

Learning statistics with R: A tutorial for
psychology students and other beginners.
(Version 0.6.1)

Danielle Navarro (bookdown translation: Emily Kothe)

2020-05-12

Contents

Overview	5
Licensing	7
Dedication	9
Preface	11
0.1 Preface to Version 0.6.1	11
0.2 Preface to Version 0.6	11
0.3 Preface to Version 0.5	12
0.4 Preface to Version 0.4	12
0.5 Preface to Version 0.3	13
1 Una breve introducción a diseños de investigación	17

Overview

Learning Statistics with R covers the contents of an introductory statistics class, as typically taught to undergraduate psychology students, focusing on the use of the R statistical software. The book discusses how to get started in R as well as giving an introduction to data manipulation and writing scripts. From a statistical perspective, the book discusses descriptive statistics and graphing first, followed by chapters on probability theory, sampling and estimation, and null hypothesis testing. After introducing the theory, the book covers the analysis of contingency tables, t-tests, ANOVAs and regression. Bayesian statistics are covered at the end of the book.

Licensing

This book is published under a Creative Commons BY-SA license (CC BY-SA) version 4.0. This means that this book can be reused, remixed, retained, revised and redistributed (including commercially) as long as appropriate credit is given to the authors. If you remix, or modify the original version of this open textbook, you must redistribute all versions of this open textbook under the same license - CC BY-SA. <https://creativecommons.org/licenses/by-sa/4.0/>

Dedication

This book was brought to you today by the letter 'R'.

Preface

0.1 Preface to Version 0.6.1

Hi! I'm not Danielle.

This version is a work in progress to transport the book from LaTeX to book-down, this allows for the code chunks shown in the text to be interactively rendered and is the first step in updating sections to make use of the **tidyverse**.

Moving from LaTeX is a bit fiddly so there are broken pieces in this version that I am updating progressively. If you notice errors or have suggestions feel free to raise an issue (or pull request) at <http://github.com/ekothe/rbook>

Cheers Emily Kothe

0.2 Preface to Version 0.6

The book hasn't changed much since 2015 when I released Version 0.5 – it's probably fair to say that I've changed more than it has. I moved from Adelaide to Sydney in 2016 and my teaching profile at UNSW is different to what it was at Adelaide, and I haven't really had a chance to work on it since arriving here! It's a little strange looking back at this actually. A few quick comments...

- Weirdly, the book *consistently* misgenders me, but I suppose I have only myself to blame for that one :-). There's now a brief footnote on page 12 that mentions this issue; in real life I've been working through a gender affirmation process for the last two years and mostly go by she/her pronouns. I am, however, just as lazy as I ever was so I haven't bothered updating the text in the book.
- For Version 0.6 I haven't changed much I've made a few minor changes when people have pointed out typos or other errors. In particular it's worth noting the issue associated with the `etaSquared` function in the **lsr** package (which isn't really being maintained any more) in Section 14.4. The function works fine for the simple examples in the book, but there

are definitely bugs in there that I haven't found time to check! So please take care with that one.

- The biggest change really is the licensing! I've released it under a Creative Commons licence (CC BY-SA 4.0, specifically), and placed all the source files to the associated GitHub repository, if anyone wants to adapt it.

Maybe someone would like to write a version that makes use of the **tidyverse**... I hear that's become rather important to R these days :-)

Best, Danielle Navarro

0.3 Preface to Version 0.5

Another year, another update. This time around, the update has focused almost entirely on the theory sections of the book. Chapters 9, 10 and 11 have been rewritten, hopefully for the better. Along the same lines, Chapter 17 is entirely new, and focuses on Bayesian statistics. I think the changes have improved the book a great deal. I've always felt uncomfortable about the fact that all the inferential statistics in the book are presented from an orthodox perspective, even though I almost always present Bayesian data analyses in my own work. Now that I've managed to squeeze Bayesian methods into the book somewhere, I'm starting to feel better about the book as a whole. I wanted to get a few other things done in this update, but as usual I'm running into teaching deadlines, so the update has to go out the way it is!

Dan Navarro February 16, 2015

0.4 Preface to Version 0.4

A year has gone by since I wrote the last preface. The book has changed in a few important ways: Chapters 3 and 4 do a better job of documenting some of the time saving features of Rstudio, Chapters 12 and 13 now make use of new functions in the `lsr` package for running chi-square tests and t tests, and the discussion of correlations has been adapted to refer to the new functions in the `lsr` package. The soft copy of 0.4 now has better internal referencing (i.e., actual hyperlinks between sections), though that was introduced in 0.3.1. There's a few tweaks here and there, and many typo corrections (thank you to everyone who pointed out typos!), but overall 0.4 isn't massively different from 0.3.

I wish I'd had more time over the last 12 months to add more content. The absence of any discussion of repeated measures ANOVA and mixed models more generally really does annoy me. My excuse for this lack of progress is that my second child was born at the start of 2013, and so I spent most of last year just trying to keep my head above water. As a consequence, unpaid side projects like this book got sidelined in favour of things that actually pay my salary! Things are a little calmer now, so with any luck version 0.5 will be a bigger step

forward.

One thing that has surprised me is the number of downloads the book gets. I finally got some basic tracking information from the website a couple of months ago, and (after excluding obvious robots) the book has been averaging about 90 downloads per day. That's encouraging: there's at least a few people who find the book useful!

Dan Navarro February 4, 2014

0.5 Preface to Version 0.3

There's a part of me that really **doesn't** want to publish this book. It's not finished.

And when I say that, I mean it. The referencing is spotty at best, the chapter summaries are just lists of section titles, there's no index, there are no exercises for the reader, the organisation is suboptimal, and the coverage of topics is just not comprehensive enough for my liking. Additionally, there are sections with content that I'm not happy with, figures that really need to be redrawn, and I've had almost no time to hunt down inconsistencies, typos, or errors. In other words, *this book is not finished*. If I didn't have a looming teaching deadline and a baby due in a few weeks, I really wouldn't be making this available at all.

What this means is that if you are an academic looking for teaching materials, a Ph.D. student looking to learn R, or just a member of the general public interested in statistics, I would advise you to be cautious. What you're looking at is a first draft, and it may not serve your purposes. If we were living in the days when publishing was expensive and the internet wasn't around, I would never consider releasing a book in this form. The thought of someone shelling out \$80 for this (which is what a commercial publisher told me it would retail for when they offered to distribute it) makes me feel more than a little uncomfortable. However, it's the 21st century, so I can post the pdf on my website for free, and I can distribute hard copies via a print-on-demand service for less than half what a textbook publisher would charge. And so my guilt is assuaged, and I'm willing to share! With that in mind, you can obtain free soft copies and cheap hard copies online, from the following webpages:

- <http://www.compcogscisydney.com/learning-statistics-with-r.html>
- <http://www.lulu.com/content/13570633>

Even so, the warning still stands: what you are looking at is Version 0.3 of a work in progress. If and when it hits Version 1.0, I would be willing to stand behind the work and say, yes, this is a textbook that I would encourage other people to use. At that point, I'll probably start shamelessly flogging the thing on the internet and generally acting like a tool. But until that day comes, I'd like it to be made clear that I'm really ambivalent about the work as it stands.

All of the above being said, there is one group of people that I can enthusiastically endorse this book to: the psychology students taking our undergraduate research methods classes (DRIP and DRIP:A) in 2013. For you, this book is ideal, because it was written to accompany your stats lectures. If a problem arises due to a shortcoming of these notes, I can and will adapt content on the fly to fix that problem. Effectively, you've got a textbook written specifically for your classes, distributed for free (electronic copy) or at near-cost prices (hard copy). Better yet, the notes have been tested: Version 0.1 of these notes was used in the 2011 class, Version 0.2 was used in the 2012 class, and now you're looking at the new and improved Version 0.3. I'm not saying these notes are titanium plated awesomeness on a stick – though if *you* wanted to say so on the student evaluation forms, then you're totally welcome to – because they're not. But I am saying that they've been tried out in previous years and they seem to work okay. Besides, there's a group of us around to troubleshoot if any problems come up, and you can guarantee that at least *one* of your lecturers has read the whole thing cover to cover!

Okay, with all that out of the way, I should say something about what the book aims to be. At its core, it is an introductory statistics textbook pitched primarily at psychology students. As such, it covers the standard topics that you'd expect of such a book: study design, descriptive statistics, the theory of hypothesis testing, t -tests, χ^2 tests, ANOVA and regression. However, there are also several chapters devoted to the R statistical package, including a chapter on data manipulation and another one on scripts and programming. Moreover, when you look at the content presented in the book, you'll notice a lot of topics that are traditionally swept under the carpet when teaching statistics to psychology students. The Bayesian/frequentist divide is openly discussed in the probability chapter, and the disagreement between Neyman and Fisher about hypothesis testing makes an appearance. The difference between probability and density is discussed. A detailed treatment of Type I, II and III sums of squares for unbalanced factorial ANOVA is provided. And if you have a look in the Epilogue, it should be clear that my intention is to add a lot more advanced content.

My reasons for pursuing this approach are pretty simple: the students can handle it, and they even seem to enjoy it. Over the last few years I've been pleasantly surprised at just how little difficulty I've had in getting undergraduate psych students to learn R. It's certainly not easy for them, and I've found I need to be a little charitable in setting marking standards, but they do eventually get there. Similarly, they don't seem to have a lot of problems tolerating ambiguity and complexity in presentation of statistical ideas, as long as they are assured that the assessment standards will be set in a fashion that is appropriate for them. So if the students can handle it, why *not* teach it? The potential gains are pretty enticing. If they learn R, the students get access to CRAN, which is perhaps the largest and most comprehensive library of statistical tools in existence. And if they learn about probability theory in detail, it's easier for them to switch from orthodox null hypothesis testing to Bayesian methods if

they want to. Better yet, they learn data analysis skills that they can take to an employer without being dependent on expensive and proprietary software.

