Learning Materials in Biosciences

Michael H. Herzog
Gregory Francis
Aaron Clarke

Understanding Statistics and Experimental Design

How to Not Lie with Statistics

OPEN

Springer

Learning Materials in Biosciences

Learning Materials in Biosciences textbooks compactly and concisely discuss a specific biological, biomedical, biochemical, bioengineering or cell biologic topic. The textbooks in this series are based on lectures for upper-level undergraduates, master's and graduate students, presented and written by authoritative figures in the field at leading universities around the globe.

The titles are organized to guide the reader to a deeper understanding of the concepts covered.

Each textbook provides readers with fundamental insights into the subject and prepares them to independently pursue further thinking and research on the topic. Colored figures, step-by-step protocols and take-home messages offer an accessible approach to learning and understanding.

In addition to being designed to benefit students, Learning Materials textbooks represent a valuable tool for lecturers and teachers, helping them to prepare their own respective coursework.

More information about this series at http://www.springer.com/series/15430

Michael H. Herzog • Gregory Francis •
Aaron Clarke

Understanding Statistics and Experimental Design

How to Not Lie with Statistics

Springer

OPEN

Michael H. Herzog
Brain Mind Institute
École Polytechnique Fédérale de Lausanne
(EPFL)
Lausanne, Switzerland

Gregory Francis
Dept. Psychological Sciences
Purdue University
West Lafayette
IN, USA

Aaron Clarke
Psychology Department
Bilkent University
Ankara, Turkey

ISSN 2509-6125 ISSN 2509-6133 (electronic)
Learning Materials in Biosciences
ISBN 978-3-030-03498-6 ISBN 978-3-030-03499-3 (eBook)
https://doi.org/10.1007/978-3-030-03499-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Science, Society, and Statistics

The modern world is inundated with statistics. Statistics determine what we eat, how we exercise, who we befriend, how we educate our children, and what type of medical treatment we use. Obviously, statistics is ubiquitous and—unfortunately—misunderstandings about statistics are too. In Chap. 1, we will report that judges at law courts come to wrong conclusions—conclusions about whether or not to send people to prison—because they lack basic statistical understanding. We will show that patients committed suicide because doctors did not know how to interpret the outcome of medical tests. Scientists are often no better. We know colleagues who blindly trust the output of their statistical computer programs, even when the results make no sense. We know of published scientific papers containing results that are incompatible with the theoretical conclusions of the authors.

There is an old saying (sometimes attributed to Mark Twain, but apparently older) that "There are lies, damn lies, and statistics." We have to concede that the saying makes a valid point. People do often misuse statistical analyses. Maybe it does not go all the way to lying (making an intentionally false statement), but statistics often seem to confuse rather than clarify issues. The title of this book reflects our efforts to improve the use of statistics so that people who perform analyses will better interpret their data and people who read statistics will better understand them. Understanding the core ideas of statistics helps immediately to reveal that many scientific results are essentially meaningless and explains why many empirical sciences are currently facing a replication crisis.

The confusion about statistics is especially frustrating because the core ideas are actually pretty simple. *Computing* statistics can be very complicated (thus the need for complicated algorithms and thick textbooks with deep theorems), but a good understanding of the basic principles of statistics can be mastered by everyone.

In 2013, with these ideas in mind, we started teaching a course at the Ecole Polytechnique Fédérale de Lausanne in Switzerland on understanding statistics and experimental design. Over the years, the course became rather popular and draws in students from biology, neuroscience, medicine, genetics, psychology, and bioengineering. Typically, these students have already had one or more statistics class that guided them through the

details of statistical analysis. In contrast, our course and this book are designed to flesh out the *basic* principles of those analyses, to succinctly explain what they do, and to promote a better understanding of their capabilities and limitations.

About This Book

Background and Goal As mentioned, misunderstandings about statistics have become a major problem in our societies. One problem is that computing statistics has become so simple that good education seems not to be necessary. The opposite is true, however. Easy to use statistical programs allow people to perform analyses without knowing what the programs do and without knowing how to interpret the results. Untenable conclusions are a common result. The reader will likely be surprised how big the problem is and how useless a large number of studies are. In addition, the reader may be surprised that even basic terms, such as the *p*-value, are likely different from what is commonly believed.

The main goal of this book is to provide a short and to-the-point exposition on the essentials of statistics. Understanding these essentials will prepare the reader to understand and critically evaluate scientific publications in many fields of science. We are not interested in teaching the reader how to *compute* statistics. This can be left to the computer.

Readership This book is for all citizens and scientists who want to understand the principles of statistics and interpret statistics without going into the detailed mathematical computations. It is perhaps surprising that this goal can be achieved with a very short book and only a few equations. We think people (not just students or scientists) either with or without previous statistics classes will benefit from the book.

We kept mathematics to the lowest level possible and provided non-mathematical intuition wherever possible. We added equations only at occasions where they improve understanding. Except for extremely basic math, only some basic notions from probability theory are needed for understanding the main ideas. Most of the notions become intuitively clear when reading the text.

What This Book Is Not About This book is not a course in mathematical statistics (e.g., Borel algebras); it is not a traditional textbook on statistics that covers the many different tests and methods; and it is not a manual of statistical analysis programs, such as SPSS and R. The book is not a compendium explaining as many tests as possible. We tried to provide just enough information to understand the fundamentals of statistics but not much more.

What This Book Is About In Part I, we outline the philosophy of statistics using a minimum of mathematics to make key concepts readily understandable. We will introduce the most basic *t*-test and show how confusions about basic probability can be avoided.

Understanding Part I helps the reader avoid the most common pitfalls of statistics and understand what the most common statistical tests actually compute. We will introduce null hypothesis testing without the complicated traditional approach and use, instead, a simpler approach via Signal Detection Theory (SDT). Part II is more traditional and introduces the classic tests ANOVA and correlations. Parts I and II provide the standard statistics as they are commonly used. Part III shows that we have a science crisis because simple concepts of statistics were misunderstood, such as the notion of replication. For example, the reader may be surprised that too many replications of an experiment can be suspicious rather than a reflection of solid science. Just the basic notions and concepts of Chap. 3 in Part I are needed to understand Part III, which includes ideas that are not presented in other basic textbooks. Even though the main bulk of our book is about statistics, we show how statistics is strongly related to experimental design. Many statistical problems can be avoided by clever, which often means simple, designs.

We believe that the unique mixture of core concepts of statistics (Part I), a short and distinct presentation of the most common statistical tests (Part II), and a new meta-statistical approach (Part III) will not only provide a solid statistical understanding of statistics but also exciting and shocking insights to what determines our daily lives.

Materials For teachers, power point presentations covering the content of the book are available on request via e-mail: michael.herzog@epfl.ch.

Acknowledgements We would like to thank Konrad Neumann and Marc Repnow for proofreading the manuscript and Eddie Christopher, Aline Cretenoud, Max, Gertrud and Heike Herzog, Maya Anna Jastrzebowska, Slim Kammoun, Ilaria Ricchi, Evelina Thunell, Richard Walker, He Xu, and Pierre Devaud for useful comments. We sadly report that Aaron Clarke passed away during the preparation of this book.

Lausanne, Switzerland — Michael H. Herzog
West Lafayette, IN, USA — Gregory Francis
Ankara, Turkey — Aaron Clarke

Contents

Part I

The Essentials of Statistics

Basic Probability Theory

1

Contents

What You Will Learn in This Chapter

Before entering the field of statistics, we warm up with basic probability theory. Without insights into the basics of probability it is difficult to interpret information as it is provided in science and everyday life. In particular, a lot of information provided in the media is essentially useless because it is based on partial information. In this chapter, we will explain what type of complete information is needed for proper conclusions and introduce Bayes' theorem. We will present the ideas using simple equations and, for readers not comfortable with mathematics, we will provide the basic intuitions with simple examples and figures.

