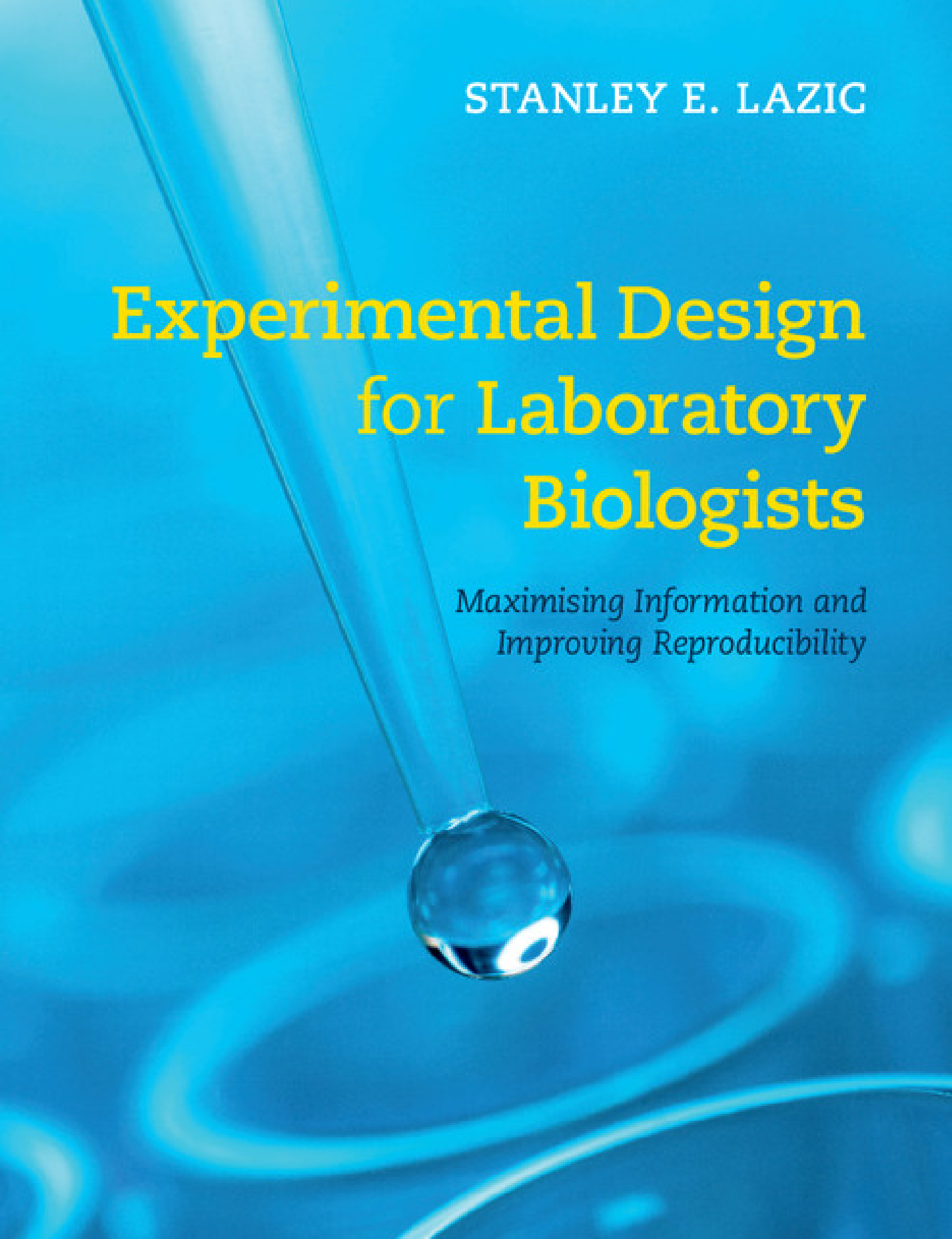
STANLEY E. LAZIC
Experimental Design for Laboratory Biologists
Maximising Information and Improving Reproducibility

Experimental Design for Laboratory Biologists

Maximising Information and Improving Reproducibility

Specifically intended for lab-based biomedical researchers, this practical guide shows how to design experiments that are reproducible, with low bias, high precision, and results that are widely applicable. With specific examples from research, using both cell cultures and model organisms, it explores key ideas in experimental design, assesses common designs, and shows how to plan a successful experiment. It demonstrates how to control biological and technical factors that can introduce bias or add noise, and covers rarely discussed topics such as graphical data exploration, choosing outcome variables, data quality control checks, and data preprocessing. It also shows how to use R for analysis, and is designed for those with no prior experience. An accompanying website (https://stanlazic.github.io/EDLB.html) includes all R code, data sets, and the labstats R package.

This is an ideal guide for anyone conducting lab-based biological research, from students to principal investigators working either in academia or industry.

Stanley E. Lazic holds a PhD in neuroscience and a Masters in computational biology from the University of Cambridge and has conducted research at Oxford, Cambridge, and Harvard. He has written several papers on reproducible research and on the design and analysis of biological experiments and has published in *Science* and *Nature*. He is currently a Team Leader in Quantitative Biology (Statistics) at AstraZeneca.

Experimental Design for Laboratory Biologists

Maximising Information and Improving Reproducibility

STANLEY E. LAZIC

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107074293

First published 2016

Printed in the United Kingdom by TJ International Ltd., Padstow, Cornwall

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

ISBN 978-1-107-07429-3 Hardback
ISBN 978-1-107-42488-3 Paperback

Additional resources for this publication at https://stanlazic.github.io/EDLB.html

To my teachers and mentors

Contents

Preface

Everything of importance has been said before by somebody who did not discover it.
Alfred North Whitehead

Everything that needs to be said has already been said. But since no one was listening, everything must be said again.
André Gide

True to the above quotes, most of this book's contents have appeared in print before, but often where biologists are unlikely to look – statistics journals and books, and methods papers in other fields. My task is to translate ideas known to statisticians into the language of experimental biology.[1] With a background in both biology (BSc, PhD, postdoc) and data analysis (MPhil in Computational Biology and over seven years working as a preclinical statistician in the pharmaceutical industry), hopefully I am fluent enough in both languages to perform a successful translation.[2]

The contents of this book have little overlap with other statistics-for-biologists books because they mostly focus on statistical analysis. Analysis is but one step of the scientific workflow (Figure 0.1), and before you can analyse data you need to do an experiment. This requires planning, good execution, and quality control checks. These critical topics are rarely taught to biologists, who are expected to learn them on their own. The consequence of this approach is predictable; some biologists obtain the necessary skills, but many do not. This book focuses on the first three steps of the scientific workflow, and data analysis is briefly discussed in Chapter 4.

This book was written to improve the quality of research conducted in academic, government, and industrial labs and institutions. Scientists and funders now recognise that bias and irreproducibility are undermining preclinical biomedical research [2, 5, 28, 30, 42, 80, 83, 84, 123, 172, 240, 251, 305, 316, 342]. There are many reasons why experiments cannot be reproduced (discussed in Chapter 1) and this book focuses on the role that experimental design and data analysis have on making results reproducible.

[1] The term *biology* refers to laboratory-based experimental biology throughout. 'Field biologists' also conduct experiments, and most statistics-for-biologists books target this audience.

[2] There are some novel ideas here, such as the distinction between front-aligned and end-aligned designs (Section 2.11) and the distinction between biological, experimental, and observational units, to replace the biological versus technical replicate distinction (all of Chapter 3).

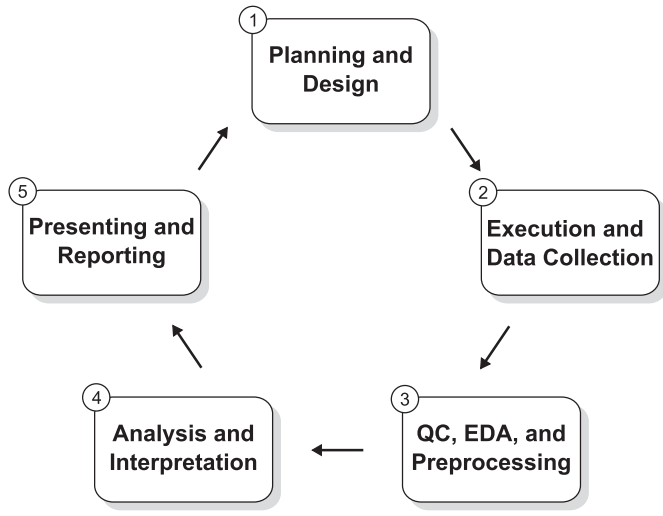

Fig. 0.1 The scientific workflow. This book focuses on steps 1–3. QC = quality control; EDA = exploratory data analysis.

Prerequisites

This book is for experimental biologists, at any level, conducting basic research or with an applied, clinical, or translational focus. Knowledge covered in an introductory statistics-for-biologists course is assumed, and concepts like the standard deviation and common statistical tests such as the *t*-test, analysis of variance (ANOVA), regression, and correlation should be familiar. It is fine if some time has passed since you formally covered these topics. Mathematical proofs are not included and equations are kept to a minimum, but given the subject, are unavoidable. The emphasis is on the ideas, concepts, and principles, and how to implement them. Hand calculations are unnecessary because statistical software is available.

Quantitative researchers who analyse biological data such as statisticians, bioinformaticians, and computational biologists might also find this book useful. Topics of interest include sources of heterogeneity and confounding in biological experiments (Section 2.12), quality control checks for biological data (Section 6.1), and understanding which types of replication address biologically interesting questions (Chapter 3).

The freely available R statistics language is used for data analysis and graphs.[3] Prior knowledge is useful, but not required. The Appendix gives a brief introduction to R and the examples in the main text assume familiarity with this material. The topics however can be followed without learning or using R. The data sets can be found in the `labstats` package on CRAN[4] and R code can be downloaded from GitHub.[5]

[3] Available at `www.r-project.com`
[4] `https://cran.r-project.org/web/packages/labstats/`
[5] `https://stanlazic.github.io/EDLB.html`

The key prerequisite to derive maximum value from this book is experience conducting biological experiments and analysing the subsequent data – and the more experience the better!