Sadly, this book isn't the silver bullet that makes all this possible. It's a work in progress, and maybe when it is finished it will be a useful tool. One among many, I would think. There are a number of other books that try to provide a basic introduction to statistics using R, and I'm not arrogant enough to believe that mine is better. Still, I rather like the book, and maybe other people will find it useful, incomplete though it is.

Dan Navarro January 13, 2013

Chapter 1

Una breve introducción a diseños de investigación

Consultar al estadístico después de que un experimento ha finalizado es a menudo meramente pedirle conducir una examinación post-mortem. El puede tal vez decir de qué murió el experimento.

– Sir Ronald Fisher¹

En este capítulo, vamos a empezar a pensar las ideas básicas que van en diseñar un estudio, recolectar datos, revisar si tus recolección de datos funciona, y así. No te dará suficiente información que te permita diseñar estudios por tu mismo, pero te dara un montón de herramientas básicas que necesitas para evaluar estudios hechos por otras personas. Sin embargo, desde que el foco de este libro es mucho más sobre análisis de datos que recolección de datos, solo estoy dando una muy breve revisión. Nota que este capítulo es “especial” en dos maneras. Primero, es mucho más específico en psicología que los últimos capítulos. Segundo, se enfoca mucho más intensamente en problema científico de metodología de investigación, y mucho menos en el problema estadístico del análisis de datos. De todos modos, los dos problemas están muy relaciones el uno con el otro, entonces es tradicional para textos de estadística discutir el problema con poco detalle. Este capítulo depende en gran medida en Campbell y Stanley (1963) para la discusión del diseño de estudio, y Stevens (1946) para la discusióboon de las escalas de medición

##Introducción a la medición psicológica{#measurement}

La primera cosa para entender la recolección de datos puede ser pensada como un tipo de **medición**. Esto es, lo que estamos intentando hacer aquí es medir

¹Presidential Address to the First Indian Statistical Congress, 1938. Source: http://en.wikiquote.org/wiki/Ronald_Fisher

algo sobre la conducta humana o la mente humana. ¿Qué quiero decir por “medición”?

1.0.1 Algunos pensamientos sobre medición psicológica

Medición en sí misma es un concepto sutil y delicado, pero básicamente viene a ser una manera de asignar números, o etiquetas, o algún otro tipo de descripciones bien definidas, a “cosas”. Entonces, alguno de los siguientes ejemplos contaría como una medición psicológica:

-Mi **edad** es *33 años*. -No me **gustan las anchoas** -Mi **género cromosómico** es *masculino* -Mi **género auto-identificado** es *masculino* ²

En breve la lista anterior, la **parte ennegrecida** es la cosa a ser “medida”, y la *parte en cursiva* es “la medición en sí misma”. De hecho, podemos expandir esto un poco más, pensando sobre un de posibles mediciones que podrían haber aparecido en cada caso:

-Mi **edad** (en años) podría haber sido *0,1,2,3...,etc.* *El límite superior que mi edad podría ser podría ser un poco difuso, peor en la práctica tu estarías seguro diciendo que la edad más alta posible es 150**, desde que ningún humano ha vivido tanto.

-Cuando se me preguntó si **me gustan las anchoas**, podría haber dicho, *Sí, o No, o No tengo opinión, o A veces*

-Mi **género cromosómico** casi ciertamente iba a ser *masculino (XY) o femenino (XX)*, pero hay unas pocas otras posibilidades- Podría tener el *Síndrome de Klinefelter (XXY), el cual es más similar a masculino que a femenino. E imagino hay otras posibilidad es también.

-Mi **género auto-identificado** es también casi muy probablemente *masculino* o *femenino*, pero no tiene que concordar con mi género cromosómico. Podría también escoger identificar con *ninguno* o explícitamente identificarme como *transgénero*.

²Bien... ahora esto incómodo ¿no? Esta sección del libro es una de las partes más antiguas del libro, y está desactualizada en una manera vergonzosa. Yo escribí esto en el 2010, un punto en el cual todos esos hechos eran verdaderos. Revisitando esto en el 2018, bien yo no tengo sigilo teniendo 33, pero no es una sorpresa supongo. No puedo imaginar que mis cromosomas han cambiado, entonces yo voy a suponer que mi cariotipo era entonces y es ahora XY. El género auto-identificado, por otro lado...ah. Supongo que el hecho de que el título refiera a mí como Danielle en vez de Daniel podría posiblemente ser una pista, pero no me identifico típicamente como “masculino” en un cuestionario de género en estos días, y prefiero el pronombre “ella” por defecto (es una larga historia) ! Pensé un poco sobre como iba a manejar esto en el libro, de hecho. El libro tenía voz autorial distinta, y sentía como podría ser un trabajo diferente si yo iba de vuelta y escribía todo como Danielle y actualizaba todos los pronombres en el trabajo. Así mismo, podría ser un montón de trabajo, entonces dejé mi nombre como “Dan” a través del libro, y en cualquier caso “Dan” es un apodo perfectamente bueno para “Danielle”, ¿No lo crees? En cualquier caso, no es un gran tema. Solo quería mencionarlo para hacer la vida un poco más fácil a los lectores que no saben como referirse a mí. Aun no me gustan las anchoas, de todos modos.

En breve la lista anterior, la **parte enegrecida** es la cosa a ser “medida”, y la *parte en cursiva* es “la medición en sí misma”. De hecho, podemos expandir esto un poco más, pensando sobre un de posibles mediciones que podrían haber aparecido en cada caso:

-Mi **edad** (en años) podría haber sido *0,1,2,3...,etc.* *El límite superior que mi edad podría ser podría ser un poco difuso, peor en la práctica tu estarías seguro diciendo que la edad más alta posible es 150**, desde que ningún humano ha vivido tanto.

-Cuando se me preguntó si **me gustan las anchoas**, podría haber dicho, *Sí, o No, o No tengo opinion, o A veces*

-Mi **género cromosómico** casi ciertamente iba a ser *masculino (XY) o femenino (XX)*, pero hay unas pocas otras posibilidades- Podría tener el *Síndrome de Klinefelter (XXY), el cuál es más similar a masculino que a femenino. E imagino hay otras posibilidad es tambien.

-Mi **genero auto-identificado** es tambien casi muy probablemente *masculino o femenino*, pero no tiene que concordar con mi género cromosómico. Podría tambien escoger identificar con *ninguno* o explicitamente identificarme como *transgénero*.

Como puedes ver, para algunas cosas (como la edad) parece bastante obvio cuál debería ser el conjunto de posibles medidas, mientras que para otras cosas se pone un poco difícil. Pero quiero señalar que incluso en el caso de la edad de alguien, es mucho más sutil que esto. Por ejemplo, en el ejemplo anterior, asumí que estaba bien medir la edad en años. Pero si eres un psicólogo del desarrollo, eso es demasiado burdo, y por eso a menudo se mide la edad en *años y meses* (si un niño tiene 2 años y 11 meses, esto suele escribirse como “2;11”). Si te interesan los recién nacidos, quizás quieras medir la edad en *días desde el nacimiento*, quizás incluso en *horas desde el nacimiento*. En otras palabras, la forma en que se especifican los valores de medición permitidos es importante.

Mirando esto un poco más de cerca, también podrías darte cuenta de que el concepto de “edad” no es tan preciso. En general, cuando decimos “edad” implícitamente queremos decir “el tiempo transcurrido desde el nacimiento”. Pero esa no es siempre la forma correcta de hacerlo. Supongamos que tenemos interés en cómo los recién nacidos controlan sus movimientos oculares. Si te interesa investigar los niños y niñas muy menores, también podrías empezar a preocuparte de que el “nacimiento” no es el único momento significativo para preocuparse. Si la Bebé Alice nace 3 semanas antes de tiempo y la Bebé Bianca nace 1 semana después, ¿tendría sentido decir que tienen la “misma edad” si los encontramos “2 horas después del nacimiento”? En un sentido, sí: por convención social, usamos el nacimiento como nuestro punto de referencia para hablar de la edad en la vida cotidiana, ya que define la cantidad de tiempo que la persona ha estado operando como una entidad independiente en el mundo, pero desde una perspectiva científica eso no es lo único que nos importa. Cuando pensamos en la biología de los seres humanos, a menudo es útil pensar en nosotros mismos como

organismos que han estado creciendo y madurando desde la concepción, y desde esa perspectiva Alice y Bianca no tienen la misma edad en absoluto. Así que tal vez quieras definir el concepto de “edad” de dos maneras diferentes: el tiempo transcurrido desde la concepción y el tiempo transcurrido desde el nacimiento. Cuando se trata de adultos, no habrá mucha diferencia, pero cuando se trata de recién nacidos sí.

Más allá de estos cuestionamientos, está la cuestión de la metodología. ¿Qué “método de medición” específico vas a usar para averiguar la edad de alguien? Como antes, hay muchas posibilidades diferentes:

- Podrías preguntarle a la gente “¿cuántos años tienes?” El método de autoinforme es rápido, barato y fácil, pero sólo funciona con personas lo suficientemente mayores para entender la pregunta, y algunas personas mienten sobre su edad.
- Podrías preguntarle a una autoridad (por ejemplo, a un padre) “¿qué edad tiene tu hijo?” Este método es rápido, y cuando se trata de niños no es tan difícil ya que el padre está casi siempre cerca. No funciona tan bien si quieres saber “la edad desde la concepción”, ya que muchos padres no pueden decir con seguridad cuándo se produjo la concepción. Para eso, puede que necesites una autoridad diferente (por ejemplo, un obstetra).
- Podrías buscar registros oficiales, como certificados de nacimiento. Esto lleva tiempo y es molesto, pero tiene sus usos (por ejemplo, si la persona ya está muerta).