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_1

1.1 Confusions About Basic Probabilities: Conditional Probabilities

1.1.1 The Basic Scenario

Some very basic probability theory

1. **Probability.** A probability assigns a value between 0 and 1 to an event A. For example, a dice is thrown. The probability of throwing a 4 is $P(4)$=1/6.
2. **Probability distribution.** There are 6 outcomes in the above example, each outcome is assigned a probability of 1/6. Assigning a probability to each possible outcome produces a probability distribution.
3. **Conditional probability.** A conditional probability $P(A|B)$ takes information of an event B into account. The vertical bar is read as "given," which indicates that this is a conditional probability statement. For example, you draw two cards, one after the other, from a standard deck of 52 cards. The probability of the first card being a spade is P(spade on first draw) = 13/52 = 1/4. Now there are only 51 remaining cards. The probability of the second card being a spade having already drawn a spade is P(spade on second draw|spade on first draw) = 12/51. In contrast, P(spade on second draw|heart on first draw) = 13/51. Here, the probability of the second draw depends on what type of card was drawn first.
4. **Independent events.** Events are independent when the conditional probability is the same as the unconditional probability: $P(A|B) = P(A)$. In this case the probability of A does not depend on B. For example, if after drawing a card you return it to the deck, then the probability of a spade on the second draw is P(spade on second draw) = 13/52, regardless of what card was drawn first.

Definitions

Consider a situation where a patient might be infected and undergoes a test for the infection. We label each of the four possible outcomes as follows:

1. **Sensitivity:** The probability that the test is positive given that the patient is infected.
2. **Specificity:** The probability that the test is negative given that the patient is not infected.
3. **False Positive Rate:** The probability that the test is positive given that the patient is not infected.
4. **Miss Rate:** The probability that the test is negative given that the patient is infected.

Let us start with an example. In the 1980s, a new disease, called acquired immune deficiency syndrome (AIDS), caused a public panic; it was caused by the HIV virus. Scientists developed a highly sensitive test to detect the virus in the blood. Suppose the HIV test has a sensitivity of 0.9999 and a specificity of 0.9999. Hence, the test is a very good test because for the vast majority of cases the test is positive when the patient is infected, and the test is negative when the patient is not infected. Further suppose that the incidence rate of HIV infection is 0.0001 in the normal population, i.e., 1 out of 10,000 people is infected with the HIV virus. Now, a randomly selected person is tested and the result is positive. Assume you are a doctor. What do you tell the person is the probability that he/she is infected? Mathematically, what is the conditional probability to be infected (HIV) given that the test is positive (T^+): $P(HIV|T^+)$?

Because the test is extremely good and makes almost no errors, many people believe that $P(HIV|T^+)$ should be very high, for example, $P(HIV|T^+) = 0.9999$. However, the reality is that $P(HIV|T^+) = 0.5$, which is no better than a coin flip. How can this happen? We can compute $P(HIV|T^+)$ using Bayes theorem, which is here in its general form:

For two events A and B

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

We can now fill in the values ($\neg HIV$ means no HIV infection):

$$\begin{aligned} P(HIV|T^+) &= \frac{P(T^+|HIV) \times P(HIV)}{P(T^+)} \\ &= \frac{P(T^+|HIV) \times P(HIV)}{P(T^+|HIV) \times P(HIV) + P(T^+|\neg HIV) \times P(\neg HIV)} \\ &= \frac{0.9999 \times 0.0001}{0.9999 \times 0.0001 + (1 - 0.9999) \times 0.9999} \\ &= 0.5 \end{aligned}$$

The mathematics gives the answer, but there is a more intuitive way to understand the situation (Fig. 1.1). Assume, 10,000 people are tested. Because the incidence rate is 0.0001, only one person is likely to be infected. Since the sensitivity of the test is extremely high (0.9999), the infection is likely detected by the test. There are 9999 non-infected persons. Even though the specificity is also extremely high (0.9999), the test still likely delivers one false positive. The false positive occurs because so many people were tested. Hence, all together there are only two people with positive test results out of 10,000 people (9998 negative test outcomes). Since only 1 out of the 2 people are infected, the probability to be infected is $\frac{1}{2}$, i.e., $P(HIV|T^+) = 0.5$.

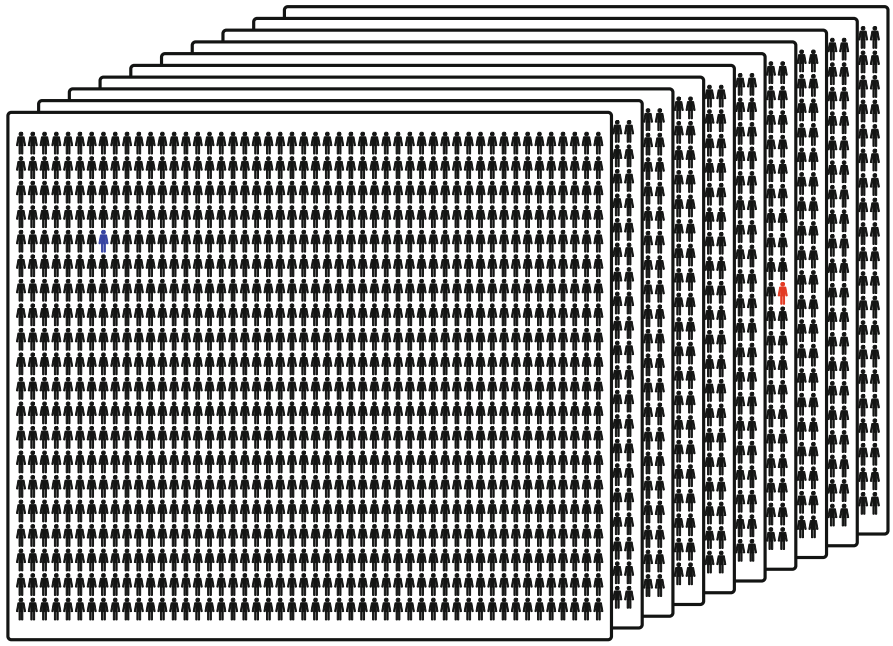


Fig. 1.1 In a sample of 10,000 people, there is likely one infected person. Because the test has a high sensitivity, the test result for this person is very likely positive (red person). If we test one arbitrary non-infected person, the test result is very likely negative because the test has high specificity. However, there are 9999 non-infected persons and, even though specificity is high, there likely is one false positive result (blue person). Hence, the test is twice positive and since only one person is infected, the probability to be infected when the test is positive is 1/2: $P(HIV|T^+) = 0.5$. It is obvious that we cannot ignore the incidence rate. It is as important as the sensitivity and specificity

Let us assume that the incidence rate is even lower, for example 1/100,000. Let us test 100,000 people. Because the incidence rate is 1/100,000, there is likely one infected person and the test likely detects the infected person. In addition, for each 10,000 people tested, there is one false positive. Hence, the test is 11 times positive[1] and the chance of being actually infected if the test is positive drops to $P(HIV|T^+) = 1/11 \approx 0.1$. On the other hand, for an incidence rate of 0.5, $P(HIV|T^+) = 0.9999$, i.e., almost 1.0. Hence, the probability $P(HIV|T^+)$ depends on the sensitivity, specificity, *and* the incidence rate. When the incidence rate changes from 0.0 to 1.0, $P(HIV|T^+)$ varies from 0.0 to 1.0. One needs to know all three terms for an educated conclusion. If any of these terms is missing,

[1] Alternatively, we may test 10,000 people. Since the incidence rate is 1/100,000, the chance is 0.1 that the sample contains one infected person. As for the incidence rate of 1/10,000, there is one false positive. Hence, we need to divide 0.1/1.1, which leads to the same result of about 0.1.

then conclusions are void. This example presents one of the main themes of this book: *Be aware of partial information!*

Comment 1 The above demonstration shows how important it is to understand basic statistical reasoning. For a patient, it matters a great deal to know how to interpret a positive outcome of a medical test. For example, in 1987, 22 recipients of a blood transfusion received a positive HIV Test statement and seven committed suicide [1]. Similarly, in one study, 16/20 German doctors told patients that there are virtually no false positives in the HIV test [2].

Comment 2 Importantly, when you are a doctor, the situation is different than in the example above because it is more likely that people who have reason to worry about being infected take the test than people who are rather sure that they are not infected. Hence, the incidence rate in a hospital is likely higher than in the above example. This larger rate means that $P(HIV|T^{+})$ may be larger than 0.5: This is a puzzling conclusion, which shows why people often have poor intuitions about statistics.

1.1.2 A Second Test

An understanding of probability also provides guidance on how to obtain better information. What happens, if we do the test a second time on only the two positively tested persons[2]? What is now the probability to be infected when the test is positive? Hence, we are looking for

$$\begin{aligned} P(HIV|T^{2+}) &= \frac{0.9999^2 \times 0.0001}{0.9999^2 \times 0.0001 + (1-0.9999)^2 \times 0.9999} \\ &= \frac{0.9999}{0.9999 + 0.0001} \\ &= 0.9999 \end{aligned}$$

A positive result now indicates that a person is almost surely infected.