How to read this book

Chapters 1–5 should be read in order as later material depends on earlier ideas, but Chapter 6 on Exploratory Data Analysis can be read at any time. Chapters 1–3 contain no R code, but for Chapters 4–6 sitting in front of a computer and running the code will reinforce the ideas.

Ideas or concepts discussed in detail later in the book will inevitably have to be mentioned earlier. To avoid excessive cross-referencing, the glossary lists the page where the main discussion of the entry is located (if there is one). For example, the term *experimental unit* is mentioned for the first time in this preface, but is discussed extensively in Section 3.2. The glossary entry for this term provides a short definition and indicates that further information can be found on page 96.

Typographical conventions

`Constant width font` is used for R code, R output, and when referring to R functions or objects. Lines of code entered by the user start with '>' or '+'. These symbols do not need to be entered, only the code that follows them. A sign like the one in the margin draws attention to a warning, a key point, a subtlety with R, or a concept that is often misunderstood.

Acknowledgements

This book has benefited greatly from comments by Maarten van Dijk, Irmgard Amrein, and especially Lutz Slomianka. Pierre Farmer and Miguel Camargo also provided constructive feedback on earlier drafts. My wife, Brynn, has read every word in this book, which is beyond the call of duty, and her comments have improved it immensely. I also thank her for her support, well, at least until page 305, at which point she declared, 'You should stop now; no one wants to read that much about statistics.' I didn't always follow everyone's good advice, but I am grateful for their input.

Katrina Halliday and Jade Scard at Cambridge University Press were a pleasure to work with and made the whole process easy and enjoyable. I also thank Judith Shaw for her expert copy-editing. Finally, I would like to thank the developers and contributors of the free software R, Emacs, LaTeX, JabRef, knitr, and Inkscape, which I used to write this book.

S.E. Lazic
Cambridge, 2016

Abbreviations

AIPE	Accuracy in parameter estimation
ALS	Amyotrophic lateral scterosis
ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
AUC	Area under the curve
BMI	Body mass index
BU	Biological unit
CCC	Concordance correlation coefficient
CCLE	Cancer Cell Line Encyclopedia
CI	Confidence (frequentist) or Credible (Bayesian) interval
CRAN	Comprehensive R archive network
CRD	Completely randomised design
CSF	Cerebrospinal fluid
CSR	Complete spatial randomness
CV	Coefficient of variation
DAMP	Damage-associated molecular pattern
df	Degrees of freedom
DoE	Design of experiments
DS	Diallyl sulfide
ED50	Median (half) effective dose
EDA	Exploratory data analysis
ES	Effect size
ESS	Emacs Speaks Statistics
EU	Experimental unit
FORE-SCI	Facilities of Research Excellence – Spinal Cord Injury
FOV	Field of view
GI	Gastrointestinal
GLM	Generalised linear model
GUI	Graphical user interface
Gst	Glutathione-S-transferase
HARKing	Hypothesising after the results are known
HSD	Honestly significant difference
ICC	Intraclass correlation coefficient
i.p.	Intraperitoneally
IQR	Interquartile range

KO	Knock out
LME	Linear mixed-effects model
LOD	Limit of detection
LSD	Least significant difference
MAD	Median absolute deviation
MAR	Missing at random
MCAR	Missing completely at random
MED	Minimum effective dose
MNAR	Missing not a random
NGS	Next generation sequencing
NHST	Null hypothesis significance testing
NIH	National Institutes of Health (USA)
NINDS	American National Institute of Neurological Disorders and Stroke
OU	Observational unit
PCA	Principal components analysis
PI	Principal investigator
PK	Pharmokinetic
QC	Quality control
QRP	Questionable research practice
qPCR	Quantitative polymerase chain reaction
RE	Relative efficiency
RIN	RNA integrity number
RM-ANOVA	Repeated measures analysis of variance
SAP	Statistical analysis plan
SD	Standard deviation
SEM	Standard error of the mean
siRNA	small interfering RNA
SNP	Single nucleotide polymorphism
SOD1	Superoxide dismutase 1 (gene)
SS	Sum of squares
RSS	Residual sum of squares
SUTVA	Stable unit-treatment value assumption
TSS	Total sum of squares
VPA	Valproic acid
WT	Wild type

1 Introduction

To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.
Sir Ronald A. Fisher, FRS

Many experiments fail because the data collectors have not been properly trained and many statisticians have their own horror stories to illustrate this.
John A. Nelder, FRS

Modeling is sometimes regarded as primarily a task for subject matter specialists, but in most fields requisite knowledge and understanding of statistics remains thinly spread.
Arthur P. Dempster

This chapter begins by defining reproducibility and discussing non-statistical – mainly psychological – sources of experimental bias. The next section assesses the quality of the published literature and discusses statistical sources of bias. The discussion will be familiar if you have been following the 'reproducibility crisis' over the past several years. The above topics are included to stimulate reflection about your own research practices and to bring together ideas that have been discussed in separate disciplines. The chapter ends with a refresher on statistical inference and a discussion on statistics software.

1.1 What is reproducibility?

An experiment is reproducible when subsequent experiments, by the same or different scientists, confirm the results. The terms *repeatability* and *replicability* are sometimes used interchangeably or with related meanings, but we will use reproducibility as an all-encompassing term. Reproducibility can occur at several levels.[1]

Analytical: Analytical or computational reproducibility refers to obtaining the same results using the original data and a description of the analysis. This is a minimum standard but is impossible to achieve when the data are unavailable. Even if the data are provided in the supplementary material or in public databases, reproducing the results may be hard if the description of the analysis is incomplete [173]. A minimum

[1] Adapted from a report on reproducibiliy by the American Society for Cell Biology: http://www.ascb.org/reproducibility

requirement for analytical reproducibility is to provide the data underlying the results and the scripts that produced them. This is simple when using R because the code can be integrated into documents such as reports and publications. For example, large portions of this book have embedded R code, which is evaluated, and then the outputs are inserted into the text document. The `knitr` and `rmarkdown` R packages make this process straightforward [402].

Direct: Direct reproducibility refers to obtaining the same results using the same experimental conditions, materials, and methods as the original experiment. The aim is to make the second experiment as similar as possible to the original, which requires an adequate description of how the original experiment was conducted. Direct replication is the focus of this book, but it may not be immediately clear how better experimental designs can improve direct reproducibility. The brief answer is that a well-designed experiment (1) can isolate the effects of interest from other factors that may influence the outcome, (2) replicates the right aspect of the experiment, and (3) can generalise the results to other times, places, conditions, and samples.

Systematic: Systematic reproducibility refers to obtaining the same results, but under different conditions; for example, using another cell line or mouse strain, or inhibiting a gene pharmacologically instead of genetically. Reasons for a lack of systematic reproducibility are harder to determine because the cell lines might be dissimilar, and what works in one will not work in another. This should not be taken as evidence of poor research practices, and one function of subsequent studies is to find the conditions under which an initial finding holds. Experimental design can help here too, as initial studies can be designed to address the question of generalisability early on.

Conceptual: Conceptual reproducibility refers to obtaining the same general results under diverse conditions, where the aim is to demonstrate the validity of a concept or a finding using another paradigm. The general concept or hypothesis might be 'stress inhibits memory formation', which could be tested in one experiment where people memorise word pairs with loud music playing and in another experiment where rats learn the location of food pellets after a corticosterone injection (a stress hormone). There are many valid reasons why some experiments support the hypothesis and others not – maybe corticosterone, while part of the stress response, is irrelevant for learning. Discrepancies between the results of such experiments do not necessarily indicate poor reproducibility.

A reproducible result was defined above as one that is confirmed by subsequent experiments, but what does *confirmed* mean? One idea is that if the original experiment has a p-value below 0.05, then the experiment is confirmed if the subsequent experiment also has a significant p-value. Although this criterion seems plausible, it has several problems. First, a study with a p-value of 0.03 would be considered irreproducible if the subsequent experiment had a p-value of 0.08. But for all practical purposes the studies may have the same effect sizes and their two confidence intervals (CIs) may overlap substantially. This relationship is shown in Figure 1.1 between the original experiment and the second experiment, New 1. A second problem is that this approach ignores the sample size and power of the experiments. Suppose that a power analysis was conducted based on the results of

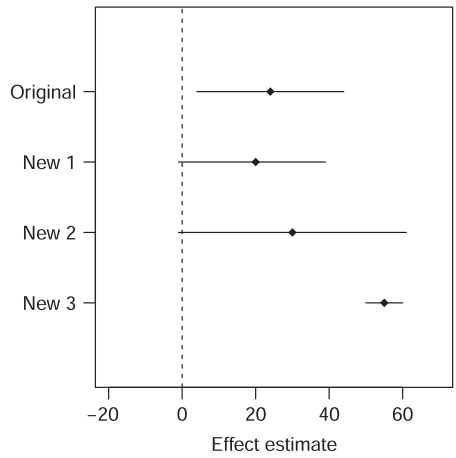

Fig. 1.1 Effect sizes and 95% CI for an original experiment and three follow-up experiments. Using significance as a criterion for reproducibility, only New 3 would be considered to reproduce the original finding, despite the different effect size.

the original experiment and the follow-up experiment uses a slightly smaller sample size and therefore the confidence intervals will be slightly wider, assuming everything else is constant (New 2 in Figure 1.1). Even though the effect size for New 2 is larger than the original, New 2 would not have reproduced the original findings by this criterion. A third problem is that a follow-up study may have a different effect size than the original but would be considered to have successfully reproduced the original if the *p*-value is significant. This situation is shown for experiment New 3, where the 95% CIs do not overlap with those of the original experiment. There is no agreed criteria for when one experiment can be said to reproduce another, but within a field, scientists 'know it when they see it'.

1.2 The psychology of scientific discovery

It is uncommon for a book on experimental design to discuss psychological aspects of research, but scientific investigations are not conducted in a vacuum; they take place in the context of previous research, are conducted by people who prefer certain outcomes over others, and are constrained by standards and conventions used by research groups and the wider scientific community. Expectations and desires of the researcher and external pressures to publish and to demonstrate creativity and innovation influence how data are analysed, interpreted, and reported. This needs to be acknowledged and discussed because improving reproducibility and 'making more published research true' [172] – one of the aims of this book – cannot be achieved by only improving scientists' maths skills.