###Operacionalización: definiendo su medición

Todas las ideas discutidas en la sección anterior se relacionan con el concepto de **operación**. Para ser un poco más precisos sobre la idea, la operacionalización es el proceso por el cual tomamos un concepto significativo pero algo vago, y lo convertimos en una medida precisa. El proceso de operacionalización puede implicar varias cosas diferentes:

- Ser preciso sobre lo que se intenta medir. Por ejemplo, ¿“edad” significa “tiempo desde el nacimiento” o “tiempo desde la concepción” en el contexto de su investigación?
- Determinar el método que se utilizará para medirlo. ¿Utilizarás el autoinforme para medir la edad, preguntarás a un padre o buscarás un registro oficial? Si usas el autoinforme, ¿cómo formularás la pregunta?
- Definiendo el conjunto de los valores permitidos que la medición puede tomar. Ten en cuenta que estos valores no siempre tienen que ser numéricos, aunque a menudo lo son. Cuando se mide la edad, los valores son numéricos, pero todavía tenemos que pensar cuidadosamente sobre qué números están permitidos. ¿Queremos la edad en años, años y meses, días, horas? Etc. Para otros tipos de mediciones (por ejemplo, el género), los valores no son numéricos. Pero, al igual que antes, tenemos que pensar en qué valores están permitidos. Si le pedimos a la gente que informe sobre su género, ¿qué opciones les permitimos elegir? ¿Es suficiente con permitir sólo “masculino” o “femenino”? ¿Necesita una opción “otra”?

¿O no deberíamos dar a las personas opciones específicas, y dejar que respondan con sus propias palabras? Y si abrimos el conjunto de posibles valores para incluir todas las respuestas verbales, ¿cómo interpretarán sus respuestas?

La operacionalización es un asunto delicado, y no hay “una manera única y verdadera” de hacerlo. La forma en que se elige para operacionalizar el concepto informal de “edad” o “género” en una medición formal depende de para qué se necesita usar la medición. A menudo encontrarás que la comunidad de científicos que trabajan en tu área tienen algunas ideas bastante bien establecidas sobre cómo hacerlo. En otras palabras, la operacionalización debe ser pensada caso por caso. Sin embargo, aunque hay muchas cuestiones que son específicas de cada proyecto de investigación individual, hay algunos aspectos que son bastante generales.

Antes de seguir adelante, quiero dedicar un momento a aclarar nuestra terminología, y en el proceso introducir un término más. Aquí hay cuatro cosas diferentes que están estrechamente relacionadas entre sí:

- Una **construcción teórica**. Esto es lo que intentas medir, como la “edad”, el “género” o una “opinión”. Un constructo teórico no puede ser observado directamente, y a menudo son en realidad un poco vagos.
- Una **medida**... La medida se refiere al método o la herramienta que se utiliza para hacer las observaciones. Una pregunta en una encuesta, una observación del comportamiento o un escáner cerebral podría contar como una medida.
- Una **operacionalización**. El término “operacionalización” se refiere a la conexión lógica entre la medida y el constructo teórico, o al proceso por el cual intentamos derivar una medida de un constructo teórico.
- Una **variable**. Finalmente, un nuevo término. Una variable es lo que obtenemos cuando aplicamos nuestra medida a algo en el mundo. Es decir, las variables son los “datos” reales con los que terminamos en nuestros conjuntos de datos.

En la práctica, incluso los científicos tienden a desdibujar la distinción entre estas cosas, pero es muy útil tratar de entender las diferencias.

##Scales of measurement{#scales}

As the previous section indicates, the outcome of a psychological measurement is called a variable. But not all variables are of the same qualitative type, and it's very useful to understand what types there are. A very useful concept for distinguishing between different types of variables is what's known as **scales of measurement**.

###Nominal scale

A **nominal scale** variable (also referred to as a **categorical** variable) is one in which there is no particular relationship between the different possibilities: for these kinds of variables it doesn't make any sense to say that one of them

is “bigger’ or”better” than any other one, and it absolutely doesn’t make any sense to average them. The classic example for this is “eye colour”. Eyes can be blue, green and brown, among other possibilities, but none of them is any “better” than any other one. As a result, it would feel really weird to talk about an “average eye colour”. Similarly, gender is nominal too: male isn’t better or worse than female, neither does it make sense to try to talk about an “average gender”. In short, nominal scale variables are those for which the only thing you can say about the different possibilities is that they are different. That’s it.

Let’s take a slightly closer look at this. Suppose I was doing research on how people commute to and from work. One variable I would have to measure would be what kind of transportation people use to get to work. This “transport type” variable could have quite a few possible values, including: “train”, “bus”, “car”, “bicycle”, etc. For now, let’s suppose that these four are the only possibilities, and suppose that when I ask 100 people how they got to work today, and I get this:

Transportation	Number of people
(1) Train	12
(2) Bus	30
(3) Car	48
(4) Bicycle	10

So, what’s the average transportation type? Obviously, the answer here is that there isn’t one. It’s a silly question to ask. You can say that travel by car is the most popular method, and travel by train is the least popular method, but that’s about all. Similarly, notice that the order in which I list the options isn’t very interesting. I could have chosen to display the data like this

Transportation	Number of people
(3) Car	48
(1) Train	12
(4) Bicycle	10
(2) Bus	30

and nothing really changes.

###Ordinal scale

Ordinal scale variables have a bit more structure than nominal scale variables, but not by a lot. An ordinal scale variable is one in which there is a natural, meaningful way to order the different possibilities, but you can’t do anything else. The usual example given of an ordinal variable is “finishing position in a race”. You *can* say that the person who finished first was faster than the person

who finished second, but you *don't* know how much faster. As a consequence we know that 1st > 2nd, and we know that 2nd > 3rd, but the difference between 1st and 2nd might be much larger than the difference between 2nd and 3rd.

Here's an more psychologically interesting example. Suppose I'm interested in people's attitudes to climate change, and I ask them to pick one of these four statements that most closely matches their beliefs:

- (1) Temperatures are rising, because of human activity
- (2) Temperatures are rising, but we don't know why
- (3) Temperatures are rising, but not because of humans
- (4) Temperatures are not rising

Notice that these four statements actually do have a natural ordering, in terms of "the extent to which they agree with the current science". Statement 1 is a close match, statement 2 is a reasonable match, statement 3 isn't a very good match, and statement 4 is in strong opposition to the science. So, in terms of the thing I'm interested in (the extent to which people endorse the science), I can order the items as $1 > 2 > 3 > 4$. Since this ordering exists, it would be very weird to list the options like this...

- (3) Temperatures are rising, but not because of humans
- (4) Temperatures are rising, because of human activity
- (5) Temperatures are not rising
- (6) Temperatures are rising, but we don't know why

... because it seems to violate the natural "structure" to the question.

So, let's suppose I asked 100 people these questions, and got the following answers:

Response	Number
(1) Temperatures are rising, because of human activity	51
(2) Temperatures are rising, but we don't know why	20
(3) Temperatures are rising, but not because of humans	10
(4) Temperatures are not rising	19

When analysing these data, it seems quite reasonable to try to group (1), (2) and (3) together, and say that 81 of 100 people were willing to *at least partially* endorse the science. And it's *also* quite reasonable to group (2), (3) and (4) together and say that 49 of 100 people registered *at least some disagreement* with the dominant scientific view. However, it would be entirely bizarre to try to group (1), (2) and (4) together and say that 90 of 100 people said... what? There's nothing sensible that allows you to group those responses together at all.

That said, notice that while we *can* use the natural ordering of these items to construct sensible groupings, what we *can't* do is average them. For instance,

in my simple example here, the “average” response to the question is 1.97. If you can tell me what that means, I’d love to know. Because that sounds like gibberish to me!

Interval scale

In contrast to nominal and ordinal scale variables, *interval scale* and ratio scale variables are variables for which the numerical value is genuinely meaningful. In the case of interval scale variables, the *differences* between the numbers are interpretable, but the variable doesn’t have a “natural” zero value. A good example of an interval scale variable is measuring temperature in degrees celsius. For instance, if it was 15° yesterday and 18° today, then the 3° difference between the two is genuinely meaningful. Moreover, that 3° difference is *exactly the same* as the 3° difference between 7° and 10°. In short, addition and subtraction are meaningful for interval scale variables.³

However, notice that the 0° does not mean “no temperature at all”: it actually means “the temperature at which water freezes”, which is pretty arbitrary. As a consequence, it becomes pointless to try to multiply and divide temperatures. It is wrong to say that 20° is *twice as hot* as 10°, just as it is weird and meaningless to try to claim that 20° is negative two times as hot as -10°.

Again, let’s look at a more psychological example. Suppose I’m interested in looking at how the attitudes of first-year university students have changed over time. Obviously, I’m going to want to record the year in which each student started. This is an interval scale variable. A student who started in 2003 did arrive 5 years before a student who started in 2008. However, it would be completely insane for me to divide 2008 by 2003 and say that the second student started “1.0024 times later” than the first one. That doesn’t make any sense at all.

Ratio scale

The fourth and final type of variable to consider is a *ratio scale* variable, in which zero really means zero, and it’s okay to multiply and divide. A good psychological example of a ratio scale variable is response time (RT). In a lot of tasks it’s very common to record the amount of time somebody takes to solve a problem or answer a question, because it’s an indicator of how difficult the task is. Suppose that Alan takes 2.3 seconds to respond to a question, whereas Ben takes 3.1 seconds. As with an interval scale variable, addition and subtraction are both meaningful here. Ben really did take $3.1 - 2.3 = 0.8$ seconds longer than Alan did. However, notice that multiplication and division also make sense here too: Ben took $3.1 / 2.3 = 1.35$ times as long as Alan did to answer the

³Actually, I’ve been informed by readers with greater physics knowledge than I that temperature isn’t strictly an interval scale, in the sense that the amount of energy required to heat something up by 3° depends on its current temperature. So in the sense that physicists care about, temperature isn’t actually interval scale. But it still makes a cute example, so I’m going to ignore this little inconvenient truth.

question. And the reason why you can do this is that, for a ratio scale variable such as RT, “zero seconds” really does mean “no time at all”.