This equation can be intuitively explained. The first test led to two positive results. Only these two persons are tested a second time. Since the test is so good, the test almost always detects the infected person and almost always delivers a conclusion of no infection for the non-infected person. Hence, a person who twice tests positive has a probability of infection that is close to 1.0.

[2]We assume that the tests are independent for a given person.

Comment 1 In reality, when doctors run the test a second time they discover that $P(HIV|T^{2+})$ is lower than 0.9999. The reason is that certain people show a consistent positive result even though they are not infected. Some molecules of these people seem to be similar to the anti-bodies that the HIV test is sensitive to.

Comment 2 Confusions about statistics occur in all fields. Coralie Colmez and Leila Schneps have dedicated an entire book "Math on Trial" to law court cases. The book shows how simple misunderstandings about statistics can lead (and have led) to wrong legal conclusions. The book reports the case of the student Amanda Knox, who was accused of killing a flatmate. A genetic analysis was carried out that gave some evidence that the flatmate was killed with a knife that Amanda's finger prints were on. When the judged learned how low the probability is that the test delivers a clear cut result, he decided to not go for a second analysis—even though, as just shown above, the second analysis would have made a big difference. The judge was simply not sufficiently educated in basic statistics [3].

1.1.3 One More Example: Guillain-Barré Syndrome

Vaccination (V) against swine flu (SF) may cause Guillain-Barré syndrome (GB) as a side effect in one out of a million cases, i.e., $P(GB|V) = \frac{1}{1{,}000{,}000}$. In severe cases, the GB syndrome resembles a locked-in syndrome, where patients are fully immobile and even unable to speak. Given how horrible GB can be, should we really go for vaccination? Again, we cannot yet properly answer the question because we have only partial information. We need to know the probability of acquiring Guillain-Barré syndrome without the vaccine ($\neg V$). Let us suppose, for the sake of the argument, that other than the vaccine, GB only comes from the swine flu (further, let us suppose that the vaccine is fully effective at preventing the swine flu). The probability of acquiring the Guillain-Barré syndrome from the swine flu is quite high: 1/3000. A probability of 1/3000 is much higher than a probability of 1/1,000,000. Thus, it is seems the vaccine is much better. However, we need to take the infection rate of swine flu into the account since not everyone gets infected. This rate varies from epidemic to epidemic; suppose it is: 1/300 for a random unvaccinated person. Thus, the probability of an unvaccinated person acquiring Guillain-Barré syndrome is:

$$P(GB|\neg V) = P(GB|SF) \times P(SF|\neg V) \times P(\neg V) = \frac{1}{3000} \times \frac{1}{300} \times 1 = \frac{1}{900{,}000} \tag{1.1}$$

Thus, in this situation, the probability of an unvaccinated person acquiring Guillain-Barré syndrome is a bit higher than for a vaccinated person. In addition, the vaccine has an added benefit of protection against the swine flu.

The key point here is that one cannot make a good decision based on just a single probability (of contracting Guillain-Barré syndrome from a vaccine). You have to also consider the probability with the complement (of contracting Guillain-Barré syndrome without the vaccine).

1.2 Confusions About Basic Probabilities: The Odds Ratio

1.2.1 Basics About Odds Ratios (OR)

Many smokers die because of heart attacks. Quit smoking? This is partial information! Counter question: How many non-smokers die because of heart attacks? Without this information, an answer to the first question is as good as: 100% of smokers will die once—as do 100% of non-smokers.

We summarize this effect by describing the odds. As a hypothetical example, out of 107 smokers seven suffer from a heart attack, i.e., 100 do not suffer (Table 1.1A). The odds are the ratio $\frac{7}{100}$. For the non-smokers, there is only 1 out of 101 and we compute the odds $\frac{1}{100}$. The idea of the Odds Ratio (OR) is to compare the two fractions by dividing them. The ratio of the two ratios tells us to what extent smokers suffer from heart attacks more often than non-smokers: $\frac{7/100}{1/100} = 7$. Thus, the odds for a smoker to suffer from a heart attack is about seven times higher than for a non-smoker, which seems substantial. As a comparison, if there is no effect, i.e, smokers suffer from heart attacks as often as non-smokers, the OR = 1.0.

Table 1.1 A hypothetical example

A	Smokers	Non-smokers	B	Smokers	Non-smokers
Heart attack	7	1	Heart attack	7	1
No heart attack	100	100	No heart attack	10000	10000

A) What are the odds to suffer from a heart attack when being a smoker? Assume out of 107 smokers, seven suffered from a heart attack. Out of 101 non-smokers, it was only one. Thus, how much higher is the odds of a smoker to suffer from a heart attack compared to a non-smoker? The Odds Ratio first divides 7/100 and 1/100 and then divides these two ratios: (7/100)/(1/100)= (7*100)/(1*100) = 7/1= 7. Hence, the odds is seven times higher, which seems substantial. B) Let us now assume there are 10,000 people without a heart attack in the smoker and non-smoker groups. The Odds Ratio is (7/10,000)/(1/10,000) = 7/1 = 7, i.e., the odds are the same as in the case before. Hence, the Odds Ratio is independent of the incidence rate. However, the likelihood to suffer from a heart attack has decreased by about a factor of 100. It makes a real difference whether the probability to suffer from a heart attack is 7 in 107 or 7 in 10,007 cases. Importantly, the Odds Ratio provides only partial information!

Table 1.2 The terms that contribute to the Odds Ratio[a]

	With risk factor	Without risk factor
Ill	a	b
Not ill	c	d

[a] As a small comment: the Odds Ratio divides $\frac{a}{b}$, $\frac{c}{d}$ and takes the ratio of the two. One could also use the proportions $\frac{a}{a+b}$, $\frac{c}{c+d}$ and take the ratio of the two

In its general form (Table 1.2), the Odds Ratio is $\frac{a/c}{b/d} = \frac{a*d}{b*c}$.

The OR is a very compact way to compare an experimental and a control condition and, indeed, the OR is one of the most frequently used measures in medicine and biology. For example, the impact of a gene on a disease is usually expressed in terms of the OR. However, decisions based on the OR are based on partial information. Here is why. Let us increase the number of people without a heart attack in both groups by a factor of 100. The Odds Ratio has not changed (Table 1.1B).

Obviously, Odds Ratios are independent of the rate of non-affected people even when the likelihood to suffer from a heart attack has substantially changed. Since the OR is incidence rate independent, a high OR is of almost no relevance if, for example, a disease is rare.

How to read ORs? First, a high OR is a reason to worry only if also the main effect, a/c, is large. For example, $\frac{\#smokers\ with\ heart\ attack}{\#smokers\ without\ heart\ attack} = 7/10{,}000$ is not an especially large effect even though an OR of seven is substantial. In the example of Table 1.1B heart attacks are simply not very frequent. Only 8 out of 20,008 have an attack. Thus, it is very unlikely for someone to suffer from heart attacks at all, contrary to the case of Table 1.1A, where 8 out of 208 people suffer. In the latter case, one may worry. Second, a high main effect a/c is only a reason to worry when also the OR is high. Here is an extreme example. If you have blue eyes your probability of dying is very high (100%). However, the probability of dying for brown eyed people is also 100%. Hence, the OR $= 1.0$, which is low. One may worry to die but one should not worry about eye color.

1.2.2 Partial Information and the World of Disease

The overall situation can be even more complicated. We have discussed the effect of one factor (smoking) on one outcome (heart attack). However, smoking may also affect other diseases in a positive or negative way (even smoking is not always deleterious). Hence, to formally answer the question whether one should quit smoking, one needs to take all diseases into account, including potentially unknown ones. In addition, one needs to take the costs into account because tooth decay is less severe than a heart attack. Thus, one would want to compute something like a morbidity effect that considers the cost of

different diseases and the probability of those diseases for the given factor:

$$\text{Morbidity(Factor)} = \sum_S P(\text{disease } S|\text{Factor}) \times \text{Cost(disease } S)$$

Hence, one needs to take all diseases into account, even diseases that are waiting to be discovered. Hence, whether to quite smoking or change a diet is almost impossible to determine, unless effect sizes are large. In practice one never has all of this information available. This does not mean that one can never use statistical reasoning to guide decisions, but it does indicate that one should be aware that decisions are based on partial information. Such knowledge should motivate you to get as much information as is practical.