Some of the topics discussed below fall within the 'heuristics and biases' field of psychological research. Cognitive biases or cognitive illusions are deviations from true or optimal answers or responses when making estimations, inferences, decisions, conclusions, or judgements [312]. They are cognitive in that they result from perceptual or cognitive

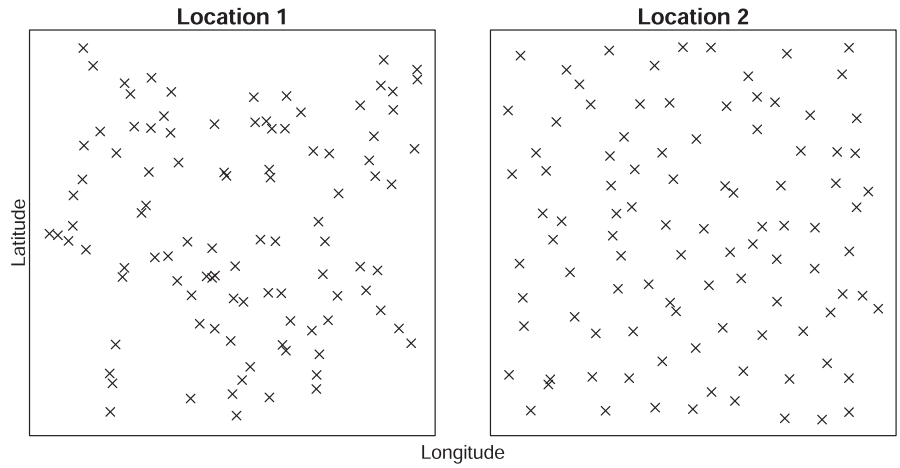

Fig. 1.2 Positions of bombs dropped for two geographic locations. Which location represents the uniform random bombing strategy?

processes instead of, for example, an uncalibrated measuring device. They are also systematic, meaning that the deviations tend to be in a certain direction. They are also hard to avoid. Cognitive biases can influence the design, analysis, interpretation, and reporting of biological experiments and are therefore relevant for scientific investigations. Several such biases are described below, with an emphasis on how they apply to experimentation and statistical inference. Methods for avoiding them are also suggested.

1.2.1 Seeing patterns in randomness

People often see patterns where none exist, including clusters, associations between variables, and sequences of similar values. A fictitious example is given in Figure 1.2. The positions (latitude and longitude) of 100 bombs dropped during a World War II bombing campaign are shown for two different geographic locations. The General wants to know if the enemy is dropping bombs at random, or if they are targeting certain positions more heavily, and he asks you to investigate. Intelligence from the front line indicates that *if* the enemy is using a random strategy, they will randomly sample a pair of latitude and longitude coordinates with equal probability anywhere within the bombing region – known as a uniform random bombing strategy.[2] Was a uniform random bombing strategy used at either of the locations in Figure 1.2? If so, which one? Furthermore, is there evidence at either location for certain positions to be bombed more heavily while others are avoided, possibly reflecting the strategic importance of the positions?

Many would say that the distribution of points at Location 2 represents the strategy of randomly picking a latitude and longitude from a uniform distribution. The positions for

[2] The name arises from sampling latitude and longitude positions from a uniform distribution, where all values between an upper and lower boundary have an equal probability of being chosen.

Location 2 were instead generated by selecting a 10-by-10 grid of equally spaced positions, and then adding some noise to these values. This makes the bomb positions evenly spaced. The random uniform strategy is only used at Location 1. This appears counter-intuitive because there are large regions with no bombs, while other regions have a denser clustering. Such clustering and empty regions are to be expected under a uniform random strategy. Intuition about what randomness looks like does not come easily or naturally.

1.2.2 Not wanting to miss anything

Potentially meaningful patterns like the above example can be formally tested with a statistical analyses, but it is important to avoid using the same data to first find an interesting pattern (such as the lower left empty region of Location 1 in Figure 1.2), and then to statistically test for this pattern. For example, we might try to calculate the probability of seeing no bombs dropped in an area the size of the empty lower left region. Random data – especially if there is a lot of it – will have local regularities and patterns. Picking out one such pattern that catches our attention and then performing a statistical test has implicitly performed many informal tests, in that all of the patterns that *could have been* interesting were examined and discarded without a formal test. For example, it does not appear that more bombs were dropped at higher latitudes compared with lower latitudes (comparing the top versus the bottom half of Location 1). If such a pattern did appear to exist, then we would test that instead. The key principle is: *if a hypothesis is derived from the data, then the ability of the data to support that hypothesis is diminished.* The ability of the data to support a hypothesis can also be compromised by what others do. For example, a PhD student is the first person to analyse a data set and explores it thoroughly. He finds a relationship but is unsure of the appropriate statistical test and so brings it to the principal investigator's (PI) attention. The PI then conducts only one analysis and feels confident that the *p*-value is valid, because she is unaware of how the data were used to discover this relationship.

Even when a visual-driven inspection of the data is not so pronounced, people want to make the most of the data and to avoid missing anything interesting. This desire is likely greater when the primary result is not significant and then we have to see 'what else we can get out of the data'. One might begin to look for correlations between variables, then again after normalising or correcting for other variables. Then checking for differences between sexes, or the old versus the young, or the less severely affected compared to the most affected, and so on until there are enough interesting findings to report. On the one hand, it seems foolish not to thoroughly examine the data, given all of the work that went into generating it. On the other hand, such a search process can generate many false positives.

There are two approaches to limit the number of false positive results that arise from data-driven discoveries. The first is to divide the analyses into confirmatory and exploratory parts. The confirmatory analysis specifies everything in advance (before seeing the data), including the hypothesis to be tested, the main outcome variable, and the analysis that will be used. The subsequent exploratory analysis allows for greater flexibility to find other relationships of interest, but with the knowledge that the findings carry less weight and are

less convincing because they were not predicted in advance, even if attempts have been made to correct for multiple testing. The second approach is to validate the findings, either by conducting a subsequent experiment, or by dividing the data into two parts. Once the experiment is complete, but before any analysis, about 20–30% of the data are removed and locked away. The remaining data are used to find interesting relationships. Once the analysis is complete, the data that were locked away are used to confirm the findings. This is a common approach in the data mining, machine learning, and predictive modelling fields, but it does require enough samples to split into two sets.

People differ in how easily they detect signals in pure noise, find patterns in randomness, or meaning in the coincidental. A sign of 'inferential maturity' is to know where you lie on the spectrum. If you find anything vaguely resembling an association or effect interesting and tend to believe that it is 'real', then pay attention to controlling false positives. If instead you are sceptical and find only large associations or effects convincing, then you risk not further exploring small but true findings.

1.2.3 Psychological cliff at $p = 0.05$

One criticism of p-values is that they encourage dichotomous thinking – the effect or relationship is either significant or it is not – even though evidence is continuous.[3] In the 1960s, Rosenthal and Gaito showed that such a psychological effect exists. They asked psychology researchers and graduate students to rate their degree of belief or confidence in a research hypothesis with p-values ranging from 0.001 to 0.90. They found a 'psychological cliff' at $p = 0.05$ – an abrupt jump in confidence just below 0.05 [331, 332]. A replication experiment by Poitevineau and Lecoutre provided a more nuanced view [313]. They found a strong cliff effect, but only in a subset of subjects; others had a linear or exponentially decreasing confidence as the p-values increased (Figure 1.3).

The cliff effect likely contributes to another misinterpretation of statistical results, which Gelman and Stern have phrased as 'The difference between "significant" and "not significant" is not itself statistically significant' [132]. They are referring to a situation where, for example, Group A is significantly different from the control group, Group B is not significantly different from the control group, and then an incorrect conclusion is made that Group A is significantly different from Group B. If differences between Group A and B are of interest, then they need to be compared directly against each other.

To the extent that a small p-value provides evidence for a research hypothesis, there is no sharp evidential distinction between 0.04 and 0.06. An obvious question is 'What is the correct relationship between a p-value and evidence for a hypothesis?' The short answer is that there is no correct relationship because a p-value says nothing about hypotheses and so the question makes no sense. If you are interested in evidence or the probability that a hypothesis is correct, then likelihood or Bayesian methods are required. These are beyond the scope of this book but good introductions can be found in references [50, 95, 139, 201–204, 269].

[3] Technically, a p-value does *not* provide evidence against a hypothesis. Informally, however, a smaller p-value suggests that an effect is present. The interpretation of a p-value is discussed in Section 1.4.

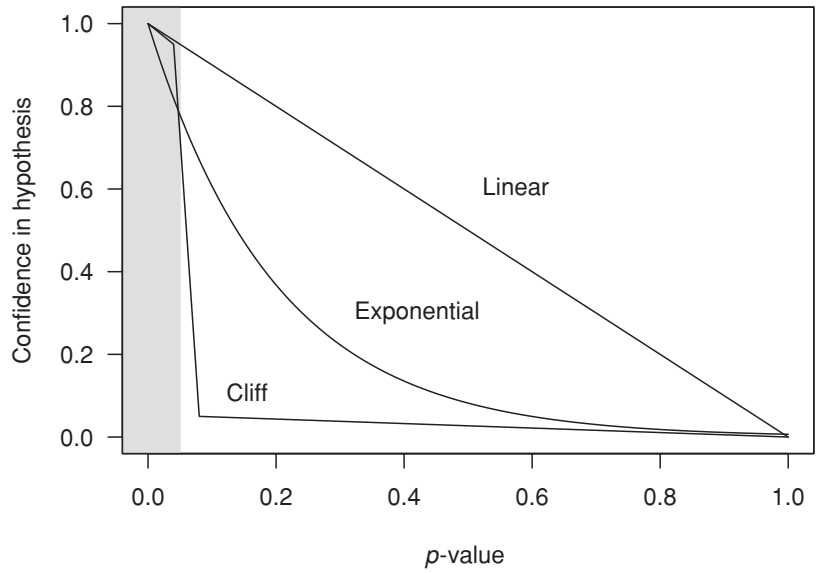

Fig. 1.3 Confidence from *p*-values. A schematic diagram from Poitevineau and Lecoutre [313]. The confidence in a hypothesis drops rapidly just past $p = 0.05$ for some people. Shaded area is $p < 0.05$.