1.0.2 Continuous versus discrete variables

There’s a second kind of distinction that you need to be aware of, regarding what types of variables you can run into. This is the distinction between continuous variables and discrete variables. The difference between these is as follows:

- A **continuous variable** is one in which, for any two values that you can think of, it’s always logically possible to have another value in between.
- A **discrete variable** is, in effect, a variable that isn’t continuous. For a discrete variable, it’s sometimes the case that there’s nothing in the middle.

These definitions probably seem a bit abstract, but they’re pretty simple once you see some examples. For instance, response time is continuous. If Alan takes 3.1 seconds and Ben takes 2.3 seconds to respond to a question, then it’s possible for Cameron’s response time to lie in between, by taking 3.0 seconds. And of course it would also be possible for David to take 3.031 seconds to respond, meaning that his RT would lie in between Cameron’s and Alan’s. And while in practice it might be impossible to measure RT that precisely, it’s certainly possible in principle. Because we can always find a new value for RT in between any two other ones, we say that RT is continuous.

Discrete variables occur when this rule is violated. For example, nominal scale variables are always discrete: there isn’t a type of transportation that falls “in between” trains and bicycles, not in the strict mathematical way that 2.3 falls in between 2 and 3. So transportation type is discrete. Similarly, ordinal scale variables are always discrete: although “2nd place” does fall between “1st place” and “3rd place”, there’s nothing that can logically fall in between “1st place” and “2nd place”. Interval scale and ratio scale variables can go either way. As we saw above, response time (a ratio scale variable) is continuous. Temperature in degrees celsius (an interval scale variable) is also continuous. However, the year you went to school (an interval scale variable) is discrete. There’s no year in between 2002 and 2003. The number of questions you get right on a true-or-false test (a ratio scale variable) is also discrete: since a true-or-false question doesn’t allow you to be “partially correct”, there’s nothing in between 5/10 and 6/10. Table 1.4 summarises the relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible. I’m trying to hammer this point home, because (a) some textbooks get this wrong, and (b) people very often say things like “discrete variable” when they mean “nominal scale variable”. It’s very unfortunate.

###Some complexities

Okay, I know you’re going to be shocked to hear this, but ... the real world is much messier than this little classification scheme suggests. Very few variables

Table 1.4: The relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible.

	continuous	discrete
nominal		\$\checkmark\$
ordinal		\$\checkmark\$
interval	\$\checkmark\$	\$\checkmark\$
ratio	\$\checkmark\$	\$\checkmark\$

in real life actually fall into these nice neat categories, so you need to be kind of careful not to treat the scales of measurement as if they were hard and fast rules. It doesn't work like that: they're guidelines, intended to help you think about the situations in which you should treat different variables differently. Nothing more.

So let's take a classic example, maybe *the* classic example, of a psychological measurement tool: the **Likert scale**. The humble Likert scale is the bread and butter tool of all survey design. You yourself have filled out hundreds, maybe thousands of them, and odds are you've even used one yourself. Suppose we have a survey question that looks like this:

Which of the following best describes your opinion of the statement
that "all pirates are freaking awesome" ...

and then the options presented to the participant are these:

- (1) Strongly disagree
- (2) Disagree
- (3) Neither agree nor disagree
- (4) Agree
- (5) Strongly agree

This set of items is an example of a 5-point Likert scale: people are asked to choose among one of several (in this case 5) clearly ordered possibilities, generally with a verbal descriptor given in each case. However, it's not necessary that all items be explicitly described. This is a perfectly good example of a 5-point Likert scale too:

- (1) Strongly disagree
- (2)
- (3)
- (4)
- (5) Strongly agree

Likert scales are very handy, if somewhat limited, tools. The question is, what kind of variable are they? They're obviously discrete, since you can't give a response of 2.5. They're obviously not nominal scale, since the items are ordered; and they're not ratio scale either, since there's no natural zero.

But are they ordinal scale or interval scale? One argument says that we can't really prove that the difference between "strongly agree" and "agree" is of the same size as the difference between "agree" and "neither agree nor disagree". In fact, in everyday life it's pretty obvious that they're not the same at all. So this suggests that we ought to treat Likert scales as ordinal variables. On the other hand, in practice most participants do seem to take the whole "on a scale from 1 to 5" part fairly seriously, and they tend to act as if the differences between the five response options were fairly similar to one another. As a consequence, a lot of researchers treat Likert scale data as if it were interval scale. It's not interval scale, but in practice it's close enough that we usually think of it as being *quasi-interval scale*.

##Assessing the reliability of a measurement{#reliability}

At this point we've thought a little bit about how to operationalise a theoretical construct and thereby create a psychological measure; and we've seen that by applying psychological measures we end up with variables, which can come in many different types. At this point, we should start discussing the obvious question: is the measurement any good? We'll do this in terms of two related ideas: *reliability* and *validity*. Put simply, the *reliability* of a measure tells you how *precisely* you are measuring something, whereas the validity of a measure tells you how *accurate* the measure is. In this section I'll talk about reliability; we'll talk about validity in the next chapter.

Reliability is actually a very simple concept: it refers to the repeatability or consistency of your measurement. The measurement of my weight by means of a "bathroom scale" is very reliable: if I step on and off the scales over and over again, it'll keep giving me the same answer. Measuring my intelligence by means of "asking my mum" is very unreliable: some days she tells me I'm a bit thick, and other days she tells me I'm a complete moron. Notice that this concept of reliability is different to the question of whether the measurements are correct (the correctness of a measurement relates to its validity). If I'm holding a sack of potatoes when I step on and off of the bathroom scales, the measurement will still be reliable: it will always give me the same answer. However, this highly reliable answer doesn't match up to my true weight at all, therefore it's wrong. In technical terms, this is a *reliable but invalid* measurement. Similarly, while my mum's estimate of my intelligence is a bit unreliable, she might be right. Maybe I'm just not too bright, and so while her estimate of my intelligence fluctuates pretty wildly from day to day, it's basically right. So that would be an *unreliable but valid* measure. Of course, to some extent, notice that if my mum's estimates are too unreliable, it's going to be very hard to figure out which one of her many claims about my intelligence is actually the right one. To some extent, then, a very unreliable measure tends to end up being invalid

for practical purposes; so much so that many people would say that reliability is necessary (but not sufficient) to ensure validity.

Okay, now that we're clear on the distinction between reliability and validity, let's have a think about the different ways in which we might measure reliability:

- **Test-retest reliability.** This relates to consistency over time: if we repeat the measurement at a later date, do we get the same answer?
- **Inter-rater reliability.** This relates to consistency across people: if someone else repeats the measurement (e.g., someone else rates my intelligence) will they produce the same answer?
- **Parallel forms reliability.** This relates to consistency across theoretically-equivalent measurements: if I use a different set of bathroom scales to measure my weight, does it give the same answer?
- **Internal consistency reliability.** If a measurement is constructed from lots of different parts that perform similar functions (e.g., a personality questionnaire result is added up across several questions) do the individual parts tend to give similar answers.

Not all measurements need to possess all forms of reliability. For instance, educational assessment can be thought of as a form of measurement. One of the subjects that I teach, *Computational Cognitive Science*, has an assessment structure that has a research component and an exam component (plus other things). The exam component is *intended* to measure something different from the research component, so the assessment as a whole has low internal consistency. However, within the exam there are several questions that are intended to (approximately) measure the same things, and those tend to produce similar outcomes; so the exam on its own has a fairly high internal consistency. Which is as it should be. You should only demand reliability in those situations where you want to measure the same thing!

##The “role” of variables: predictors and outcomes{#ivdv}

Okay, I've got one last piece of terminology that I need to explain to you before moving away from variables. Normally, when we do some research we end up with lots of different variables. Then, when we analyse our data we usually try to explain some of the variables in terms of some of the other variables. It's important to keep the two roles “thing doing the explaining” and “thing being explained” distinct. So let's be clear about this now. Firstly, we might as well get used to the idea of using mathematical symbols to describe variables, since it's going to happen over and over again. Let's denote the “to be explained” variable Y , and denote the variables “doing the explaining” as X_1, X_2 , etc.

Now, when we doing an analysis, we have different names for X and Y , since they play different roles in the analysis. The classical names for these roles are **independent variable** (IV) and **dependent variable** (DV). The IV is the variable that you use to do the explaining (i.e., X) and the DV is the variable being explained (i.e., Y). The logic behind these names goes like this: if there really is a relationship between X and Y then we can say that Y depends on

Table 1.5: The terminology used to distinguish between different roles that a variable can play when analysing a data set. Note that this book will tend to avoid the classical terminology in favour of the newer names.

role of the variable	classical name	modern name
to be explained	dependent variable (DV)	outcome
to do the explaining	independent variable (IV)	predictor

X , and if we have designed our study “properly” then X isn’t dependent on anything else. However, I personally find those names horrible: they’re hard to remember and they’re highly misleading, because (a) the IV is never actually “independent of everything else” and (b) if there’s no relationship, then the DV doesn’t actually depend on the IV. And in fact, because I’m not the only person who thinks that IV and DV are just awful names, there are a number of alternatives that I find more appealing. The terms that I’ll use in these notes are ***predictors*** and ***outcomes***. The idea here is that what you’re trying to do is use X (the predictors) to make guesses about Y (the outcomes).⁴ This is summarised in Table 1.5.