Take Home Messages

1. Be aware of partial information and make sure you have full information for proper conclusions. For example, the Odds Ratio usually provides too little information.
2. Incidence rates of diseases are usually low, except for diseases such as tooth decay.

References

1. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz L, Woloshin S. Glaub keiner Statistik, die du nicht verstanden hast. Geist und Gehirn. 2009;10:34–39.
2. Gigerenzer G, Hoffrage U, Ebert A. AIDS counselling for low-risk clients. AIDS Care. 1998;10:197–211.
3. Colmez C, Schneps L. Math on trial: how numbers get used and abused in the courtroom. Basic Books: New York; 2013.

Experimental Design and the Basics of Statistics: Signal Detection Theory (SDT)

2

Contents

What You Will Learn in This Chapter

What is a good performance measure? The most frequently used measure is the percentage of correct responses. Here, we will see that percent correct confuses two variables, namely, sensitivity and criterion, and should therefore be used with care. We will introduce the sensitivity measure d', which turns out to be a crucial term in much of statistics.

2.1 The Classic Scenario of SDT

Assume we are in a yellow submarine cruising through the ocean. It can be quite dangerous to hit a rock and for this reason, the submarine is equipped with a sonar device. Sonar waves are emitted and their reflections are recorded by a receiver. These reflections are combined to form what we call the "sonar measure." If there is a rock, the sonar measure is larger than when there is no rock. However, the situation is noisy and hence even under the very same rock or no-rock condition the sonar measure varies quite a bit across recordings (Fig. 2.1).

A probability distribution corresponds to each of the two conditions, rock vs. no-rock, that indicates how likely it is that a certain value of the sonar measure, indicated on the

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_2

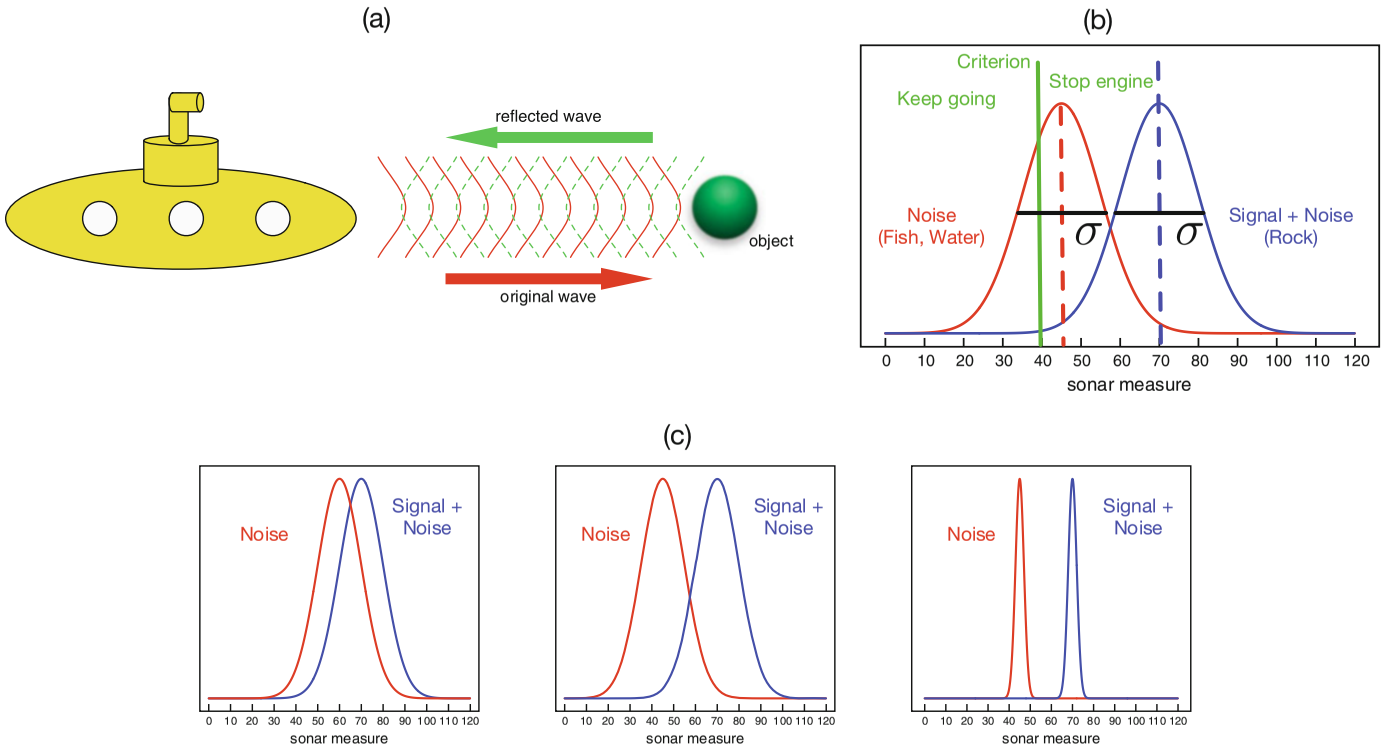


Fig. 2.1 (**a**) A submarine sends out sonar signals and measures a response based on the echoes. The sonar measure is noisy, i.e., for the very same rock, the echoes can differ from sonar measure to sonar measure. For example, fish may change the reflectance. The same is true when there is no rock. How to decide whether or not there is a rock? (**b**) The classic scenario of SDT. To both the rock and no-rock condition belongs a probability distribution, which shows how likely it is to receive a sonar measure at various values. SDT assumes that these probability distributions are Gaussians.

Fig. 2.1 (continued) In this example, the mean value of the sonar measure is 70 when a rock is present and 45 when there is no rock. The spread of a distribution is described by its standard deviation σ. Here, $\sigma = 10$. If we record a sonar measure of 80 it is much more likely that there is a rock than not. A sonar measure of 57.5 is equally likely from a situation with a rock and a situation without a rock. To make a decision, a criterion is needed. When the sonar measure is larger than the criterion, we decide to stop the engine (conclude there is a rock), otherwise we continue cruising (conclude there is no rock). Where to set the criterion is our choice. If we are conservative, we set the criterion to low values, when we are more risky we set it to higher values. (**c**) How well we can discriminate between the rock and no-rock conditions depends on the overlap between the two probability distributions. The overlap is denoted by d', which is the difference between the mean values divided by the standard deviation. A large overlap means low discriminability, a low overlap means high discriminability. For a fixed standard deviation, d' increases when the mean difference increases (left vs. center figure). For a fixed mean value difference, discriminability increases when the standard deviation decreases (center vs. right figure)

x-axis, is elicited.[1] As is often the case in statistics, we assume that Gaussian distributions adequately describe the situation. A Gaussian distribution is fully determined by its mean value μ and the standard deviation σ, which determines the width of the Gaussian. A large σ means that for the very same rock condition, very different signals can be reflected with a high probability. As the other extreme, if $\sigma = 0$, there is no variability. We always receive the same value of the sonar measure. Hence, σ reflects how noisy the situation is and that is why σ is also called the noise. We call the no-rock condition the *noise alone condition* because there is no signal from the rock and we call the rock condition the *signal plus noise* condition.

How well can we distinguish between the rock and no rock conditions? It depends on the overlap between the Gaussians. If the Gaussians fully overlap, we cannot discriminate between the two cases. The sonar is useless. When there is almost no-overlap, it is easy to discriminate the two situations because a certain sonar measure value originates very likely from only one of the Gaussians. For example, Fig. 2.1b indicates that a sonar measure of 80 occurs from the rock condition with a much higher likelihood than from the no-rock condition. The overlap can be captured by the difference between the means, μ_1 and μ_2, of the two Gaussians divided by the standard deviation σ, which is assumed to be the same for both distributions[2]:

$$d' = \frac{\mu_1 - \mu_2}{\sigma} \tag{2.2}$$

d' is called sensitivity or discriminability and it measures how well we can discriminate between two alternatives, i.e., d' is a measure of the signal $(\mu_1 - \mu_2)$ to noise (σ) ratio. The overlap depends on both the difference of the means and the standard deviation. Hence, d' can be increased by both increasing the mean value difference or decreasing the standard deviation (Fig. 2.1c). Importantly, the notion of sensitivity here is different from the notion in Chap. 1, where it is identical with the Hit rate. In the following, we will use sensitivity only for Hit rate and not d'.