The key point is that there is nothing special about 0.05, or values on either side, that indicates an abrupt change in what the data have to say about a hypothesis. Gelman and Loken raise two related points about interpreting statistical results [130]. The first is that effects cannot be divided into those that are 'real' and those that are 'not real', based on a *p*-value. The presence and magnitude of effects and associations are conditional on (1) the sample material used, (2) background variables and conditions (such as laboratory equipment and experimenter), (3) the experimental design (was a blocking factor incorporated), and (4) data preprocessing and the statistical analysis. Since the (true) magnitude of an effect or association is always conditional on so many factors it makes sense to consider how the effect or association varies across diverse situations. Under some conditions the effect may be smaller and the *p*-value above 0.05, and this does not imply a lack of reproducibility.

Their second point is that the statistical analysis does not determine whether an effect is 'real', just as microscopes do not determine whether bacteria are real, but both microscopes and statistics can help one see things that are not obvious with the naked eye. Effects are determined by the biological process under investigation, the experiment used to probe it, and the data derived from it. Occasionally, effects are so large and clear that no statistical analysis is necessary. When the experiment is more complex and the results less obvious, a statistical analysis only helps one to interpret what is already there. Interpretation of the results may differ depending on the analysis, but so too may a conclusion about a phenomenon depending on the microscope (e.g. light, confocal, or electron). Do not believe a result just because 'the statistics said. . . '.

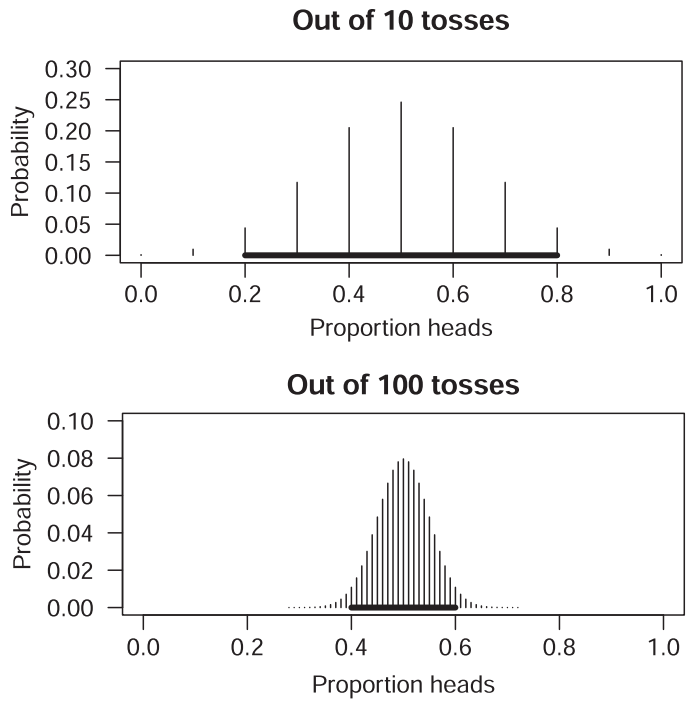

Fig. 1.4 Sampling variability and its dependence on sample size. When tossing a coin 10 times, we expect the proportion of heads to be 0.5, but it would not be unusual to obtain a proportion between 0.2 and 0.8 (thick black line on the *x*-axis). When tossing a coin 100 times (larger sample size), the range of likely values for the proportion is narrower, between 0.4 and 0.6.

1.2.4 Neglect of sampling variability

Sampling variability is the reason that the outcome of a random process differs from run to run. If a fair coin is tossed 10 times, we would expect, on average, five heads and five tails. On any given trial we may get more or less heads, but we would expect most tosses to have between two and eight heads. Rephrasing, the expected proportion of heads is 0.5, with the majority between 0.2 and 0.8. This is sampling variability: we do not always get five heads, and is illustrated in Figure 1.4. Furthermore, as the sample size increases, the variability in the outcome decreases. When the sample size is increased to 100 tosses, we still expect the proportion of heads to be 0.5, but now the majority will lie between 0.4 and 0.6 – a narrower interval. This is the dependence on sample size: the larger the sample size, the narrower the interval of values that we are likely to see. As the sample size increases, we converge to the true proportion of heads when tossing a fair coin. These simple ideas appear often and can lead to incorrect inferences and conclusions if not taken into account. Some examples are discussed below.

Wainer provides two real-world examples [384]. In his first example he shows a map of kidney cancer death rates in the USA, with the top (worst) 10% of counties highlighted.

Counties with high death rates tend to be rural and in the west or midwest, and one might speculate that people in these regions have unhealthy lifestyles or less access to high quality medical care, leading to higher death rates. The interesting twist is that if the counties with the *lowest* 10% of death rates are plotted, they also tend to be in the west or midwest. Counties with the highest and lowest rates are often right beside each other! What is going on? Counties in the west and midwest are sparsely populated, so the addition or subtraction of a few cases will have a larger influence on the cancer rate than in a larger population. Just as in the coin example above, with a small sample size the value of whatever is calculated will fluctuate more widely around the true value. Thus, sparsely populated counties are over-represented at both ends of the death rate distribution.

Wainer's second example discusses the billions of dollars wasted on supporting smaller schools in the USA by educational charities. Smaller schools tend to outperform larger schools on student achievement, and so a logical conclusion is that if larger schools were split into several smaller ones, then student achievement would increase. However, smaller schools are expected to be over-represented in both the top and bottom of the achievement distribution, and that is what Wainer found. Again, with a small sample size (number of students), achievement results will have a wider spread and therefore the tails of the distributions will mostly contain smaller schools.

This phenomenon of larger variances with smaller sample sizes is also relevant for biological experiments. The following situation is often seen in high-throughput screening experiments. The data are simulated, but the example is from a real experiment. Suppose 5000 compounds are tested in a cellular assay and the goal is to find compounds that increase the number of cells expressing a key protein marker. Cells are plated in high-density microtitre plates and images are taken after treatment with the compounds. The total number of cells and the number of cells positive for the marker (which is a subset of the total cell count) are obtained from the images. There are no active compounds in this simulation and so the results are what would be expected from random fluctuations. The distribution of total cell counts across all 5000 wells is shown in Figure 1.5A. The average number of cells per well is 122, but it ranges from 3 to 425. The distribution of positive cell counts across the 5000 wells is shown in 1.5B. The number of positive cells is related to the total number of cells (Figure 1.5C), and the proportion (or percentage) of positive cells is calculated by dividing the number of positive cells by the total number of cells (Figure 1.5D). The proportion was calculated because it supposedly removes the dependence on total cell count. The mean proportion is 0.25 and the range is 0 to 0.8 (the few high values are hard to see in this graph). A cut-off is made based on some criterion such as three standard deviations above the mean of the distribution (dashed line in Figure 1.5D), and all compounds above that are considered 'hits' and will be tested in further experiments. The criterion or threshold used to determine a hit is not important for this example.

Figure 1.5E shows how the *variation* in the proportion of positive cells is dependent on the total cell count, even though the mean no longer is. When cell counts are low, variation is high, and vice versa. The horizontal dashed line is the threshold for hit calling, and all of the hits (high proportion of positive cells) are from wells with few cells. One compound

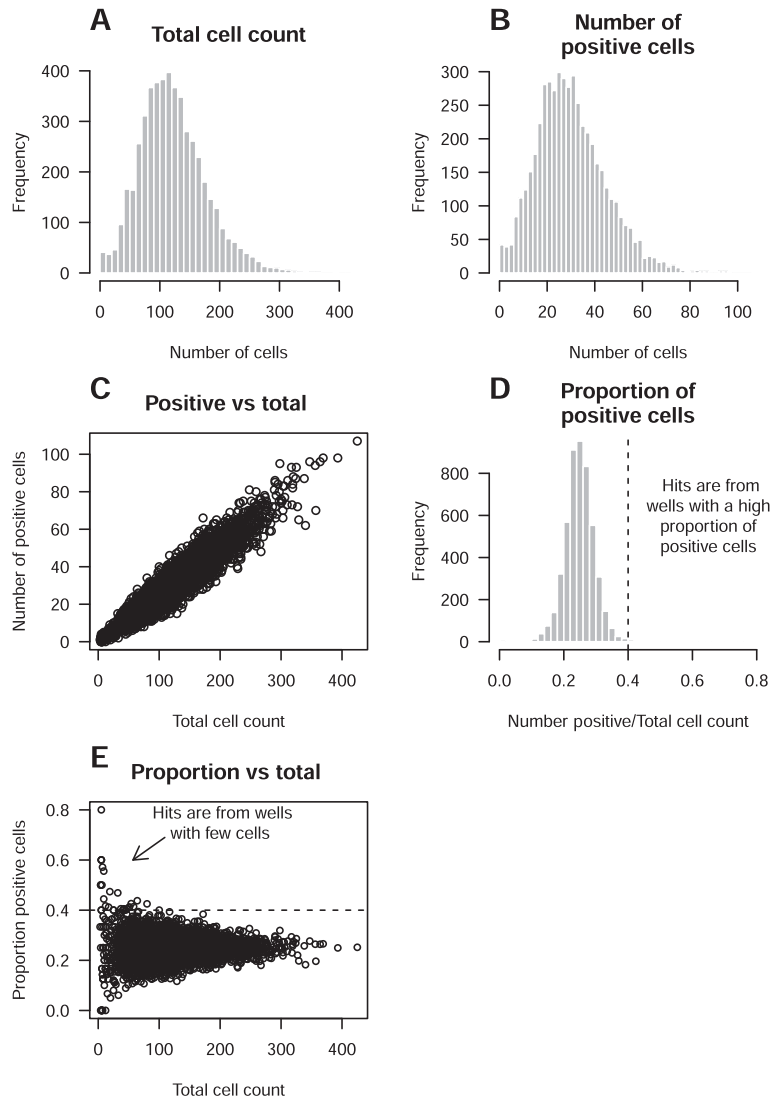

Fig. 1.5 Sensitivity to sample size. The two measured variables are the total cell count and the number of positive cells (A, B). There is a correlation between these variables because the positive cells are a subset of the total number of cells (C). A common strategy to remove the dependence on total count is to take the ratio of positive to total cells, and to look for high ratios (D). However, wells with few cells have the highest ratios (E), which are statistical artefacts.

has a very high proportion of 0.8. This is a statistical artefact but is routinely seen in real experiments and can be mistaken for a true hit. Resources could then be wasted in trying to validate it. But why does this occur? Think of each cell as having a probability of being positive, determined by the flip of an unbalanced coin. Each cell has a 25% chance of being positive (coin lands heads) and a 75% chance of being negative (coin lands tails). If there are only three coins, it is not unusual that all three of them land heads (proportion = 1),

especially when 5000 trys are made.[4] If there are 100 coins, it is unlikely that all will land heads.