##Experimental and non-experimental research{#researchdesigns}

One of the big distinctions that you should be aware of is the distinction between “experimental research” and “non-experimental research”. When we make this distinction, what we’re really talking about is the degree of control that the researcher exercises over the people and events in the study.

###Experimental research

The key features of ***experimental research*** is that the researcher controls all aspects of the study, especially what participants experience during the study. In particular, the researcher manipulates or varies the predictor variables (IVs), and then allows the outcome variable (DV) to vary naturally. The idea here is to deliberately vary the predictors (IVs) to see if they have any causal effects on the outcomes. Moreover, in order to ensure that there’s no chance that something other than the predictor variables is causing the outcomes, everything else is kept constant or is in some other way “balanced” to ensure that they have no effect on the results. In practice, it’s almost impossible to *think* of everything else that might have an influence on the outcome of an experiment, much less keep it constant. The standard solution to this is ***randomisation***: that is, we randomly assign people to different groups, and then give each group a different treatment (i.e., assign them different values of the predictor variables). We’ll talk more about randomisation later in this course, but for now, it’s enough to

⁴Annoyingly, though, there’s a lot of different names used out there. I won’t list all of them – there would be no point in doing that – other than to note that R often uses “response variable” where I’ve used “outcome”, and a traditionalist would use “dependent variable”. Sigh. This sort of terminological confusion is very common, I’m afraid.

say that what randomisation does is minimise (but not eliminate) the chances that there are any systematic difference between groups.

Let's consider a very simple, completely unrealistic and grossly unethical example. Suppose you wanted to find out if smoking causes lung cancer. One way to do this would be to find people who smoke and people who don't smoke, and look to see if smokers have a higher rate of lung cancer. This is *not* a proper experiment, since the researcher doesn't have a lot of control over who is and isn't a smoker. And this really matters: for instance, it might be that people who choose to smoke cigarettes also tend to have poor diets, or maybe they tend to work in asbestos mines, or whatever. The point here is that the groups (smokers and non-smokers) actually differ on lots of things, not *just* smoking. So it might be that the higher incidence of lung cancer among smokers is caused by something else, not by smoking per se. In technical terms, these other things (e.g. diet) are called "confounds", and we'll talk about those in just a moment.

In the meantime, let's now consider what a proper experiment might look like. Recall that our concern was that smokers and non-smokers might differ in lots of ways. The solution, as long as you have no ethics, is to *control* who smokes and who doesn't. Specifically, if we randomly divide participants into two groups, and force half of them to become smokers, then it's very unlikely that the groups will differ in any respect other than the fact that half of them smoke. That way, if our smoking group gets cancer at a higher rate than the non-smoking group, then we can feel pretty confident that (a) smoking does cause cancer and (b) we're murderers.

Non-experimental research

Non-experimental research is a broad term that covers "any study in which the researcher doesn't have quite as much control as they do in an experiment". Obviously, control is something that scientists like to have, but as the previous example illustrates, there are lots of situations in which you can't or shouldn't try to obtain that control. Since it's grossly unethical (and almost certainly criminal) to force people to smoke in order to find out if they get cancer, this is a good example of a situation in which you really shouldn't try to obtain experimental control. But there are other reasons too. Even leaving aside the ethical issues, our "smoking experiment" does have a few other issues. For instance, when I suggested that we "force" half of the people to become smokers, I must have been talking about *starting* with a sample of non-smokers, and then forcing them to become smokers. While this sounds like the kind of solid, evil experimental design that a mad scientist would love, it might not be a very sound way of investigating the effect in the real world. For instance, suppose that smoking only causes lung cancer when people have poor diets, and suppose also that people who normally smoke do tend to have poor diets. However, since the "smokers" in our experiment aren't "natural" smokers (i.e., we forced non-smokers to become smokers; they didn't take on all of the other normal, real life characteristics that smokers might tend to possess) they probably have better diets. As such, in this silly example they wouldn't get lung cancer, and

our experiment will fail, because it violates the structure of the “natural” world (the technical name for this is an “artifactual” result; see later).

One distinction worth making between two types of non-experimental research is the difference between *quasi-experimental research* and *case studies*. The example I discussed earlier – in which we wanted to examine incidence of lung cancer among smokers and non-smokers, without trying to control who smokes and who doesn’t – is a quasi-experimental design. That is, it’s the same as an experiment, but we don’t control the predictors (IVs). We can still use statistics to analyse the results, it’s just that we have to be a lot more careful.

The alternative approach, case studies, aims to provide a very detailed description of one or a few instances. In general, you can’t use statistics to analyse the results of case studies, and it’s usually very hard to draw any general conclusions about “people in general” from a few isolated examples. However, case studies are very useful in some situations. Firstly, there are situations where you don’t have any alternative: neuropsychology has this issue a lot. Sometimes, you just can’t find a lot of people with brain damage in a specific area, so the only thing you can do is describe those cases that you do have in as much detail and with as much care as you can. However, there’s also some genuine advantages to case studies: because you don’t have as many people to study, you have the ability to invest lots of time and effort trying to understand the specific factors at play in each case. This is a very valuable thing to do. As a consequence, case studies can complement the more statistically-oriented approaches that you see in experimental and quasi-experimental designs. We won’t talk much about case studies in these lectures, but they are nevertheless very valuable tools!

##Assessing the validity of a study{#validity}

More than any other thing, a scientist wants their research to be “valid”. The conceptual idea behind *validity* is very simple: can you trust the results of your study? If not, the study is invalid. However, while it’s easy to state, in practice it’s much harder to check validity than it is to check reliability. And in all honesty, there’s no precise, clearly agreed upon notion of what validity actually is. In fact, there’s lots of different kinds of validity, each of which raises it’s own issues, and not all forms of validity are relevant to all studies. I’m going to talk about five different types:

- Internal validity
- External validity
- Construct validity
- Face validity
- Ecological validity

To give you a quick guide as to what matters here... (1) Internal and external validity are the most important, since they tie directly to the fundamental question of whether your study really works. (2) Construct validity asks whether you’re measuring what you think you are. (3) Face validity isn’t terribly important except insofar as you care about “appearances”. (4) Ecological validity

is a special case of face validity that corresponds to a kind of appearance that you might care about a lot.

Internal validity

Internal validity refers to the extent to which you are able draw the correct conclusions about the causal relationships between variables. It's called "internal" because it refers to the relationships between things "inside" the study. Let's illustrate the concept with a simple example. Suppose you're interested in finding out whether a university education makes you write better. To do so, you get a group of first year students, ask them to write a 1000 word essay, and count the number of spelling and grammatical errors they make. Then you find some third-year students, who obviously have had more of a university education than the first-years, and repeat the exercise. And let's suppose it turns out that the third-year students produce fewer errors. And so you conclude that a university education improves writing skills. Right? Except... the big problem that you have with this experiment is that the third-year students are older, and they've had more experience with writing things. So it's hard to know for sure what the causal relationship is: Do older people write better? Or people who have had more writing experience? Or people who have had more education? Which of the above is the true *cause* of the superior performance of the third-years? Age? Experience? Education? You can't tell. This is an example of a failure of internal validity, because your study doesn't properly tease apart the *causal* relationships between the different variables.

External validity

External validity relates to the *generalisability* of your findings. That is, to what extent do you expect to see the same pattern of results in "real life" as you saw in your study. To put it a bit more precisely, any study that you do in psychology will involve a fairly specific set of questions or tasks, will occur in a specific environment, and will involve participants that are drawn from a particular subgroup. So, if it turns out that the results don't actually generalise to people and situations beyond the ones that you studied, then what you've got is a lack of external validity.

The classic example of this issue is the fact that a very large proportion of studies in psychology will use undergraduate psychology students as the participants. Obviously, however, the researchers don't care *only* about psychology students; they care about people in general. Given that, a study that uses only psych students as participants always carries a risk of lacking external validity. That is, if there's something "special" about psychology students that makes them different to the general populace in some *relevant* respect, then we may start worrying about a lack of external validity.

That said, it is absolutely critical to realise that a study that uses only psychology students does not necessarily have a problem with external validity. I'll talk about this again later, but it's such a common mistake that I'm going to mention it here. The external validity is threatened by the choice of population

if (a) the population from which you sample your participants is very narrow (e.g., psych students), and (b) the narrow population that you sampled from is systematically different from the general population, *in some respect that is relevant to the psychological phenomenon that you intend to study*. The italicised part is the bit that lots of people forget: it is true that psychology undergraduates differ from the general population in lots of ways, and so a study that uses only psych students *may* have problems with external validity. However, if those differences aren't very relevant to the phenomenon that you're studying, then there's nothing to worry about. To make this a bit more concrete, here's two extreme examples:

- You want to measure “attitudes of the general public towards psychotherapy”, but all of your participants are psychology students. This study would almost certainly have a problem with external validity.
- You want to measure the effectiveness of a visual illusion, and your participants are all psychology students. This study is very unlikely to have a problem with external validity

Having just spent the last couple of paragraphs focusing on the choice of participants (since that's the big issue that everyone tends to worry most about), it's worth remembering that external validity is a broader concept. The following are also examples of things that might pose a threat to external validity, depending on what kind of study you're doing:

- People might answer a “psychology questionnaire” in a manner that doesn't reflect what they would do in real life.
- Your lab experiment on (say) “human learning” has a different structure to the learning problems people face in real life.

Construct validity

Construct validity is basically a question of whether you're measuring what you want to be measuring. A measurement has good construct validity if it is actually measuring the correct theoretical construct, and bad construct validity if it doesn't. To give very simple (if ridiculous) example, suppose I'm trying to investigate the rates with which university students cheat on their exams. And the way I attempt to measure it is by asking the cheating students to stand up in the lecture theatre so that I can count them. When I do this with a class of 300 students, 0 people claim to be cheaters. So I therefore conclude that the proportion of cheaters in my class is 0%. Clearly this is a bit ridiculous. But the point here is not that this is a very deep methodological example, but rather to explain what construct validity is. The problem with my measure is that while I'm *trying* to measure “the proportion of people who cheat” what I'm actually measuring is “the proportion of people stupid enough to own up to cheating, or bloody minded enough to pretend that they do”. Obviously, these aren't the same thing! So my study has gone wrong, because my measurement has very poor construct validity.