When should we stop the engine? A decision criterion c is needed. In Fig. 2.1b, the criterion is set at 40, i.e., if the sonar measure is larger than 40, we stop the engine, if

[1] Strictly speaking, the probability is zero that the sonar measure is *exactly* 80. The probability function shows the probability that a value close to 80 occurs (a value from a very narrow interval around 80). Whereas these aspects are crucial for mathematical statistics, they hardly play a role for the basic understanding of statistics.

[2] In Signal Detection Theory (SDT), d' is usually defined by the absolute value of the difference of the means:

$$d' = |\frac{\mu_1 - \mu_2}{\sigma}| \tag{2.1}$$

We use the definition without the absolute values because it is better suited when we apply d' to statistics in Chap. 3.

the sonar measure is smaller than 40, we continue cruising. Where we set the criterion is our choice. If a rock is as likely as a no-rock situation, then the optimal criterion is at the intersection point of the two Gaussians, in the sense that it maximizes the number of correct decisions.

2.2 SDT and the Percentage of Correct Responses

Let us apply SDT to a behavioral experiment. Consider a typical detection experiment. You are looking at a computer screen and either a faint light patch is presented (stimulus present) or the screen is blank (stimulus absent). As in the yellow submarine and HIV test examples, there are four possible outcomes (Table 2.1).

If the stimulus is presented in half of the trials, the percentage of correct responses is computed by the average of the Hit rate and Correct Rejection rate: $\frac{Hit+CR}{2}$. Let us consider "percent correct" in terms of SDT. As in the submarine example, we assume that decision making is noisy. Contrary to the submarine example, we do not know how the perceptual processes are coded in the human brain, i.e., we do not explicitly know the probability distributions. The good thing about SDT is that it is very flexible. For example, we assume that we can focus on one neuron, which codes for the brightness of the stimuli. A value of 0.0 corresponds to a blank screen and a positive value to a light patch with a certain brightness. We use a decision criterion that determines whether we decide for the light patch or the blank. If we are more conservative, we set the criterion to a higher value, i.e., we respond 'light patch present' only when we are quite sure. If we are more risky, we set the criterion to a lower value, i.e., respond 'light patch present' when there is the slightest evidence for a light being present. We can set the criterion to any value we wish. For example to optimise the percentage of correct responses, we can set the criterion at

Table 2.1 There are four outcomes in a classic SDT experiment

Response	Stimulus present	Stimulus absent
Present	Hit	False Alarm (FA)
Absent	Miss	Correct Rejection (CR)

A rock/light patch is present and we decide a rock/light patch is present (Hit). A rock/light patch is present and we decide no rock/light patch is present (Miss). No rock/light patch is present and we decide a rock/light patch is present (False Alarm). No rock/light patch is present and we decide no rock/light patch is present (Correct Rejection)

the intersection of the Gaussians. In this case, we respond equally often for 'light patch present' and 'light patch absent' when both stimulus alternatives are presented with the same frequency. If there are more patch absent than patch present situations, we may want to move the criterion towards more "patch absent" responses, i.e., towards the right in this case. In addition, we may want to take the costs of our decision into account. If a "patch absent" response is less rewarded than a "patch present" response, we may want to move the criterion so that more "patch present" responses occur.

Let us change the criterion smoothly and see how the percentage of correct responses changes with d' (Fig. 2.2). We start with a criterion set to 2.0, i.e., at the intersection point. If we now move the criterion a bit from optimal, e.g., to the right, the percentage

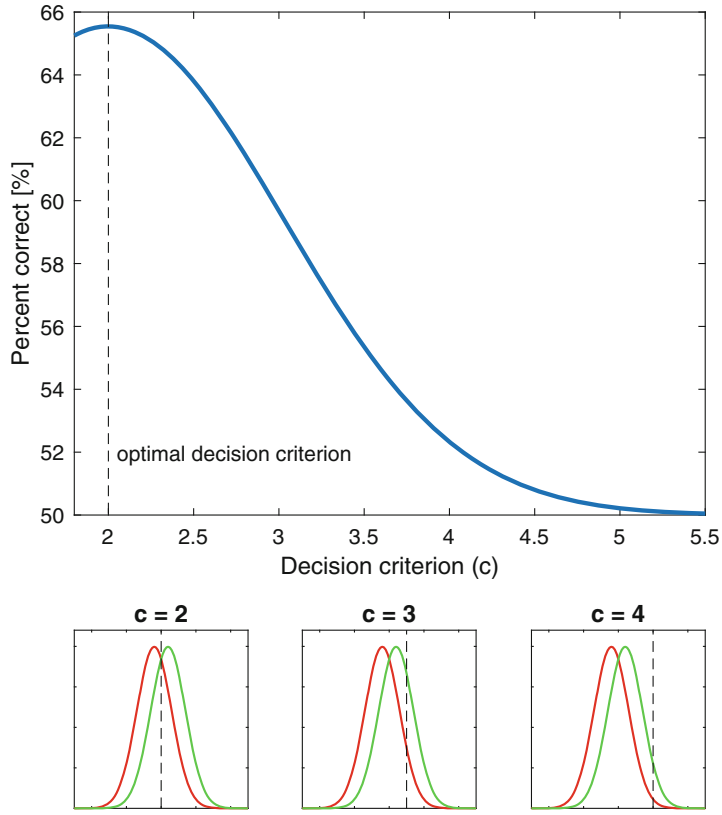


Fig. 2.2 Percent correct depends on the criterion. Let us, first, put the criterion at the intersection of the two Gaussians, which we set at the value 2 (lower left panel). Let us now move the criterion to the right (center and right). The upper panel shows that the percentage of correct responses decreases the further we move the criterion from optimal. If the criterion is set to the most right, we almost always respond for one response alternative, such as "patch absent." In this case, the Hit rate is 0 but the Correct Rejection rate is 1. Hence, percent correct is: $(0 + 1)/2 = 0.5$, i.e., we are at chance level, even though we can, in principle, well discriminate between the two situations

of correct responses deteriorates and it does so the more we move the criterion. If we move the criterion to the far right, performance approaches 50%, i.e., chance level. In this case we always give the same response, say, "light patch absent", i.e., we have a large *response bias*. Importantly, the discriminability d' has not changed, i.e., our visual abilities are constant. Only our decision behavior has changed.

Hence, the percentage of correct responses confounds decision criterion and discriminability. For a certain performance level, say 75%, we cannot tell whether there is high discriminability and a suboptimal criterion or an optimal criterion with a lower discriminability. For this reason, the percentage of correct responses can be a dangerous measure, which can easily obliterate true effects or lead to false positives. Conclusions based on the percentage of correct responses are based on partial information!

2.3 The Empirical d'

From only the percentage of correct responses, it is impossible to infer the discriminability in an experiment. Aren't then all experiments hopeless? Surprisingly, one can disentangle discriminability and criterion by separately estimating d' and b, the *bias* of the criterion:

$$d'_{emp} = z(Hit) - z(FA) \tag{2.3}$$

$$b_{emp} = -\frac{z(Hit) + z(FA)}{2} \tag{2.4}$$

To compute d', we simply need to z-transform the Hit and False Alarm rate. The z-transformation is the inverse cumulative Gaussian function. If you are not familiar with the z-transformation just treat it as a function you can find on your computer. b_{emp} tells you how much the current criterion differs from the optimal one, i.e., how far it is from the intersection point of the Gaussians. Hence, b_{emp} measures the bias of the criterion.

Importantly, d'_{emp} does not change when the criterion (and thus response bias) changes. However, d' is criterion free but not model free. There are three assumptions:

1. The underlying probability distributions are Gaussians.
2. The Gaussians have the same variance.
3. The criterion is not changed during the measurements.

Assumption 1 is crucial since we compute the z-transform, i.e., the inverse Gaussian distribution, which only makes sense when the data are Gaussian distributed. Assumption 1 is often fulfilled. Assumption 2 is usually fulfilled since the stimulus alternatives are similar. Assumption 3 is crucial and is not easy to check.

Attention
The term "sensitivity" is used in two different ways:

1. Mainly in the medical literature, sensitivity is the same as Hit Rate.
2. In SDT, sensitivity corresponds to discriminability, i.e., $d' = z(Hit) - z(FA)$. We will use the term discriminability rather than sensitivity in the context of SDT.