A subtler example occurs when classifying active compounds according to the reason the compound was selected for inclusion in the assay.[5] For example, some compounds are selected for testing because they bind to proteins in a biochemical pathway believed to be relevant for the disease. Alternatively, epigenetic mechanisms might be important and so any compound that is known to affect DNA methylation or histone acetylation is included. Also, a set of chemically diverse compounds are often used to cover a broad range of the chemical space. At the end of the screen it is usually of interest to see if one of the three compound sets (pathway set, epigenetic set, diverse set) is enriched for hits. For example, if 12% of the epigenetic compounds tested are hits while only 4.5% of the pathway set and 6% of the diverse set are hits, then epigenetic mechanisms might be important and further effort should be focused here. But it would not be surprising for the epigenetic set to have an unusually high or low percentage of hits if it contained only 50 compounds, while the pathway set had 5000 and the diverse set 50 000. Thus, the percentage of hits does not provide enough information to make conclusions about the compound sets. Can you think of examples from your own experiments where sampling variability might provide a different interpretation of the results?

Sampling variability is also important when assessing the reproducibility of results. Some researchers are surprised when they obtain different results after repeating an experiment, especially when great effort was spent to make the replication as similar as possible to the original experiment. 'Different results' are usually defined as one with $p < 0.05$ and the other with $p > 0.05$. Comparing p-values is not a good way of assessing reproducibility, either within or between laboratories (see Figure 1.1). It is impossible to exactly reproduce an experiment, but even if it were possible, the results would not be identical because of sampling variability. Just like tossing a coin 10 times gives a different number of heads every time, estimating the difference between two groups will give a different estimate every time the experiment is conducted. An example is shown in Figure 1.6, where hypothetical data were generated with a mean difference of 0.5 in each of the 10 experiments. The mean differences and 95% CI are shown, along with the p-values. Half of the p-values are significant and the other half are not. This does not indicate lack of reproducibility and reflects the variation that one would expect due to sampling variability alone, as this was a simulated example and so all other sources of variation are under complete control. Compare Experiment 9, which is significant, and 10, which is not. The estimated mean difference is nearly identical in the two experiments, but Experiment 9 just happened to have better precision (narrower 95% CI) and therefore a smaller p-value. The mean differences from the 10 experiments range from 0.25 to 0.95 and the p-values range from 0.008 to 0.256, and are entirely due to sampling variation. *Large differences in significance can be associated with small differences in the underlying effect.*

[4] The probability of obtaining three heads when the probability of each is 0.25 is $0.25 \times 0.25 \times 0.25 = 0.015625$. If you try this 5000 times, then you would expect $5000 \times 0.015625 = 78$ of those times to get all heads.

[5] I thank Pierre Farmer for this example.

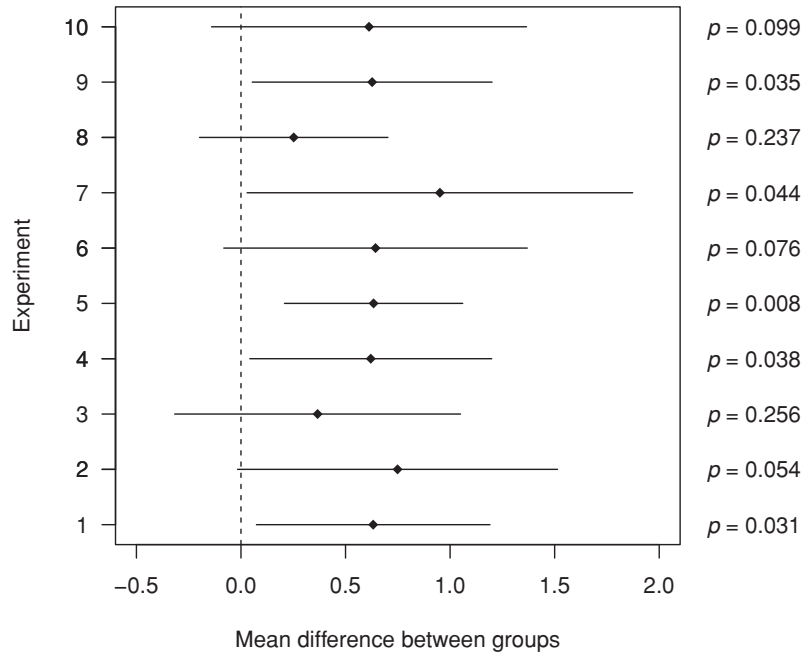

Fig. 1.6 Sampling variability for experimental outcomes. Simulated data from 10 perfectly controlled experiments with a true mean difference of 0.5, and a constant variance and sample size. Variation in the estimated mean difference, 95% CIs, and *p*-values are just a reflection of sampling variability and are within the range of expected values, despite only half of the *p*-values being significant.

1.2.5 Independence bias

Independence bias is the tendency to overestimate the evidential value of new data, especially when the data are correlated with existing data. It leads to the belief that many significant results provide much stronger support for a hypothesis than they actually do. Suppose we are interested in whether a compound increases muscle strength in humans. We randomly assign 20 people to either the compound or placebo control group and assess their strength after 4 weeks of treatment.[6] The subjects' strength is measured on three barbell exercises: bench press, deadlift, and squat. The three strength variables have the following *p*-values when testing for the effect of the compound: $p = 0.01$, $p = 0.03$, and $p = 0.02$. How convincing are the results? How likely is it that all three *p*-values would be significant if the compound was inactive? Many people informally reason that although there is a 0.05 chance of a false positive result, three significant results provides convincing evidence for the effect of the compound, even if the *p*-values are not very small. More formally, if there is a 0.05 chance of a false positive result, the probability of three false positives is $p = 0.05 \times 0.05 \times 0.05 = 0.05^3 = 0.000125$. Since this total probability is

[6] This experiment could be improved by taking the subjects' baseline strength into account, but we ignore this for the sake of simplicity.

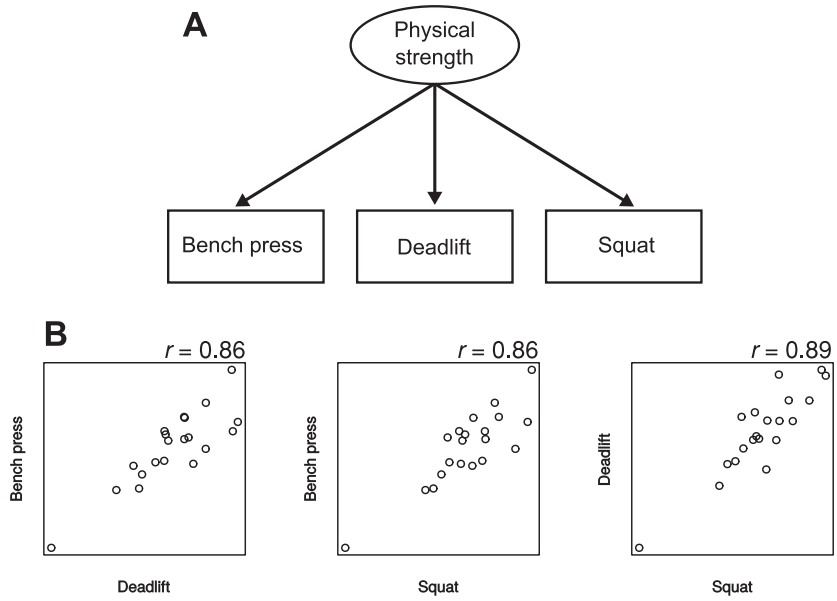

Fig. 1.7 Latent and measured variables. Three measured variables are measuring different aspects of physical strength – the underlying latent variable – and are therefore highly correlated (A). The scatterplots (B) show the relationships between the measured variables along with the Pearson correlation (*r*).

small, it is unlikely that all three tests would be significant if the compound is inactive. The mental bias arises because the three measured outcomes do not provide *independent* pieces of information about the effect of the compound. The total probability calculated is only valid if the three *p*-values are independent, which would occur, for example, if the values were from three different experiments using different people. At the other extreme, if the three measured outcomes are perfectly correlated, then they would all be significant or none of them would, the second and third *p*-value provide no new information once we known the first. The higher the correlation between the variables the greater the redundancy in the information they provide.

The three variables in this example are highly correlated because they are all measuring the same thing – the people's strength. The three exercises are all capturing different aspects of physical strength, but the high correlations are expected. Strength, like many aspects of biology, is hard to define and often cannot be measured directly. Other examples include inflammation, disease severity or stage, cognitive functioning, psychological and emotional states, and even gene expression. These are not directly observed or measured like a person's weight. Variables that cannot be directly measured are called *latent variables*.[7] Figure 1.7A shows the relationship between latent and observed variables for the physical strength example. Figure 1.7B shows scatterplots of 20 simulated values for the three measured variables, with an average correlation of 0.87 between them. Assume

[7] The observed variables are called *manifest* or *indicator* variables, but these terms tend to be discipline specific. We use the terms *observed* or *measured* variables.

that 10 of these values are the controls and the other 10 are the treated people, and that the compound has no effect. What is the probability *all three* statistical tests will have a p-value of less than 0.05? Based on a simulation (not shown) it is about 0.02, just below the usual 0.05 cut-off, but much higher than 0.000125 if independence is assumed. The three significant p-values provide some evidence for an effect of the compound, but it is almost as likely as obtaining one significant p-value.

Highly correlated variables do not provide independent evidence for an effect because they are often different measures of a single underlying effect. The above example used physical strength, but the same problem arises with gene, protein, or metabolite levels as these tend to be co-regulated.