Face validity

Face validity simply refers to whether or not a measure “looks like” it’s doing what it’s supposed to, nothing more. If I design a test of intelligence, and people look at it and they say “no, that test doesn’t measure intelligence”, then the measure lacks face validity. It’s as simple as that. Obviously, face validity isn’t very important from a pure scientific perspective. After all, what we care about is whether or not the measure *actually* does what it’s supposed to do, not whether it *looks like* it does what it’s supposed to do. As a consequence, we generally don’t care very much about face validity. That said, the concept of face validity serves three useful pragmatic purposes:

- Sometimes, an experienced scientist will have a “hunch” that a particular measure won’t work. While these sorts of hunches have no strict evidentiary value, it’s often worth paying attention to them. Because often times people have knowledge that they can’t quite verbalise, so there might be something to worry about even if you can’t quite say why. In other words, when someone you trust criticises the face validity of your study, it’s worth taking the time to think more carefully about your design to see if you can think of reasons why it might go awry. Mind you, if you don’t find any reason for concern, then you should probably not worry: after all, face validity really doesn’t matter much.
- Often (very often), completely uninformed people will also have a “hunch” that your research is crap. And they’ll criticise it on the internet or something. On close inspection, you’ll often notice that these criticisms are actually focused entirely on how the study “looks”, but not on anything deeper. The concept of face validity is useful for gently explaining to people that they need to substantiate their arguments further.
- Expanding on the last point, if the beliefs of untrained people are critical (e.g., this is often the case for applied research where you actually want to convince policy makers of something or other) then you *have* to care about face validity. Simply because – whether you like it or not – a lot of people will use face validity as a proxy for real validity. If you want the government to change a law on scientific, psychological grounds, then it won’t matter how good your studies “really” are. If they lack face validity, you’ll find that politicians ignore you. Of course, it’s somewhat unfair that policy often depends more on appearance than fact, but that’s how things go.

Ecological validity

Ecological validity is a different notion of validity, which is similar to external validity, but less important. The idea is that, in order to be ecologically valid, the entire set up of the study should closely approximate the real world scenario that is being investigated. In a sense, ecological validity is a kind of face validity – it relates mostly to whether the study “looks” right, but with a bit more rigour to it. To be ecologically valid, the study has to look right in a fairly specific way. The idea behind it is the intuition that a study that is ecologically valid is more likely to be externally valid. It’s no guarantee, of course. But the

nice thing about ecological validity is that it's much easier to check whether a study is ecologically valid than it is to check whether a study is externally valid. An simple example would be eyewitness identification studies. Most of these studies tend to be done in a university setting, often with fairly simple array of faces to look at rather than a line up. The length of time between seeing the "criminal" and being asked to identify the suspect in the "line up" is usually shorter. The "crime" isn't real, so there's no chance that the witness being scared, and there's no police officers present, so there's not as much chance of feeling pressured. These things all mean that the study *definitely* lacks ecological validity. They might (but might not) mean that it also lacks external validity.

##Confounds, artifacts and other threats to validity

If we look at the issue of validity in the most general fashion, the two biggest worries that we have are *confounds* and *artifact*. These two terms are defined in the following way:

- **Confound:** A confound is an additional, often unmeasured variable⁵ that turns out to be related to both the predictors and the outcomes. The existence of confounds threatens the internal validity of the study because you can't tell whether the predictor causes the outcome, or if the confounding variable causes it, etc.
- **Artifact:** A result is said to be "artifactual" if it only holds in the special situation that you happened to test in your study. The possibility that your result is an artifact describes a threat to your external validity, because it raises the possibility that you can't generalise your results to the actual population that you care about.

As a general rule confounds are a bigger concern for non-experimental studies, precisely because they're not proper experiments: by definition, you're leaving lots of things uncontrolled, so there's a lot of scope for confounds working their way into your study. Experimental research tends to be much less vulnerable to confounds: the more control you have over what happens during the study, the more you can prevent confounds from appearing.

However, there's always swings and roundabouts, and when we start thinking about artifacts rather than confounds, the shoe is very firmly on the other foot. For the most part, artifactual results tend to be a concern for experimental studies than for non-experimental studies. To see this, it helps to realise that the reason that a lot of studies are non-experimental is precisely because what the researcher is trying to do is examine human behaviour in a more naturalistic context. By working in a more real-world context, you lose experimental control (making yourself vulnerable to confounds) but because you tend to be studying

⁵The reason why I say that it's unmeasured is that if you *have* measured it, then you can use some fancy statistical tricks to deal with the confound. Because of the existence of these statistical solutions to the problem of confounds, we often refer to a confound that we have measured and dealt with as a *covariate*. Dealing with covariates is a topic for a more advanced course, but I thought I'd mention it in passing, since it's kind of comforting to at least know that this stuff exists.

human psychology “in the wild” you reduce the chances of getting an artifactual result. Or, to put it another way, when you take psychology out of the wild and bring it into the lab (which we usually have to do to gain our experimental control), you always run the risk of accidentally studying something different than you wanted to study: which is more or less the definition of an artifact.

Be warned though: the above is a rough guide only. It’s absolutely possible to have confounds in an experiment, and to get artifactual results with non-experimental studies. This can happen for all sorts of reasons, not least of which is researcher error. In practice, it’s really hard to think everything through ahead of time, and even very good researchers make mistakes. But other times it’s unavoidable, simply because the researcher has ethics (e.g., see ??).

Okay. There’s a sense in which almost any threat to validity can be characterised as a confound or an artifact: they’re pretty vague concepts. So let’s have a look at some of the most common examples...

History effects

History effects refer to the possibility that specific events may occur during the study itself that might influence the outcomes. For instance, something might happen in between a pre-test and a post-test. Or, in between testing participant 23 and participant 24. Alternatively, it might be that you’re looking at an older study, which was perfectly valid for its time, but the world has changed enough since then that the conclusions are no longer trustworthy. Examples of things that would count as history effects:

- You’re interested in how people think about risk and uncertainty. You started your data collection in December 2010. But finding participants and collecting data takes time, so you’re still finding new people in February 2011. Unfortunately for you (and even more unfortunately for others), the Queensland floods occurred in January 2011, causing billions of dollars of damage and killing many people. Not surprisingly, the people tested in February 2011 express quite different beliefs about handling risk than the people tested in December 2010. Which (if any) of these reflects the “true” beliefs of participants? I think the answer is probably both: the Queensland floods genuinely changed the beliefs of the Australian public, though possibly only temporarily. The key thing here is that the “history” of the people tested in February is quite different to people tested in December.
- You’re testing the psychological effects of a new anti-anxiety drug. So what you do is measure anxiety before administering the drug (e.g., by self-report, and taking physiological measures, let’s say), then you administer the drug, and then you take the same measures afterwards. In the middle, however, because your labs are in Los Angeles, there’s an earthquake, which increases the anxiety of the participants.

Maturation effects

As with history effects, *maturational effects* are fundamentally about change over time. However, maturation effects aren't in response to specific events. Rather, they relate to how people change on their own over time: we get older, we get tired, we get bored, etc. Some examples of maturation effects:

- When doing developmental psychology research, you need to be aware that children grow up quite rapidly. So, suppose that you want to find out whether some educational trick helps with vocabulary size among 3 year olds. One thing that you need to be aware of is that the vocabulary size of children that age is growing at an incredible rate (multiple words per day), all on its own. If you design your study without taking this maturational effect into account, then you won't be able to tell if your educational trick works.
- When running a very long experiment in the lab (say, something that goes for 3 hours), it's very likely that people will begin to get bored and tired, and that this maturational effect will cause performance to decline, regardless of anything else going on in the experiment

Repeated testing effects

An important type of history effect is the effect of *repeated testing*. Suppose I want to take two measurements of some psychological construct (e.g., anxiety). One thing I might be worried about is if the first measurement has an effect on the second measurement. In other words, this is a history effect in which the "event" that influences the second measurement is the first measurement itself! This is not at all uncommon. Examples of this include:

- *Learning and practice*: e.g., "intelligence" at time 2 might appear to go up relative to time 1 because participants learned the general rules of how to solve "intelligence-test-style" questions during the first testing session.
- *Familiarity with the testing situation*: e.g., if people are nervous at time 1, this might make performance go down; after sitting through the first testing situation, they might calm down a lot precisely because they've seen what the testing looks like.
- *Auxiliary changes caused by testing*: e.g., if a questionnaire assessing mood is boring, then mood at measurement at time 2 is more likely to become "bored", precisely because of the boring measurement made at time 1.

Selection bias

Selection bias is a pretty broad term. Suppose that you're running an experiment with two groups of participants, where each group gets a different "treatment", and you want to see if the different treatments lead to different outcomes. However, suppose that, despite your best efforts, you've ended up with a gender imbalance across groups (say, group A has 80% females and group B has 50% females). It might sound like this could never happen, but trust me, it can. This is an example of a selection bias, in which the people "selected into" the two groups have different characteristics. If any of those characteris-

tics turns out to be relevant (say, your treatment works better on females than males) then you're in a lot of trouble.

