R package version 0.9.11-1. Retrieved from http://CRAN.R-project.org/package=BayesFactor

Myung, J. I. & Pitt, M. A. (2016). Model comparison in psychology. In J. Wixted & E.-J. Wagenmakers (Eds.), *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (Fourth Edition), Volume 5: Methodology.* New York, NY: John Wiley & Sons.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*(2), 101–122.

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones, & X. L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*. NW: Chapman & Hall.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *236*(767), 333–380.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Novick, M. R. (1965). The axioms and principal results of classical test theory. *ETS Research Report Series*, *1965*(1).

Novick, M. R. & Lewis, C. (1966). Coefficient alpha and the reliability of composite measurements. *ETS Research Report Series*, *1966*(1).

Olsson, G. & Knorring, A.-L. (1999). Adolescent depression: Prevalence in Swedish high-school students. *Acta Psychiatrica Scandinavica*, *99*(5), 324–331.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Payne, R. & Jones, H. G. (1957). Statistics for the investigation of individual cases. *Journal of Clinical Psychology*, *13*(2), 115–121.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*, 240–242.

Pearson, K. (1920). The fundamental problem of practical statistics. *Biometrika*, *13*(1), 1–16.

Pitman, J. & Yor, M. (1981). Bessel processes and infinitely divisible laws. In W. D. (Ed.), *Stochastic integrals. lecture notes in mathematics* (Vol. 851, pp. 285–370). Berlin, Heidelberg: Springer.

Plummer, M. (2015). JAGS Version 4.0.0 user manual. *International Agency for Research on Cancer, Lyon, France.*

Plummer, M. (2016). *Rjags: Bayesian graphical models using MCMC.* R package version 4-6. Retrieved from https://CRAN.R-project.org/package=rjags

R Core Team. (2015). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401.

Radloff, L. S. (1991). The use of the Center for Epidemiologic Studies Depression Scale in adolescents and young adults. *Journal of Youth and Adolescence*, *20*(2), 149–166.

Rahman, M. M. & Govindarajulu, Z. (1997). A modification of the test of Shapiro and Wilk for normality. *Journal of Applied Statistics*, *24*(2), 219–236.

Ratcliff, R. & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481.

Reiss, J. & Sprenger, J. (2017). Scientific objectivity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University.

Revuelta, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(4), 311–327.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Ross, J. S., Fletcher, J. A., Bloom, K. J., Linette, G. P., Stec, J., Symmans, W. F., ... Hortobagyi, G. N. (2004). Targeted therapy in breast cancer the HER-2/neu gene and protein. *Molecular & Cellular Proteomics*, *3*(4), 379–398.

Rouder, J. N. & Morey, R. D. (2012). Default Bayes Factors for model selection in regression. *Multivariate Behavioral Research*, *47*(6), 877–903.

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing*, *2*(3), 117–119.

Rubin, D. B. et al. (1981). The Bayesian bootstrap. *The Annals of Statistics*, *9*(1), 130–134.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Schwind, C. & Buder, J. (2012). Reducing confirmation bias and evaluation bias: When are preference-inconsistent recommendations effective–and when not? *Computers in Human Behavior*, *28*(6), 2280–2290.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922–932.

Shek, D. T. (1990). Reliability and factorial structure of the Chinese version of the Beck Depression Inventory. *Journal of Clinical Psychology*, *46*(1), 35–43.

Shou, Y. & Smithson, M. (2016). *Cdfquantreg: Quantile regression for random variables on the unit interval.* R package version 1.0.4. Retrieved from http://CRAN.R-project.org/package=cdfquantreg

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Smithson, M. & Shou, Y. (2017). CDF-quantile distributions for modelling random variables on the unit interval. *British Journal of Mathematical and Statistical Psychology*, *70*(3), 412–438.

Souery, D., Papakostas, G. I., & Trivedi, M. H. (2006). Treatment-resistant depression. *Journal of Clinical Psychiatry*, *67*, 16–22.

Spearman, C. (1904a). "General intelligence" Objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292.

Spearman, C. (1904b). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Stan Development Team. (2016). The Stan C++ Library, Version 2.15.0. http://mc-stan.org.

Steckler, A. & McLeroy, K. R. (2008). The importance of external validity. *American Public Health Association*, *98*, 9–10.

Steer, R. A. & Clark, D. A. (1997). Psychometric characteristics of the Beck Depression Inventory-II with college students. *Measurement and Evaluation in Counseling and Development*, *30*(3), 128–136.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103.

Student. (1908). The probable error of a mean. *Biometrika*, *6*, 1–25.

Tukey, J. W. (1960). Contributions to probability and statistics: Essays in honor of Harold Hotelling. In I. Olkin (Ed.), (pp. 448–485). Stanford, CA: Stanford University Press.

Tversky, A. & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. *Progress in Social Psychology*, *1*, 49–72.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes Factor. *Journal of Mathematical Psychology*, *54*(6), 491–498.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wagenmakers, E.-J., Lee, M. D., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York, NY: Statistics for Social and Behavioral Sciences.

Wagenmakers, E.-J., Lee, M. D., Rouder, J., & Morey, R. (2015). Another statistical paradox. *Manuscript submitted for publication.*

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169–176.

Wallace, C. S. & Boulton, D. M. (1968). An information measure for classification. *The Computer Journal*, *11*(2), 185–194.

Wang, C. & Blei, D. M. (2015). A general method for robust Bayesian modeling. *Bayesian Analysis*, *In Press.*

Ward, L. (2015). Health and ageing. *[Unpublished raw data].*

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Welsh, M. & Navarro, D. J. (2012). Seeing is believing: Priors, trust, and base rate neglect. *Organizational Behavior and Human Decision Processes*, *119*, 1–14.

Wetzels, R., Raaijmakers, J. G., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, *16*(4), 752–760.

Wiebe, J. S. & Penley, J. A. (2005). A psychometric comparison of the Beck Depression Inventory-II in English and Spanish. *Psychological Assessment*, *17*(4), 481.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing.* Amsterdam: Academic Press.

Wilcox, R. R. & Schönbrodt, F. D. (2015). *The WRS package for robust statistics in R (version 0.27.5)*. Retrieved from https://github.com/nicebread/WRS

Wong, D., Dahm, J., & Ponsford, J. (2013). Factor structure of the Depression Anxiety Stress Scales in individuals with traumatic brain injury. *Brain Injury*, *27*(12), 1377–1382.

Wyrwich, K. W. (2004). Minimal important difference thresholds and the standard error of measurement: Is there a connection? *Journal of Biopharmaceutical Statistics*, *14*(1), 97–110.

Yap, B. W. & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, *81*(12), 2141–2155.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*, *61*(1), 165–170.

Zeigenfuse, M. D. & Lee, M. D. (2010). A general latent assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*, *54*(4), 352–362.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α Revelle's β, and Mcdonald's ω h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133.