There are several ways to avoid over-interpreting the degree of evidence provided by many correlated variables. First, one variable can be defined beforehand as the primary outcome, and then only this variable is used for testing hypotheses or making a decision. A primary outcome can be chosen based on the literature, after pilot experiments, or during assay development. For example, the outcome with the greatest sensitivity to discriminate between negative control and positive control samples could be chosen, or the variable with the smallest coefficient of variation (CV). The primary outcome is then used for all subsequent experiments because it has been validated as the best variable. Criteria to consider when choosing a primary outcome are discussed in Section 5.1. This approach works well in theory, but often not in practice, because if the result for the primary variable is not significant, but one of the other variables is, many researchers would find it hard to stick with the original plan and base their conclusions on the primary outcome, especially if a significant result is desired. A drawback of the primary outcome approach is that the other variables have some additional information that remains unused. It seems wasteful and inefficient that data are collected but not put to use. One way the additional data could be used is to reduce measurement error. For the strength example, one person may have a stiff shoulder on the day of testing, another may have a sore knee, and another a bad back. Depending on how much an exercise involves an injured body part, a single measurement can underestimate the true strength of a person. If the three measurements could be combined, then a better estimate of overall strength could be obtained.

This brings us to the second method of dealing with many correlated variables: combine them into a set of fewer variables that still relate to the latent variable of interest. We could for example add the amount lifted in the three exercises to obtain a new variable called 'total weight':

$$\text{Total weight} = \text{Bench press} + \text{Deadlift} + \text{Squat}$$

Total weight is a linear combination of the original three variables and can be used as the primary outcome in an analysis. The measured variables are all of the same kind and have the same units (e.g. kilograms) so a simple summation is meaningful. When the observed variables have different units, summing them creates a variable that is hard to interpret. Another problem with adding variables is that variables with larger values will have a larger influence on the total than variables with smaller values. Exercises where a lot of weight can be lifted will contribute more to the total than exercises where less can be lifted. This is undesirable and we would like all variables to contribute equally to the total. Ideally, we

would like a general method to combine variables of different types, where each variable makes the same contribution. Fortunately, many methods are available and the two most common are *principal components analysis* (PCA) and *factor analysis*. The popularity of each method varies widely by scientific discipline. Shipley [354] and Grace [140] discuss the use of such latent variable models in biology to reduce measurement error.

Related to this are *composite measures*, which are a combination of several measured variables, but they may be on different scales. For example, an overall assessment of disease severity or disease burden is composed of several subscales such as the degree of cognitive and motor impairment, which are assessed on a zero (no impairment) to five (complete impairment) scale. The sum of the subscales gives the overall disease severity score. Although composite scores have the advantage of reducing the number of outcome variables, different patterns on the subscales will be obscured. For example, one patient has mostly cognitive impairments, another mostly motor impairments, and a third is mildly impaired on both. Their composite score can be the same, despite differences in their disease manifestation. For this reason, composite scores are especially problematic when looking for associations between clinical outcomes and gene expression or imaging biomarkers.

1.2.6 Confirmation bias

Confirmation bias is the tendency to search for, interpret, focus on, and recall information that confirms a research hypothesis and can manifest itself in many ways. For example, suppose a microarray study is conducted and a list of differentially expressed genes between diseased patients and healthy controls is derived. It is common to provide support for a gene in this list by citing previous studies that found an association between the gene and the disease in question. A PubMed search[8] is conducted using the disease name plus a gene of interest as search terms. The papers found will tend to be those that show a link or association with the gene and the disease, and this may seem to provide support for the findings of the microarray study. But how convincing is this approach (ignoring for the moment that the papers found will be of varying quality and provide various degrees of support for the gene of interest)? What about the studies that examined the gene (or protein) in this disease but *did not* find it to be relevant? There could be many papers that addressed the same research question, but only those that mention the disease and gene in the title or abstract will be found with a PubMed search, and they will tend to be ones with statistically significant results (Figure 1.8). The studies returned from a literature search could be mostly those with false positive results. This is especially true with -omics experiments, where all genes are examined but only a few will be mentioned in the title or abstract, and will therefore be found with a PubMed search. In the above example, there has been no attempt to find negative results, for example, by searching for the disease and gene using the Gene Expression Omnibus (GEO) database.[9] A GEO search will find the

[8] www.ncbi.nlm.nih.gov/pubmed.

[9] The GEO homepage is http://www.ncbi.nlm.nih.gov/geo/, and GEO Profiles can be used to find the expression profiles for a gene in a disease or other experimental condition by using 'gene AND disease' as search terms (http://www.ncbi.nlm.nih.gov/geoprofiles/).

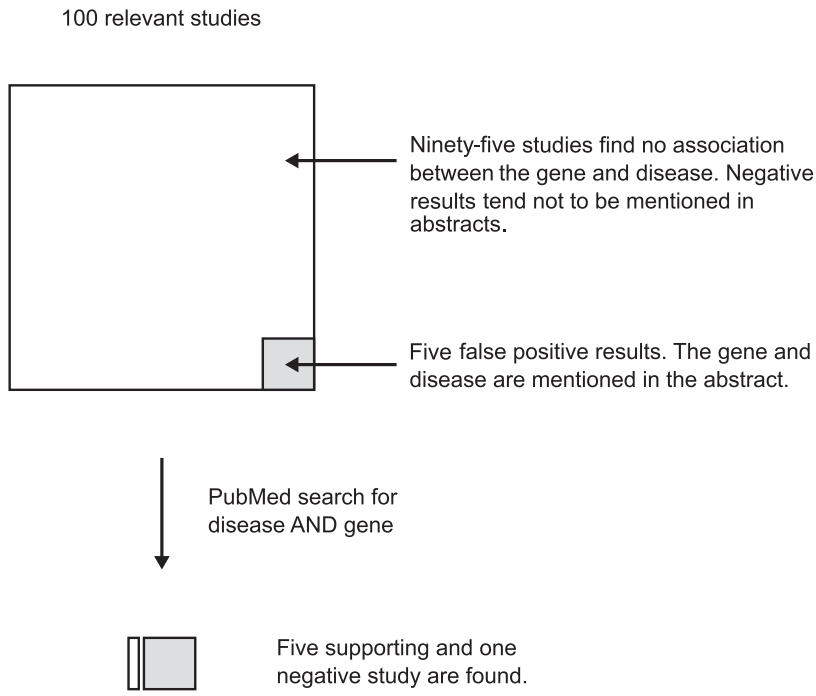

Fig. 1.8 Confirmation bias in literature searches. The tendency to only search for papers that support findings instead of all papers that potentially disconfirm findings. PubMed searches will tend to return positive results because they are more likely to be mentioned in the title or abstract.

gene of interest across all of the microarray and NGS studies it contains that mention the disease. One can then see how many relevant studies were conducted and the number that did not find the gene to be differentially expressed between patients and healthy controls.

Confirmation bias can also occur during the analysis and interpretation of data. Suppose that the data are slightly skewed, and a decision is made to log-transform it. An analysis is conducted on both the raw and transformed data, and if only one of these analyses provides a significant result for the main comparison of interest, then it is the result that is reported. It is as if *by definition* the correct analysis is the one that gives the significant result; the other analyses that were conducted are disregarded because they do not support the research hypothesis.

Another source of confirmation bias occurs when deciding which data to include in a publication to 'tell a story', and which references to cite when supporting a claim. Data that do not fit with the overall story may be excluded. There is also the tendency to minimise the importance of negative results and to cite them less – especially in the biological sciences [108]. Confirmation bias is hard to avoid, but efforts can be made to find negative results and ensure that they are not discounted and are appropriately cited. Tetlock suggests that confirmation bias can be mitigated by simply rewording the question [367]. For example, in addition to asking 'is there an association between this gene and disease?', which biases

one to find supporting evidence, also ask 'why is there no association between this gene and disease?', which focuses the search on finding negative evidence.

1.2.7 Expectancy effects

Expectancy effects occur when a scientist's expectations influence measurements or assessments. For example, if transgenic mice are expected to exhibit more of one behaviour compared with wild-type controls, then subjective ratings of behaviour can overestimate their prevalence in the transgenic mice. Or, if no differences are expected between litters of animals or batches of samples processed separately, then subtle clustering of data points may be ignored or attributed to random noise.

A classic example of expectancy effects is the story of N-rays – a story rooted in poor experimental design [194, 296]. In 1903, French physicist Rene Prosper Blondlot claimed to have discovered a new type of radiation, which he called N-rays, and many respected scientists reproduced his results. It was an important finding and received a lot of attention from the scientific community. N-rays were supposedly emitted from organic objects and biologists and medical doctors became interested.

N-rays were detected by subjectively assessing the brightness of a spark or the darkness of photographic plates. The researchers were not blinded during these assessments and they saw what they expected and wanted to see: when we do X, the spark gets brighter. The effects however disappeared in later experiments when researchers were unaware of the experimental condition when assessing the brightness of the spark, that is, when they were *blinded* [194].

Between 1903 and 1906, some 300 papers on N-rays were published by 100 scientists [296]. A couple of years after their discovery, few believed in the existence of N-rays, and science appears to be working as it should: a claim was made, scientists investigated, and eventually the truth was found. On the other hand, if expectancy effects had been controlled from the start, 100 scientists need not have wasted their time. Blinding (Section 2.5) and randomisation (Section 2.3) could have prevented the expectancy effects.

1.2.8 Hindsight bias

Hindsight bias is the tendency to find explanations for results that were not predicted – often consistent with one's hypothesis or the prevailing paradigm. It is mainly a problem for exploratory experiments (as opposed to confirmatory experiments, see Section 2.1) and occurs whenever an unplanned or unanticipated effect or association is statistically significant, and we conclude 'that makes sense' or 'I knew it all along' [390]. Sometimes the explanation is not obvious, but after some thought and a PubMed search, one can appeal to some theory or find a couple of papers that can be marshalled in support of an explanation. With almost 25 million entries in PubMed, something useful can be found. Furthermore, given the number of false positive results in the literature, it is not hard to find at least one paper that supports any given finding. The only way to validate such results is to admit that they were not predicted, make a prediction about what will be found in a *new* experiment, and then conduct the new experiment [135]. When an explanation comes

after the result, it is extremely weak and unconvincing. Researchers may develop different explanations for the same results depending on their background knowledge and the papers they stumbled across in their internet search. The proposed explanation may be true, but the only way to know for sure is to test it in a subsequent experiment.