###Differential attrition{#differentialattrition}

One quite subtle danger to be aware of is called *differential attrition*, which is a kind of selection bias that is caused by the study itself. Suppose that, for the first time ever in the history of psychology, I manage to find the perfectly balanced and representative sample of people. I start running “Dan’s incredibly long and tedious experiment” on my perfect sample, but then, because my study is incredibly long and tedious, lots of people start dropping out. I can’t stop this: as we’ll discuss later in the chapter on research ethics, participants absolutely have the right to stop doing any experiment, any time, for whatever reason they feel like, and as researchers we are morally (and professionally) obliged to remind people that they do have this right. So, suppose that “Dan’s incredibly long and tedious experiment” has a very high drop out rate. What do you suppose the odds are that this drop out is random? Answer: zero. Almost certainly, the people who remain are more conscientious, more tolerant of boredom etc than those that leave. To the extent that (say) conscientiousness is relevant to the psychological phenomenon that I care about, this attrition can decrease the validity of my results.

When thinking about the effects of differential attrition, it is sometimes helpful to distinguish between two different types. The first is *homogeneous attrition*, in which the attrition effect is the same for all groups, treatments or conditions. In the example I gave above, the differential attrition would be homogeneous if (and only if) the easily bored participants are dropping out of all of the conditions in my experiment at about the same rate. In general, the main effect of homogeneous attrition is likely to be that it makes your sample unrepresentative. As such, the biggest worry that you’ll have is that the generalisability of the results decreases: in other words, you lose external validity.

The second type of differential attrition is *heterogeneous attrition*, in which the attrition effect is different for different groups. This is a much bigger problem: not only do you have to worry about your external validity, you also have to worry about your internal validity too. To see why this is the case, let’s consider a very dumb study in which I want to see if insulting people makes them act in a more obedient way. Why anyone would actually want to study that I don’t know, but let’s suppose I really, deeply cared about this. So, I design my experiment with two conditions. In the “treatment” condition, the experimenter insults the participant and then gives them a questionnaire designed to measure obedience. In the “control” condition, the experimenter engages in a bit of pointless chitchat and then gives them the questionnaire. Leaving aside the questionable scientific merits and dubious ethics of such a study, let’s have a think about what might go wrong here. As a general rule, when someone insults me to my face, I tend to get much less co-operative. So, there’s a pretty good chance that a lot more people are going to drop out of the treatment condition than the control condition. And this drop out isn’t going to be ran-

dom. The people most likely to drop out would probably be the people who don't care all that much about the importance of obediently sitting through the experiment. Since the most bloody minded and disobedient people all left the treatment group but not the control group, we've introduced a confound: the people who actually took the questionnaire in the treatment group were *already* more likely to be dutiful and obedient than the people in the control group. In short, in this study insulting people doesn't make them more obedient: it makes the more disobedient people leave the experiment! The internal validity of this experiment is completely shot.

###Non-response bias

Non-response bias is closely related to selection bias, and to differential attrition. The simplest version of the problem goes like this. You mail out a survey to 1000 people, and only 300 of them reply. The 300 people who replied are almost certainly not a random subsample. People who respond to surveys are systematically different to people who don't. This introduces a problem when trying to generalise from those 300 people who replied, to the population at large; since you now have a very non-random sample. The issue of non-response bias is more general than this, though. Among the (say) 300 people that did respond to the survey, you might find that not everyone answers every question. If (say) 80 people chose not to answer one of your questions, does this introduce problems? As always, the answer is maybe. If the question that wasn't answered was on the last page of the questionnaire, and those 80 surveys were returned with the last page missing, there's a good chance that the missing data isn't a big deal: probably the pages just fell off. However, if the question that 80 people didn't answer was the most confrontational or invasive personal question in the questionnaire, then almost certainly you've got a problem. In essence, what you're dealing with here is what's called the problem of *missing data*. If the data that is missing was "lost" randomly, then it's not a big problem. If it's missing systematically, then it can be a big problem.

###Regression to the mean

Regression to the mean is a curious variation on selection bias. It refers to any situation where you select data based on an extreme value on some measure. Because the measure has natural variation, it almost certainly means that when you take a subsequent measurement, that later measurement will be less extreme than the first one, purely by chance.

Here's an example. Suppose I'm interested in whether a psychology education has an adverse effect on very smart kids. To do this, I find the 20 psych I students with the best high school grades and look at how well they're doing at university. It turns out that they're doing a lot better than average, but they're not topping the class at university, even though they did top their classes at high school. What's going on? The natural first thought is that this must mean that the psychology classes must be having an adverse effect on those students. However, while that might very well be the explanation, it's more

likely that what you're seeing is an example of "regression to the mean". To see how it works, let's take a moment to think about what is required to get the best mark in a class, regardless of whether that class be at high school or at university. When you've got a big class, there are going to be *lots* of very smart people enrolled. To get the best mark you have to be very smart, work very hard, and be a bit lucky. The exam has to ask just the right questions for your idiosyncratic skills, and you have to not make any dumb mistakes (we all do that sometimes) when answering them. And that's the thing: intelligence and hard work are transferrable from one class to the next. Luck isn't. The people who got lucky in high school won't be the same as the people who get lucky at university. That's the very definition of "luck". The consequence of this is that, when you select people at the very extreme values of one measurement (the top 20 students), you're selecting for hard work, skill and luck. But because the luck doesn't transfer to the second measurement (only the skill and work), these people will all be expected to drop a little bit when you measure them a second time (at university). So their scores fall back a little bit, back towards everyone else. This is regression to the mean.

Regression to the mean is surprisingly common. For instance, if two very tall people have kids, their children will tend to be taller than average, but not as tall as the parents. The reverse happens with very short parents: two very short parents will tend to have short children, but nevertheless those kids will tend to be taller than the parents. It can also be extremely subtle. For instance, there have been studies done that suggested that people learn better from negative feedback than from positive feedback. However, the way that people tried to show this was to give people positive reinforcement whenever they did good, and negative reinforcement when they did bad. And what you see is that after the positive reinforcement, people tended to do worse; but after the negative reinforcement they tended to do better. But! Notice that there's a selection bias here: when people do very well, you're selecting for "high" values, and so you should *expect* (because of regression to the mean) that performance on the next trial should be worse, regardless of whether reinforcement is given. Similarly, after a bad trial, people will tend to improve all on their own. The apparent superiority of negative feedback is an artifact caused by regression to the mean (see Kahneman and Tversky, 1973, for discussion).

Experimenter bias

Experimenter bias can come in multiple forms. The basic idea is that the experimenter, despite the best of intentions, can accidentally end up influencing the results of the experiment by subtly communicating the "right answer" or the "desired behaviour" to the participants. Typically, this occurs because the experimenter has special knowledge that the participant does not – either the right answer to the questions being asked, or knowledge of the expected pattern of performance for the condition that the participant is in, and so on. The classic example of this happening is the case study of "Clever Hans", which dates back to 1907 (Pfungst, 1911; Hothersall, 2004). Clever Hans was a horse

that apparently was able to read and count, and perform other human like feats of intelligence. After Clever Hans became famous, psychologists started examining his behaviour more closely. It turned out that – not surprisingly – Hans didn’t know how to do maths. Rather, Hans was responding to the human observers around him. Because they did know how to count, and the horse had learned to change its behaviour when people changed theirs.

The general solution to the problem of experimenter bias is to engage in double blind studies, where neither the experimenter nor the participant knows which condition the participant is in, or knows what the desired behaviour is. This provides a very good solution to the problem, but it’s important to recognise that it’s not quite ideal, and hard to pull off perfectly. For instance, the obvious way that I could try to construct a double blind study is to have one of my Ph.D. students (one who doesn’t know anything about the experiment) run the study. That feels like it should be enough. The only person (me) who knows all the details (e.g., correct answers to the questions, assignments of participants to conditions) has no interaction with the participants, and the person who does all the talking to people (the Ph.D. student) doesn’t know anything. Except, that last part is very unlikely to be true. In order for the Ph.D. student to run the study effectively, they need to have been briefed by me, the researcher. And, as it happens, the Ph.D. student also knows me, and knows a bit about my general beliefs about people and psychology (e.g., I tend to think humans are much smarter than psychologists give them credit for). As a result of all this, it’s almost impossible for the experimenter to avoid knowing a little bit about what expectations I have. And even a little bit of knowledge can have an effect: suppose the experimenter accidentally conveys the fact that the participants are expected to do well in this task. Well, there’s a thing called the “Pygmalion effect”: if you expect great things of people, they’ll rise to the occasion; but if you expect them to fail, they’ll do that too. In other words, the expectations become a self-fulfilling prophesy.

Demand effects and reactivity

When talking about experimenter bias, the worry is that the experimenter’s knowledge or desires for the experiment are communicated to the participants, and that these effect people’s behaviour (Rosenthal, 1966). However, even if you manage to stop this from happening, it’s almost impossible to stop people from knowing that they’re part of a psychological study. And the mere fact of knowing that someone is watching/studying you can have a pretty big effect on behaviour. This is generally referred to as *reactivity* or *demand effects*. The basic idea is captured by the Hawthorne effect: people alter their performance because of the attention that the study focuses on them. The effect takes its name from a the “Hawthorne Works” factory outside of Chicago (see Adair, 1984). A study done in the 1920s looking at the effects of lighting on worker productivity at the factory turned out to be an effect of the fact that the workers knew they were being studied, rather than the lighting.

To get a bit more specific about some of the ways in which the mere fact of

being in a study can change how people behave, it helps to think like a social psychologist and look at some of the *roles* that people might adopt during an experiment, but might not adopt if the corresponding events were occurring in the real world:

- The *good participant* tries to be too helpful to the researcher: he or she seeks to figure out the experimenter's hypotheses and confirm them.
- The *negative participant* does the exact opposite of the good participant: he or she seeks to break or destroy the study or the hypothesis in some way.
- The *faithful participant* is unnaturally obedient: he or she seeks to follow instructions perfectly, regardless of what might have happened in a more realistic setting.
- The *apprehensive participant* gets nervous about being tested or studied, so much so that his or her behaviour becomes highly unnatural, or overly socially desirable.