Example 1 (Automated System) Performance (Hit, FA, Miss, CR) of both a doctor and an artificial intelligence system (AI) for diagnosing a disease is shown in Fig. 2.3. The overall percentage of correct responses is 80% for the two. d' is almost as simple to compute as percent correct: one only needs to z-transform Hit and FA. As it turns out, performance in terms of d', contrary to the percentage correct, is strongly different. Is the doctor or the AI system better? Usually, discriminability is an inbuilt, hard to change, characteristic of a system. Your eyes are as good as they are. On the other hand, changing the decision criterion is easy, one just needs to respond more often for one than the other alternative. Obviously, the AI system is strongly biased towards "yes" responses, thereby avoiding misses but leading to a higher false alarm rate. Hence, the AI's system criterion is far from being optimal. Setting the criterion to optimal strongly increases performance in terms of percent correct.

Doctor's performance			Automated recognition		
Signal	Present	Absent	Signal	Present	Absent
Yes	80	20	Yes	98	38
No	20	80	No	2	62
	P	z		P	z
Hit	0.8	0.842	Hit	0.98	2.054
FA	0.2	-0.842	FA	0.38	-0.305
Sensitivity, d'	1.683		Sensitivity, d'	2.359	
Bias, b	0.000		Bias, b	-0.874	
P(correct)	0.800		P(correct)	0.800	

Fig. 2.3 Doctor vs. Machine. The percentage of correct responses is identical for both the doctor and the machine. d' is almost as simple to compute as percent correct: one only needs to z-transform Hit and FA and subtract. Performance in terms of d', i.e, discriminability, is strongly different. Is the doctor or the AI system better? Obviously, the AI system is strongly biased towards "yes" responses, which avoids Misses but also leads to a higher False Alarm rate. The criterion for the AI system is far from optimal. Courtesy: Mark Georgeson

Example 2 (Learning) In a learning experiment, observers are shown either a left or right tilted line. Since the difference is small, observers perform poorly. To improve performance, observers train on the task with 10 blocks, each containing 80 trials. We consider the number of correct responses for each of the 10 blocks, by averaging the 80 trials. Performance improves strongly. Does this imply that perception has improved? Improvements can be caused by changes in both discriminability or criterion. For example, training with the stimuli may lead to a decrease of the variance σ of the Gaussians, i.e., people can more clearly discriminate the tilt of the lines. A decrease of the variance leads to an increase in d', i.e., an increase in discriminability (Fig. 2.2). An increase in discriminability can also occur when the means of the Gaussians are pulled apart. Performance can also improve when the decision criterion of the participants is not optimal in block 1. During training, participants may learn to adjust the criterion. A change of *perception* is generally considered to be related to a change of discriminability. When we analyze the data with the percentage of correct responses, we cannot make proper conclusions since we cannot disentangle changes in discriminability from changes in criterion. Hence, for all learning experiments, it is important to plot the results as d' and $bias$.

Example 3 (Sensitivity and Specificity) In Chap. 1, the HIV test had a very high sensitivity and specificity. Determining sensitivity and specificity also depends on a criterion. In fact, it is in no way different than determining percent correct. Recall that sensitivity is Hit rate and specificity is the rate of Correct rejections. Thus, the situation is exactly the same as in the above example with the submarine, except that on the x-axis we have HIV anti-body concentration (as measured by the test). We need a criterion that determines whether or not the test is positive for a certain antibody concentration. Hence, we can increase sensitivity at the cost of specificity and vice versa. Just to mention, (sensitivity + specificity)/2 is percent correct.

Example 4 (Speed-Accuracy Trade-Off) In many experiments, responses are speeded, i.e., observers need to respond as fast as possible. Often, slow observers, e.g., older people, have a higher d' than fast observers, e.g., younger people; a so called speed-accuracy trade-off. So the situation is even more complicated because we need to pit Reaction Times against d' and bias to reach proper conclusions. Experiments with a clear speed-accuracy trade-off are often hard to interpret.

Example 5 (Floor and Ceiling Effects) One additional issue involves so called floor and ceiling effects. In a (non-sense) experiment, the experimenter holds up one hand, with all fingers clearly visible. When asked, all observers correctly identify the 5 fingers, i.e., 100% correct responses. Can we conclude that all observers have the same good eyes, i.e., the same discriminability? Of course not; the task was too simple and for this reason observers performed in a ceiling regime, i.e., close to 100%. Conclusions are useless. Computing d' does not help in this situation because the false alarm rate is 0.0 and d' is infinity.

The very same is true for floor effects, i.e., when performance is close to chance level (50%). It is therefore important to make sure that stimulus alternatives are in a range where the differences between participants can be detected.

Example 6 (Standardized Effects) d' is also often called a standardized effect because the division by σ converts the measurement to units of standard deviation. As a result, d' is insensitive to the original units (e.g., it does not matter whether the original measurements are in meters or inches). Moreover, a standardized effect size can also be insensitive to some experimental variations. For example, if a reaction time experiment can be manipulated to make everything slower by a common factor, then both the difference of means will increase (signal) and the standard deviation will increase (noise) by an equal factor. The ratio, d', will be unchanged.

Take Home Messages

1. Be aware of partial information. The percentage of correct responses confounds discriminability d' and decision criterion c.
2. Be aware of partial information. The same is true for many other measures such as Sensitivity and Specificity in medical tests.
3. You can disentangle discriminability and criterion by using d'_{emp}.
4. d'_{emp} is criterion free but not model free. There is no free lunch.

The Core Concept of Statistics

3

Contents

What You Will Learn in This Chapter

In Chaps. 1 and 2, we showed that proper conclusions need full information and that many popular measures, such as the Odds Ratio or the Percentage of Correct Responses, provide only partial information. In this chapter, we use the framework of SDT to understand statistical inference, including the role of the *p*-value, a dominant term in statistics. As we show, the *p*-value confounds effect size and sample size and, hence, also provides only partial information.

This chapter is about the essentials of statistics. We explain these essentials with the example of the *t*-test, which is the most popular and basic statistical test. This is the only chapter of this book where we go into details because, we think, the details help a great deal in understanding the fundamental aspects of statistics. Still, only basic math knowledge is required. The hasty or math phobic reader can go directly to Sect. 3.3 *Summary*, where we summarize the main findings and the key steps. Understanding at least this Summary is necessary to understand the rest of the book.

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_3

3.1 Another Way to Estimate the Signal-to-Noise Ratio

In Chap. 2 we defined d' as the distance between population distribution means, divided by the population standard deviation. Typically, we label one of the populations as noise-alone, with mean μ_N, and the other population as signal-and-noise, with mean μ_{SN}. In statistics, the d' of populations is also often referred to as Cohen's δ or effect size. The calculation is the same as in Chap. 2:

$$\delta = d' = \frac{\mu_{SN} - \mu_N}{\sigma} \tag{3.1}$$

Oftentimes we do not have population information, and we want to estimate δ (i.e., d') from empirical data. For example in Chap. 2, we could estimate d' in the patch present vs. absent experiment by computing z(Hit)-z(FA) just from the behavioral data. We did not have any knowledge about the underlying means and variance of the Gaussians. This estimation approach is useful when we cannot directly measure the underlying variables driving the performance of the system but we can measure decision outcomes. In other situations we cannot easily measure decision outcomes but we can estimate the means and variance of the Gaussians directly. For example, we can use a sonar and record the sonar responses in many trials when a rock is present. We plot the results and obtain a graph, from which we can estimate the mean and variance of the Gaussian. We can obtain the mean and variance also for the no-rock condition. Next, the sample means $\overline{x}_{SN}$ and $\overline{x}_N$ and the standard deviation s can be used to compute an estimated effect size called Cohen's d:

$$d = \frac{\overline{x}_{SN} - \overline{x}_N}{s} \tag{3.2}$$

Once again, this standardized effect d is simply an estimate of the d' of the population distributions. If d is large, it will be fairly easy to distinguish a single measurement as coming from the signal-and-noise distribution or from the noise-alone distribution. Regrettably, for many situations that scientists care about the value of d is quite small. For example, within psychology, a value of around $d = 0.8$ is considered to be "large," a value around $d = 0.5$ is considered to be "medium," and a value around $d = 0.2$ is considered to be "small." As Fig. 3.1 shows, even "large" values of d correspond to considerable overlap between distributions. For situations like these we are never going to have a good ability to correctly discriminate *single* measurements. All is not lost, though, as long as you are willing to discriminate the *means* of those measurements. As we will see, the properties of SDT apply for discriminating means similarly to discriminating single measurements.