Hindsight bias can be avoided by making a prediction before the experiment is conducted. It helps to write down the expected results of the experiment; for example, that the treated group will express Gene X at a higher level. One could go a step further and predict the size of the effect and its uncertainty, such as a 2-fold increase with a 95% CI between 1.5- and 2.5-fold. The estimate can be based on effect sizes commonly seen in the literature for related experiments or a pilot study. Maybe write down a value for an unbelievably large effect. This helps to calibrate predictions and intuitions. If the result is a 10-fold increase, then the usual reaction of excitement might be tempered with concern of why the prediction was so different and why the effect is so large. Is there another source of variation that is influencing the results? Furthermore, if Gene X is the same between groups but Gene Y and Gene Z are differentially expressed, then one cannot claim these results were expected.

1.2.9 Herding effect

Herding behaviour is the tendency to follow the scientific crowd when it comes to theories believed and methods used (either experimental or statistical) – science is not immune to fashionable trends. For example, a protein or biological process is believed to be critically important for a disease. Then, everyone moves on to study another protein or process that is now thought to be of greater importance. In drug discovery, many companies chase after the same targets at about the same time. Such herding behaviour makes it hard to publish findings that go against the prevailing view. And when such results are published, they may be ignored by the research community. Olson argues that findings are ignored when they do not contribute to the overarching narrative that scientists use to understand and interpret results,[10] and cites his own work in marine biology as an example [300, p. 211].

The herding effect can also occur at the level of a research group, where all members have the same beliefs, are trained in the same methods, and conduct experiments to support a single point of view. This can lead to *scientific inbreeding*, where replications are not truly independent because they have the same biases and errors as the original experiment [170].

Herding behaviour is a problem because there may be few people that take a critical view of a research area, demand that assumptions are checked, and seriously consider alternative explanations.[11] A risk is that contradictory or negative results are suppressed and experiments are run again until the 'right' answer is obtained. Such behaviour is hard to overcome because we often take basic premises as given, and then try to extend scientific knowledge

[10] A similar idea to Kuhn's concept of a *paradigm* [206].

[11] Not fully appreciating how alternative explanations can account for observed results is a problem in the field of adult neurogenesis [217, 218, 223], where I spent my PhD and postdoc years.

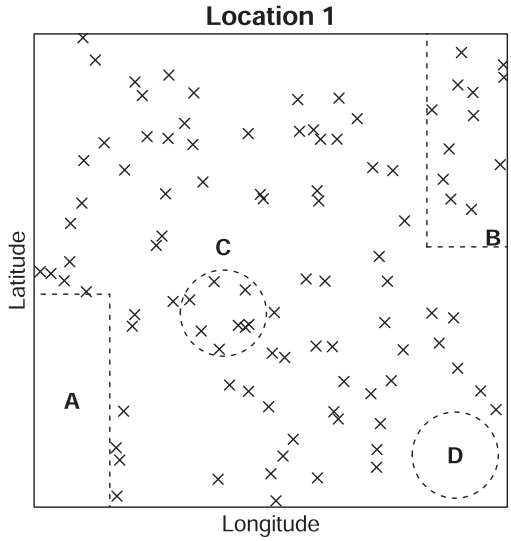

Fig. 1.9 Positions of dropped bombs. Regions (A) and (B) are fields, Region (C) is a power plant, and Region (D) is a munitions factory.

in a certain direction. The admonishment to be an independent thinker is unlikely to be helpful, as most people – scientist or otherwise – already believe that they are.

1.2.10 How the biases combine

The above biases and mental misrepresentations often work together to distort scientific results. As an illustration, we return to the example of whether bombs are dropped at random at Location 1 from Figure 1.2. The graph is reproduced in Figure 1.9 and further information is added. Recall that the General asked you to determine if the bombs are dropped at random, or if some positions are targeted more heavily while others are avoided. Also recall that these positions were generated by randomly selecting latitude and longitude pairs from a uniform distribution.

After discussing with colleagues, reading the literature on bombing strategies, and looking at what others have concluded about previous bombings, the consensus is that the enemy is unlikely to randomly drop bombs (herding effect). So when you start to analyse the data there is an expectation that you will find further evidence, and this is what you set out to do (confirmation bias). A first look at the bomb positions suggests that there may be some areas that are more heavily bombed while others have few bombs (seeing patterns in randomness). For example, you notice that Region A in Figure 1.9 has no bombs in it. After some further investigation you find that this region is mostly an empty field, which makes perfect sense because a field is not of strategic interest (hindsight bias). Furthermore, you notice that Region C has a denser clustering of bombs. This region is close to – but not exactly encompassing – a power plant, which is of definite strategic importance (hindsight bias). You know the importance of a formal statistical analysis and develop the idea that by

calculating the area of a region relative to the total area, and then given the total number of bombs dropped, you can work out if the region has significantly greater or fewer bombs than expected. Your analysis confirms your intuitions that Region A has significantly fewer bombs than expected and Region C has significantly more! You examine other areas that might suggest an over- or under-representation of bombs (not wanting to miss anything), but nothing else is significant, and so you conclude that the power station was the only specific position targeted (psychological cliff at 0.05).

It is clear how several biases combined to produce the incorrect conclusion. There are additional issues. First, the absence of bombs in Region A is taken as evidence for avoiding locations with no strategic importance, such as an empty field. But Region B is also an empty field, and the bombs here are not taken as evidence *against* this hypothesis, either because the number of bombs in this field was not checked, or if it was checked, the results were downplayed because they were inconsistent with expectations. In either case, it is another example of confirmation bias.

Second, Region A was defined as not of strategic interest because it was mostly an empty field. But what if it contains a dairy farm, and the field is for the cows to graze? Is it now of strategic importance? What if the dairy farm produces high protein rations from milk products for the army? If there were more bombs than expected in Region A, then we could have made the story that the enemy is targeting the food supply (hindsight bias). What is considered strategically important is not defined beforehand. There was nothing *a priori* interesting about the patch of land contained in Region A until it was observed that there were fewer bombs there. Third, the clustering of bombs in Region C is taken as evidence for targeting important positions (the power plant), but the lack of bombs in Region D is not taken as evidence against this hypothesis. Suppose Region D is a munitions factory. Again, what is of strategic importance is not defined beforehand, and how close do the bombs have to be to a strategic location to be included in the count? Recall that Region C was *close to* a power plant. How close is close? Furthermore, why is Region A defined as a rectangle and Region C a circle?

A fourth, much subtler point, is that the analyses are not independent, and so the two significant p-values are correlated (independence bias). If there is an unusually high number of bombs in one region, then by definition, there must be fewer in the remaining region. It is similar to testing if eight heads out of 10 coin tosses is significant, and then testing if two tails out of 10 tosses is also significant. Since the number of tails is equal to the number of tosses minus the number of heads, the two p-values are redundant. Since the regions are not defined beforehand, the dependence is not as strong as in the coin example.

Pleased with your analysis you decide to share the results with the General, who is happy to see that his belief in a non-random bombing strategy has been vindicated. He will recommend you for a promotion and also suggests that you distribute the findings more widely. You decide to deposit them in either the repository of **NAT**ional **U**seful **RE**ports or the **C**entral **EL**aborated **L**ists, making sure to support your conclusions by citing earlier reports that came to the same conclusion (but not those that found the opposite results; confirmation bias). There is now one more piece of evidence to add to the collection that argues in favour of targeted non-random bombing by the enemy. And the cycle continues...

This was a fictitious example, but the parallels with experimental biology are clear. There was no fraudulent activity, the analysis was technically correct and, superficially, many of the steps seem sensible. But taken together, the approach led us astray. The question remains about how we should have proceeded. Certainly, more could have been defined before the analysis was started, including what is a strategic target (there can be several priority levels), how close do bombs have to be to the target to be included in the count (information on the accuracy of the enemy bombers can be used). This narrows the options available for finding patterns in pure noise. In addition, fake data generated by random uniform sampling of positions could have been included, and the analyst not told which data are real and which are fake [64]. This introduces blinding at the analysis stage. One could even withhold information that fake data are included. If the analyst can make a convincing argument that the fake data show signs of non-random bombing, then their conclusions about the real data become suspect.

A key mistake was trying to prove non-randomness from the start instead of trying to disprove randomness. This is, again, an example of confirmation bias, where we are trying to confirm our hypothesis of targeted bombing instead of trying to disprove the opposite hypothesis of random bombing. The problem with trying to confirm non-randomness is that we have to specify what particular type of non-randomness we want to test, and this led to the data-driven selection of whatever looked interesting. To disprove the hypothesis of random bombing, we only have to reject the hypothesis of *complete spatial randomness* (CSR) without specifying how exactly non-randomness will manifest itself. Methods exist for testing CSR but are not discussed here (see [96]). If the data are consistent with CSR, then there is no need to investigate further.

1.3 Are most published results wrong?

The heading for this section is inspired by Ioannidis' conclusion that most published results are false [168] (and in a subsequent paper he argued that the effect size is often overestimated in those studies that reached the correct conclusion [169]). The short answer is that many papers provide insufficient information to assess whether the experimental design and statistical analysis are appropriate or whether the investigators engaged in any questionable research practices, such as reporting only favourable results. There has been a good deal of introspection by the biomedical research community lately regarding bias and the lack of reproducibility [2, 5, 28, 30, 42, 80, 83, 84, 123, 172, 240, 251, 305, 316, 342], which suggests that there are serious problems with current biomedical research. These discussions have led top journals (e.g. *Science*, *Science Translational Medicine*, *Nature*, and the *Nature* series of journals) to implement stricter reporting standards [3, 183, 271], and the National Institutes of Health (NIH) to pilot various methods for improving the reproducibility of research – including training in experimental design for NIH intramural postdoctoral fellows [80]. In addition, *Science* has recently introduced a Statistical Board of Reviewing Editors, who will check manuscripts and recommend those that should receive a more thorough review [270]. Below we discuss several lines of evidence suggesting

that much of the literature is of questionable value, including anecdotal comments from statisticians and scientists, empirical studies examining the quality of published reports, large-scale attempts at replication, and anonymous questionnaires given to scientists asking about questionable research practices that they have observed or engaged in.