Placebo effects

The *placebo effect* is a specific type of demand effect that we worry a lot about. It refers to the situation where the mere fact of being treated causes an improvement in outcomes. The classic example comes from clinical trials: if you give people a completely chemically inert drug and tell them that it's a cure for a disease, they will tend to get better faster than people who aren't treated at all. In other words, it is people's belief that they are being treated that causes the improved outcomes, not the drug.

Situation, measurement and subpopulation effects

In some respects, these terms are a catch-all term for "all other threats to external validity". They refer to the fact that the choice of subpopulation from which you draw your participants, the location, timing and manner in which you run your study (including who collects the data) and the tools that you use to make your measurements might all be influencing the results. Specifically, the worry is that these things might be influencing the results in such a way that the results won't generalise to a wider array of people, places and measures.

Fraud, deception and self-deception

It is difficult to get a man to understand something, when his salary depends on his not understanding it.

– Upton Sinclair

One final thing that I feel like I should mention. While reading what the textbooks often have to say about assessing the validity of the study, I couldn't help but notice that they seem to make the assumption that the researcher is honest. I find this hilarious. While the vast majority of scientists are honest, in my experience at least, some are not.⁶ Not only that, as I mentioned earlier,

⁶Some people might argue that if you're not honest then you're not a real scientist. Which

scientists are not immune to belief bias – it’s easy for a researcher to end up deceiving themselves into believing the wrong thing, and this can lead them to conduct subtly flawed research, and then hide those flaws when they write it up. So you need to consider not only the (probably unlikely) possibility of outright fraud, but also the (probably quite common) possibility that the research is unintentionally “slanted”. I opened a few standard textbooks and didn’t find much of a discussion of this problem, so here’s my own attempt to list a few ways in which these issues can arise are:

- ***Data fabrication.*** Sometimes, people just make up the data. This is occasionally done with “good” intentions. For instance, the researcher believes that the fabricated data do reflect the truth, and may actually reflect “slightly cleaned up” versions of actual data. On other occasions, the fraud is deliberate and malicious. Some high-profile examples where data fabrication has been alleged or shown include Cyril Burt (a psychologist who is thought to have fabricated some of his data), Andrew Wakefield (who has been accused of fabricating his data connecting the MMR vaccine to autism) and Hwang Woo-suk (who falsified a lot of his data on stem cell research).
- ***Hoaxes.*** Hoaxes share a lot of similarities with data fabrication, but they differ in the intended purpose. A hoax is often a joke, and many of them are intended to be (eventually) discovered. Often, the point of a hoax is to discredit someone or some field. There’s quite a few well known scientific hoaxes that have occurred over the years (e.g., Piltdown man) some of were deliberate attempts to discredit particular fields of research (e.g., the Sokal affair).
- ***Data misrepresentation.*** While fraud gets most of the headlines, it’s much more common in my experience to see data being misrepresented. When I say this, I’m not referring to newspapers getting it wrong (which they do, almost always). I’m referring to the fact that often, the data don’t actually say what the researchers think they say. My guess is that, almost always, this isn’t the result of deliberate dishonesty, it’s due to a lack of sophistication in the data analyses. For instance, think back to the example of Simpson’s paradox that I discussed in the beginning of these notes. It’s very common to see people present “aggregated” data of some kind; and sometimes, when you dig deeper and find the raw data yourself, you find that the aggregated data tell a different story to the disaggregated data. Alternatively, you might find that some aspect of the data is being hidden, because it tells an inconvenient story (e.g., the researcher might choose not to refer to a particular variable). There’s a lot of variants on this; many of which are very hard to detect.

does have some truth to it I guess, but that’s disingenuous (google the “No true Scotsman” fallacy). The fact is that there are lots of people who are employed ostensibly as scientists, and whose work has all of the trappings of science, but who are outright fraudulent. Pretending that they don’t exist by saying that they’re not scientists is just childish.

- **Study “misdesign”.** Okay, this one is subtle. Basically, the issue here is that a researcher designs a study that has built-in flaws, and those flaws are never reported in the paper. The data that are reported are completely real, and are correctly analysed, but they are produced by a study that is actually quite wrongly put together. The researcher really wants to find a particular effect, and so the study is set up in such a way as to make it “easy” to (artificially) observe that effect. One sneaky way to do this – in case you’re feeling like dabbling in a bit of fraud yourself – is to design an experiment in which it’s obvious to the participants what they’re “supposed” to be doing, and then let reactivity work its magic for you. If you want, you can add all the trappings of double blind experimentation etc. It won’t make a difference, since the study materials themselves are subtly telling people what you want them to do. When you write up the results, the fraud won’t be obvious to the reader: what’s obvious to the participant when they’re in the experimental context isn’t always obvious to the person reading the paper. Of course, the way I’ve described this makes it sound like it’s always fraud: probably there are cases where this is done deliberately, but in my experience the bigger concern has been with unintentional misdesign. The researcher *believes* ... and so the study just happens to end up with a built in flaw, and that flaw then magically erases itself when the study is written up for publication.
- **Data mining & post hoc hypothesising.** Another way in which the authors of a study can more or less lie about what they found is by engaging in what’s referred to as “data mining”. As we’ll discuss later in the class, if you keep trying to analyse your data in lots of different ways, you’ll eventually find something that “looks” like a real effect but isn’t. This is referred to as “data mining”. It used to be quite rare because data analysis used to take weeks, but now that everyone has very powerful statistical software on their computers, it’s becoming very common. Data mining per se isn’t “wrong”, but the more that you do it, the bigger the risk you’re taking. The thing that is wrong, and I suspect is very common, is *unacknowledged* data mining. That is, the researcher run every possible analysis known to humanity, finds the one that works, and then pretends that this was the only analysis that they ever conducted. Worse yet, they often “invent” a hypothesis after looking at the data, to cover up the data mining. To be clear: it’s not wrong to change your beliefs after looking at the data, and to reanalyse your data using your new “post hoc” hypotheses. What is wrong (and, I suspect, common) is failing to acknowledge that you’ve done so. If you acknowledge that you did it, then other researchers are able to take your behaviour into account. If you don’t, then they can’t. And that makes your behaviour deceptive. Bad!
- **Publication bias & self-censoring.** Finally, a pervasive bias is “non-reporting” of negative results. This is almost impossible to prevent. Journals don’t publish every article that is submitted to them: they prefer to publish articles that find “something”. So, if 20 people run an experiment looking at whether reading *Finnegans Wake* causes insanity in humans,

and 19 of them find that it doesn't, which one do you think is going to get published? Obviously, it's the one study that did find that *Finnegans Wake* causes insanity⁷. This is an example of a *publication bias*: since no-one ever published the 19 studies that didn't find an effect, a naive reader would never know that they existed. Worse yet, most researchers "internalise" this bias, and end up *self-censoring* their research. Knowing that negative results aren't going to be accepted for publication, they never even try to report them. As a friend of mine says "for every experiment that you get published, you also have 10 failures". And she's right. The catch is, while some (maybe most) of those studies are failures for boring reasons (e.g. you stuffed something up) others might be genuine "null" results that you ought to acknowledge when you write up the "good" experiment. And telling which is which is often hard to do. A good place to start is a paper by Ioannidis (2005) with the depressing title "Why most published research findings are false". I'd also suggest taking a look at work by Kühberger et al. (2014) presenting statistical evidence that this actually happens in psychology.

There's probably a lot more issues like this to think about, but that'll do to start with. What I really want to point out is the blindingly obvious truth that real world science is conducted by actual humans, and only the most gullible of people automatically assumes that everyone else is honest and impartial. Actual scientists aren't usually *that* naive, but for some reason the world likes to pretend that we are, and the textbooks we usually write seem to reinforce that stereotype.

##Summary

This chapter isn't really meant to provide a comprehensive discussion of psychological research methods: it would require another volume just as long as this one to do justice to the topic. However, in real life statistics and study design are tightly intertwined, so it's very handy to discuss some of the key topics. In this chapter, I've briefly discussed the following topics:

- Introduction to psychological measurement. What does it mean to operationalise a theoretical construct? What does it mean to have variables and take measurements?
- Scales of measurement and types of variables. Remember that there are *two* different distinctions here: there's the difference between discrete and continuous data, and there's the difference between the four different scale types (nominal, ordinal, interval and ratio).
- Reliability of a measurement. If I measure the "same" thing twice, should I expect to see the same result? Only if my measure is reliable. But what does it mean to talk about doing the "same" thing? Well, that's why we have different types of reliability. Make sure you remember what they are.
- Terminology: predictors and outcomes. What roles do variables play in

⁷Clearly, the real effect is that only insane people would even try to read *Finnegans Wake*.

an analysis? Can you remember the difference between predictors and outcomes? Dependent and independent variables? Etc.

- Experimental and non-experimental research designs. What makes an experiment an experiment? Is it a nice white lab coat, or does it have something to do with researcher control over variables?
- Validity and its threats. Does your study measure what you want it to? How might things go wrong? And is it my imagination, or was that a very long list of possible ways in which things can go wrong?

All this should make clear to you that study design is a critical part of research methodology. I built this chapter from the classic little book by Campbell and Stanley (1963), but there are of course a large number of textbooks out there on research design. Spend a few minutes with your favourite search engine and you'll find dozens.

Bibliography

- Adair, G. (1984). The hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, 69:334–345.
- Campbell, D. T. and Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin, Boston, MA.
- Hothersall, D. (2004). *History of Psychology*. McGraw-Hill.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, 2(8):697–701.
- Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80:237–251.
- Kühberger, A., Fritz, A., and Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *Public Library of Science One*, 9:1–8.
- Pfungst, O. (1911). *Clever Hans (The horse of Mr. von Osten): A contribution to experimental animal and human psychology*. Henry Holt.
- Rosenthal, R. (1966). *Experimenter effects in behavioral research*. Appleton.