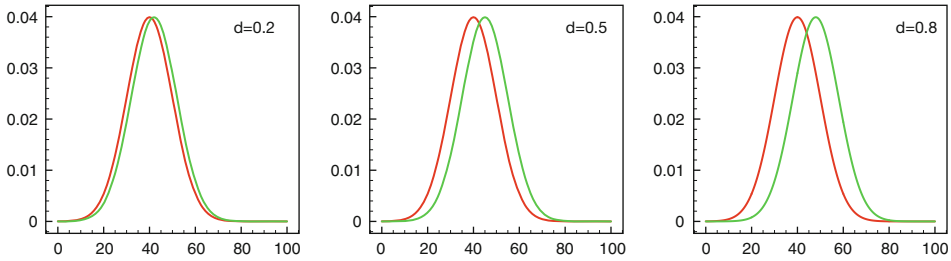


Fig. 3.1 Population distributions with small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) effect sizes

Terms

Because very similar concepts were developed in many fields, there are many identical terms, which we list here:

- Hit Rate = Power
- False Positive Rate = False Alarm = Type I error
- Miss Rate = Type II error
- d' = Cohen's δ = effect size = standardized effect size
- Gaussian distribution = Normal distribution = Bell curve
- Sample values, such as tree height, are also called Scores

Some definitions

We collected a sample of n scores x_i

- Sample mean
  $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
  where the symbol $\sum$ means "add up all following terms"
- Sample variance
  $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$
- Sample standard deviation
  $s = \sqrt{s^2}$
- Standard error
  $s_{\overline{x}} = s/\sqrt{n}$

Facts about sample means
For $n \to \infty$:

1. The distribution of sample means $\overline{x}$ is Gaussian (Central Limit Theorem; CLT)
2. $\overline{x} \to \mu$
3. $s_{\overline{x}} \to 0$

3.2 Undersampling

Let us start with an example. We are interested in the hypothesis that the mean height of Alpine oak trees at the Northern rim is different than the mean height of oaks from the Southern rim. The straightforward way to address this hypothesis is to measure the height of *all* trees in the North and South rims, compute the means, and compare them. If the means are different, they are different. If they are the same, they are the same. It is that easy.

Fig. 3.2 A small section of a forest in the Swiss Alps

Unfortunately, there are many oaks (Fig. 3.2) and our resources are limited, so we can measure only a certain number, n, of both the North and South trees and measure their heights. The collected trees of a population are called a *sample*, i.e., in this case we collected two samples, one from the North and one from the South trees. The tree heights we measured are also called *scores* and the mean height in a sample is called the *sample mean*. The number n of trees of a sample is called the *sample size*. For each sample, there is most likely a difference between the *mean* height of the *sampled* trees and the *true* mean height of *all* trees of the population. For example, we may have, just by chance, sampled more large than small trees. This difference is called the *sampling error*. Thus, because of *undersampling* (measuring fewer than all trees), we likely do not obtain accurate estimates of the two means. Importantly, we choose the trees for our samples randomly, a procedure called random sampling.

Let us now collect both a sample from the North and a sample from the South trees. If we find a difference of sample means we cannot know whether it was caused by a true difference of the tree population means or whether the population mean heights were the same but the difference was caused by undersampling. Hence, undersampling may lead to wrong conclusions. For example, even though there is no difference between the means for the North and South tree populations, we may decide that there is a difference because there was a difference in the sample means. In this case, we are making a False Alarm, also called a Type I error.

To understand how undersampling influences decisions, we will first study how likely it is that a sample mean deviates from the true mean by a certain amount. As we will see, the sampling error is determined by the standard deviation of the population, σ, and the sample size, n. Second, we will study how undersampling affects how well we can discriminate whether or not there is a difference in the mean height of the *two* tree populations. A simple equation gives the answer. The equation is nothing else but a d for mean values. Hence, we are in a SDT situation. Third, we want to control the Type I error rate. We will see that the famous p-value just sets a criterion for the Type I error.

3.2.1 Sampling Distribution of a Mean

To begin with, let us focus on the North tree population. To arrive at a sample mean, we collect a sample of North trees, measure the height x_i for each tree, sum these heights up, and divide by the sample size n. The sample mean is an estimate of the true mean μ_{North}:

$$\bar{x}_{North} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.3}$$

where x_i is the height of the i-th tree we sampled. Similarly, we can estimate the variance of the tree heights, s^2 (see box). The difference between this sampled mean and the true mean is

$$\bar{x}_{North} - \mu_{North} \tag{3.4}$$

How large is this difference *on average*? To answer this question, assume—for the sake of the argument—we are going many times to the woods and randomly choose a sample of a fixed sample size, n. The various samples most likely contain different trees because we chose the trees randomly. How many sample means are close to the true mean and how many are far from it? Let us assume that we know the true population mean and standard deviation. Let us first collect a sample of only two trees and compute the mean. In the example of Fig. 3.3, the true mean is 20 m and the mean of the sample is 19 m. Thus, the difference is 1 m. Let us collect another sample of two trees. This error is likely different from the previous one because we have likely chosen two trees with different heights. Let us continue measuring two trees and see how the sample means are distributed. The Central Limit Theorem tells us that the distribution of the sample means is similar to a Gaussian function. The Gaussian is centered around the true mean, thus many sample means reflect well the true mean. However, the Gaussian is quite broad, i.e., the standard deviation is large, and hence quite some sample means deviate substantially from the true mean.

Let us now collect 9 instead of 2 samples and repeat the procedure as before. Again, we obtain a Gaussian distribution, which is, however, narrower than for the sample of two trees, i.e., the standard deviation is smaller, and, thus, it is much more unlikely that the mean of a randomly chosen sample deviates strongly from the true mean. In general, for each sample size n, there is such a sampling distribution. The larger the sample size n, the smaller is the standard deviation of the sampling distribution. This is not surprising because the error is zero if we measure all trees and small if we fail to measure only a few trees. Hence, the standard deviation $\sigma_{\bar{x}}$ of the sampling distributions is a measure of how good we expect our estimate of the mean to be. $\sigma_{\bar{x}}$ is called the *standard error of the mean*, and it can be shown that $\sigma_{\bar{x}}$ is equal to the standard deviation of the true population distribution σ divided by the square root of the sample size n:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{3.5}$$

If we do not know σ, then we can estimate the standard error by using the sample estimate of the standard deviation:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{3.6}$$

The equation shows again why with larger sample sizes the sampling error becomes smaller: as $\sqrt{n}$ goes larger, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ goes to zero. $\sigma_{\bar{x}}$ depends on both n and σ. Suppose