1.3.1 What statisticians say

This chapter opened with quotes from eminent statisticians on the quality of experiments and analyses conducted by scientists. Statisticians are in a unique position because they see the raw data and interact with scientists before papers are published. They are often exposed to 'the real story' instead of the sanitised version that gets published. They are also less invested in the outcome – it's not their hypothesis being tested – which enables them to be more impartial. Applied biostatisticians also work with many research groups and have a broad exposure to how biologists conduct experiments. Given statisticians' knowledge on designing experiments and making inferences from data, such comments should cause one to pause and reflect on the quality of experiments in the biological sciences. Nelder's comment opening this chapter 'and many statisticians have their own horror stories' alludes to statisticians' widespread concern. It is noteworthy that Fisher's comment was made in 1938 and Nelder's and Dempster's over 60 years later, suggesting that the quality of experiments has not improved over time. Fisher made his statement at a statistical congress and Nelder and Dempster commented in statistics journals [93, 290]. In all cases the comments were directed at fellow statisticians – who perhaps nodded their heads knowingly – but their concerns were not communicated to biologists. Further comments from statisticians on experimental quality and statistical inference are below and most were published in journals or books that biologists are unlikely to read.

> *In practical situations many scientific or industrial investigations are doomed to fail. There are many and varied reasons for this, but the most often encountered reason is simply that the investigation was not properly planned. Many investigators fail to understand that careful pre-planning is essential for a successful experiment.* ([153, p. 30])

> *. . . the statistician should request a detailed description of the experiment and its objectives. It may then become evident that no inferences can be made or that those which can be made do not answer the questions to which the experimenter had hoped to find answers. In these unhappy circumstances, about all that can be done is to indicate, if possible, how to avoid this outcome in future experiments.* ([373, p. 559])

> *. . . how many investigators are fully aware of all of the potential sources of bias in their experimental protocol and understand how to incorporate them into the design and allow for them in the analysis to avoid any negative impact on the conclusions?* ([305, p. 734])

> *There is a disturbing tendency for techniques to be used by people who do not fully understand them and the standard of statistical argument in scientific journals can best be described as variable.* ([70, p. 214])

> *One of us worked for many years across multiple research groups at a large medical school and was occasionally asked to evaluate the work of other research groups, with*

> *access to all data and staff. It was alarming how often mistakes occurred and went undetected.* ([131, p. 4])

> *Today's medical journals are full of factual errors and false conclusions arising from lack of statistical common sense.* ([59, p. 335])

> *Statistical thinking is the statistical incarnation of 'common sense'. 'We know it when we see it', or perhaps more truthfully, its* absence *is often glaringly obvious.* ([397, p. 223])

> *We think that the problem is that often researchers do not admit uncertainty or variation; they think they've already made their discovery, and they think of various data-collection and data-analysis rules as technicalities that should not get in the way of science. After all, if you've published a paper with nine statistically significant results, it would seem like you've discovered a pattern that could only occur once in* $(1/20)^9$ *by chance, a probability that would seem too extreme to be seriously whittled away by minor methodological issues.* ([130, p. 54])

> *It is often the case (e.g., in biology) that the experimental worker shows a certain, indeed a strong, prejudice against statistical work.* ([411, p. 3])

> *One thing has become painfully clear to me in twenty years of extensive teaching, statistical consulting, reviewing, and interacting in ecology: ecologists' understanding of statistics is abysmally poor. Statistics should naturally be a source of strength and confidence. . . Unfortunately, it is all too frequently a source of weakness, insecurity, and embarrassment.* ([94, p. 332])

> *There was little interest for the principles of sound statistical design and analysis of experiments.* (From a workshop on statistical thinking for scientists in the pharmaceutical industry.) ([375, p. 65])

Some biologists may be offended by these comments, but consider what biologists might say if statisticians started conducting experiments in the lab without regard for good practices such as using experimental controls and validating reagents. The point remains, designing experiments and analysing data are core activities in experimental biology, and many who are expert in these areas of scientific discovery think that the general level of knowledge among biologists is insufficient. Biologists are not at fault – formal education and training in these areas are minimal and biologists have to go out of their way to learn more about topics that many find intrinsically boring and disconnected from their scientific activities. That many biologists do not see how statistics and experimental design are critical for their research is a failing of the statistics community. Biologists want to do good science and will readily learn something if it will improve their research. The rigid formality of setting up a null hypothesis (that is known to be false), comparing it with an alternative hypothesis (that is uninteresting), calculating p-values that do not say anything about hypotheses and confidence intervals that do not provide confidence, is enough to convince many biologists that statistics has nothing useful to offer.

It is not obvious how statistics and experimental design can improve scientific discovery and it is the responsibility of statisticians (or those writing about statistics) to make this

connection for scientists.[12] Focusing on experimental design is one way of making statistics more relevant for biologists – the approach taken in this book – because it is closely related to scientists' day-to-day activities. A second approach is to use Bayesian methods for statistical inference instead of frequentist methods. Introducing an unfamiliar method of inference in a book primarily about experimental design would take us too far afield and that is why frequentist methods have been used in Chapter 4. McElreath [269] and Kruschke [203] provide excellent introductions to Bayesian analysis.

1.3.2 What scientists say

Biologists may be dismissive of what statisticians think, maybe because statisticians are perceived as lacking biological knowledge or are pedantically obsessed with trivial matters and miss the bigger picture.[13] Concerns that scientists have about their own areas of research are often not captured in the literature but traded over a beer at conferences, although a few reports are now appearing. One example is from Glenn Begley; he relates how he was unable to replicate results published in *Cancer Cell*, and when discussing this with the (respected) senior author of the paper, Begley was told that the experiment was conducted many times, and only the one positive experiment was published [83]. Professor Ulrich Dirnagl discusses a paper he reviewed that excluded three animals in the treated group without mentioning it [83]. The methods stated that 10 animals were used in the treated group, but only seven animals were shown in a figure. The study was testing a new compound to see if it would protect the brain after a stroke. As it turns out, the missing three animals in the treated group had died of a massive stroke, and their exclusion made the treatment appear effective.

A detailed account of science gone wrong is by John Maddox (editor of *Nature* at the time) and colleagues, who had visited the lab of Dr Jacques Benveniste to scientifically audit a remarkable yet unlikely finding [255]. Benveniste had published a paper in *Nature* demonstrating that highly diluted solutions of anti-IgE showed biological activity, which could be interpreted as evidence in support of homeopathic medicine [91]. The paper came with an editorial reservation at the end and an accompanying opinion piece on when to believe the unbelievable [1]. During their week-long stay in Benveniste's lab, Maddox and colleagues described a remarkable list of poor experimental practices, including measuring control samples a second time if the results were higher than the treated samples – because the first reading 'must have been incorrect' – or excluding results that failed to 'work', defined as not in line with expectations. Experimenters were not blinded and thus observer bias could have been introduced. In addition, there was an undeclared conflict of interest as a company that produced homeopathic medicines funded two of the authors on the paper. Perhaps the most interesting feature is the psychology of the scientists and how

[12] Many statisticians have successfully made this connection – George Box and J. Stuart Hunter are personal favourites. Google 'Stu Hunter Teaches Statistics' and watch the first YouTube video on *What is Design of Experiments – Part 1* (`https://www.youtube.com/playlist?list=PL335F9F2DE78A358B`) as an example.

[13] Fisher was just as much a biologist as a statistician, his appointment at Cambridge University was the Balfour Chair of Genetics; he was not based in a mathematics or statistics department.

they were able to explain away results that were evidence against their hypothesis, accept uncritically results that supported their views, and how this justified recounting until the 'correct' answer was obtained or excluding data altogether.[14] The most troubling aspect of such research practices is just how common they are.

Lack of reproducibility is not a recent phenomenon, but it is interesting to speculate why attention has increased recently. In 1673 Robert Boyle wrote:

> *You will find* [...] *many of the experiments published by authors, or related to you by the persons you converse with, false and unsuccessful* [...] *you will meet with several observations and experiments which, though communicated for true by candid authors or undistrusted eye-witnesses, or perhaps recommended by our own experience, may, upon further trial, disappoint your expectation, either not at all succeeding constantly, or at least varying much from what you expected.* (Quoted in Fisher [121, p. xv])

1.3.3 Empirical evidence I: questionable research practices

Martinson and colleagues anonymously asked 3247 NIH-funded scientists if they had engaged in a list of questionable research practices (QRPs) during the previous 3 years [260]. Overall, they found that 33% of scientists reported engaging in at least one of their top 10 most serious behaviours, 0.3% of respondents admitted to falsifying data, the most serious item on Martinson's list; 6% admitted to not presenting data that contradicted their previous results, 12.5% said that they overlooked others' flawed or questionable interpretation of data, 15.5% said that they changed the design, methods, or results in response to pressure from a funding source, 13.5% admitted using inadequate or inappropriate research designs, and 15.3% dropped observations from analyses based on a gut feeling that they were inaccurate. A meta-analysis of 18 studies by Fanelli found that 2% of scientists admitted to fabricating, falsifying, or modifying data or results at least once, and 34% admitted to other QRPs [106]. When asked about colleagues, admission rates were 14% for falsification and up to 72% for other QRPs.

More recently, John and colleagues anonymously asked 2155 psychologists if they had engaged in a list of 10 QRPs [177]. Unlike Martinson's study there was no time limit, and so scientists responded if they had *ever* engaged in a behaviour. This led to much higher estimates, and the results are grim. Overall, 0.6% admitted to falsifying data, but 63.4% failed to report all dependent measures (i.e. selective reporting), 15.6% stopped collecting data earlier than planned because the desired results were obtained, 45.8% reported 'rounding off' p-values (e.g. reporting 0.054 as <0.05), 45.8% admitted to selectively reporting studies that 'worked', 38.2% said that they excluded data after looking at the impact of doing so on the results, and 27% admitted to reporting an unexpected finding as having been predicted from the start. Lest you think that psychologists are an exceptionally unethical group, when the results were broken down by discipline, neuroscience (the closest group

[14] Dr Benveniste's rebuttal can be found in *Nature* and *Science* [34, 35]. The Wikipedia page for Jacques Benveniste contains more information about the story (http://en.wikipedia.org/wiki/Jacques_Benveniste). His findings have not been substantiated during the subsequent 25 years.