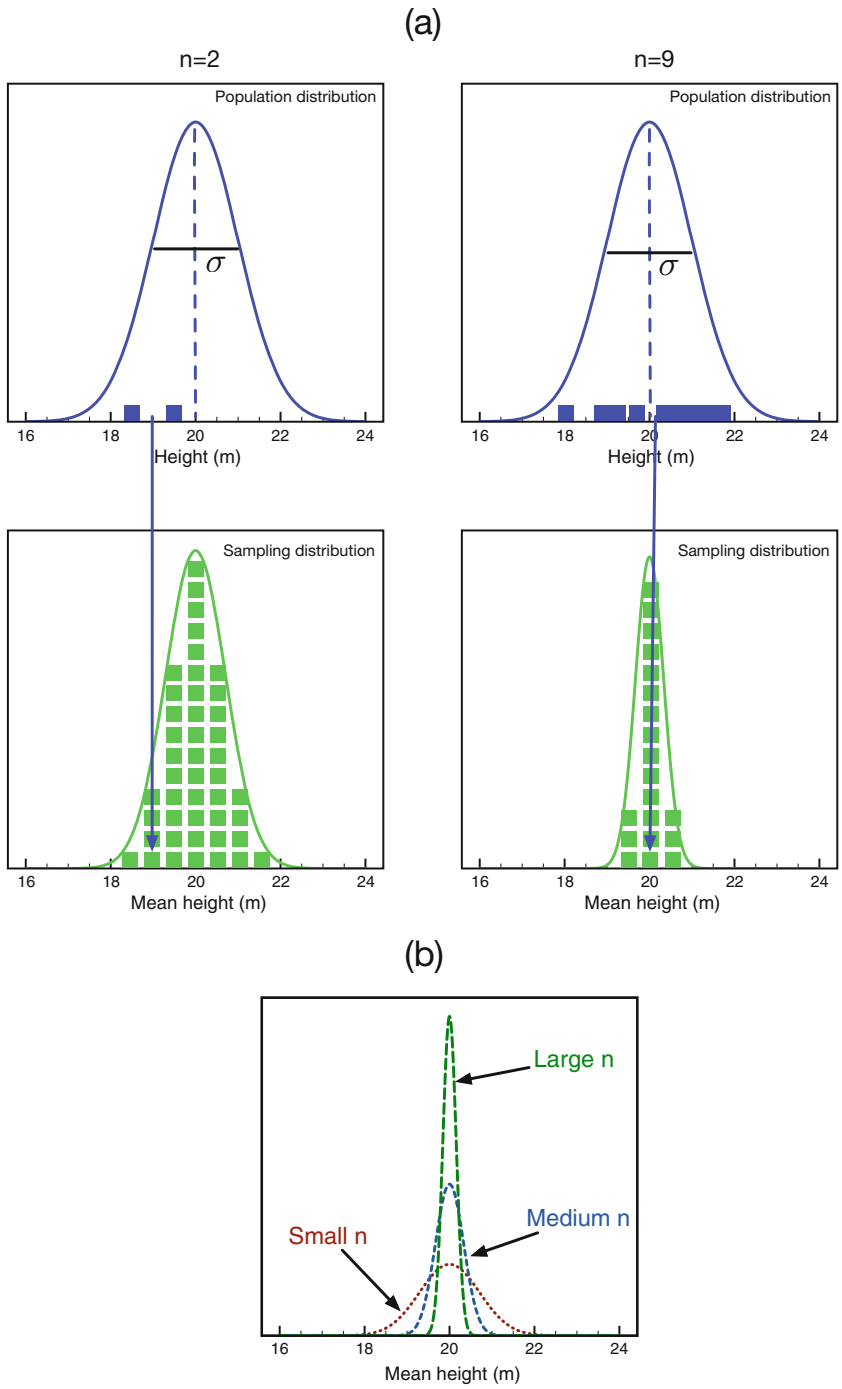


Fig. 3.3 Let us focus on the North trees only. Since we cannot measure the height of all trees, we collect samples with a mean value that, because of undersampling, is likely different from the

σ is zero, then all trees in our sample have the same height, which means all trees have the mean height μ_{North}, and hence we need to only measure the height of one tree. On the other hand, if σ is large, we need to sample many trees to obtain a good estimate of the population mean.

Summary Because of undersampling, sample means likely differ from the true mean. The standard error, $s_{\overline{x}}$, is a measure for the expected sampling error.

3.2.2 Comparing Means

Let us next see how undersampling affects a *comparison* of the means of the North and South trees. Obviously, there is a family of sampling distributions for the South trees too. In the following, we assume that the sample sizes and the population variances are the same for both tree populations. As mentioned, if our two samples contain all trees from both populations, we can simply compare the means and note any difference. For smaller samples, both sample means may strongly differ from the true means. First, we subtract the two sample means: $\overline{x}_{North} - \overline{x}_{South}$. For each pair of samples of North and South trees, we can compare the difference of sample means with the difference of the true means $\mu_{North} - \mu_{South}$. Hence, we have only one sampling distribution and the same situation as in the last subsection.

As in the last subsection, the sampling distribution is a Gaussian with a mean equal to the difference of the population means, $\mu_{North} - \mu_{South}$. Moreover, something called the "variance sum law" tells us that the standard deviation of this sampling distribution is

Fig. 3.3 (continued) true mean. (**a**) Left. The Gaussian at the top row shows the true population distribution. Let us first measure the height of only two randomly collected trees (blue boxes). The heights are 18.5 m and 19.5 m. Hence, the mean of the two samples is 19 m, shown as the green box in the graph below, highlighted by the arrow. Hence, the sampling error of the mean is 1 m since the true mean is 20 m. Let us now go to the woods and collect more samples of two trees. Each green box shows one of the corresponding sample means. After having collected many samples, we obtain a Gaussian function. The *y*-axis shows the probability how likely it is that a certain mean value occurs. Right. Now let us sample nine trees. The true population distribution is the same as on the Left and shown again on the top row along with one sample of nine trees. The corresponding sample mean is shown by the arrow. Again, if we sample more trees, we arrive at more mean values (green boxes). The Gaussian is more narrow, i.e., the standard deviation of the sample means is much smaller. Hence, whereas for samples of two trees, an error of 2 m is not unlikely, such an error is unlikely for samples of nine trees. (**b**) For any population distribution there is a family of sampling distributions, one for each sample size n. All sampling distributions are Gaussians. For increasingly large sample sizes n, we obtain sampling distributions with decreasingly smaller standard deviations. Hence, when n increases it becomes much more unlikely that the sample mean strongly differs from the true mean. This is not surprising because if we measure the height of all trees, there is no sampling error. The standard deviation is zero. If we fail to measure just a few trees, the error is very small

related to the standard deviation of the populations and the sample size:

$$\sigma_{\overline{x}_{North}-\overline{x}_{South}} = \sigma\sqrt{\frac{2}{n}} \tag{3.7}$$

This term is the *standard error* for the difference of sample means.[1] If we do not know the population means and standard deviation, we consider the estimate:

$$s_{\overline{x}_{North}-\overline{x}_{South}} = s\sqrt{\frac{2}{n}} \tag{3.8}$$

Let us recall our main question. We have collected one sample from the North trees and one sample from the South trees, respectively, with a sample size of n. Most likely there is a difference between the two sample means, i.e., $\overline{x}_{North} - \overline{x}_{South} \neq 0$. Does this difference come from undersampling even though there is no difference between the population means, or does this difference reflect a true difference in mean height? This is a classic SDT situation—just with means instead of single measurements. How well can we discriminate between the two alternatives? We can answer the question by computing the d' or Cohen's δ between the two alternatives. For the first alternative $\mu_{North} - \mu_{South} = 0$, meaning that there is no difference between the mean heights of the North and South trees, i.e., the noise alone distribution. The corresponding sampling distribution is centered around 0 since there is no difference. For the second alternative, there is a real difference and the sampling distribution is centered at $\mu_{North} - \mu_{South}$. Because we do not know the true values, we use estimates.[2]

So, we have now estimated two sampling distributions: one for when there is a real difference (signal-and-noise) with mean $\overline{x}_{North} - \overline{x}_{South}$ and one when there is no difference (noise-alone) with mean 0. Hence, we have exactly the same situation as in the yellow submarine example and estimate d' or Cohen's δ of the sampling distributions, which is usually called t, by:

$$t = \frac{(\overline{x}_{North} - \overline{x}_{South}) - (0)}{s_{\overline{x}_{North}-\overline{x}_{South}}} \tag{3.9}$$

The t-value is nothing else as a d' applied to sampling distributions. Just as for all SDT situations, if t is large, it is fairly easy to distinguish whether a difference of means comes

[1] If the samples from the North and South trees are of different sizes, then the formula is $\sigma_{\overline{x}_{North}-\overline{x}_{South}} = \sigma\sqrt{\frac{1}{n_{North}} + \frac{1}{n_{South}}}$.

[2] Typically, the value s is computed by pooling the variances from the two samples. We describe one way of doing this pooling in Sect. 3.4.

from the signal with noise distribution or from the noise-alone distribution. If t is small, then it will be difficult to determine whether there is a true difference.[3]

Let us substitute the estimate of the standard error into the t-equation:

$$t = \frac{(\overline{x}_{North} - \overline{x}_{South}) - (0)}{s_{\overline{x}_{North} - \overline{x}_{South}}} = \frac{(\overline{x}_{North} - \overline{x}_{South})}{s\sqrt{\frac{2}{n}}} = \frac{(\overline{x}_{North} - \overline{x}_{South})}{s}\sqrt{\frac{n}{2}} = d\sqrt{\frac{n}{2}} \tag{3.10}$$

By splitting up the standard error, which is the measure for the sampling error, into s and n, we see that the t-value is d (the estimated δ of the population distributions) multiplied by the square root of half the sample size.

We can interpret the t-value in two ways. First, we are interested in whether there is a real difference between the mean values. For this reason, we subtract the estimates of the two means to see how much they differ. However, a large difference is meaningless when the noise, i.e., the standard deviation is high. For this reason, as in Chap. 2, we divide by an estimate of the standard deviation, which in this case is the estimated standard deviation of the sampling distribution of the difference of means. The estimated standard deviation of the sampling distribution is the standard error:

$$s_{\overline{x}_{North} - \overline{x}_{South}} = \frac{s}{\sqrt{\frac{n}{2}}} \tag{3.11}$$

Hence, the standard error of the sampling distribution of the means combines both sources of uncertainty, namely the population standard deviation and the uncertainty from undersampling. Second, we see that the t-value is the product of the estimated d of the population distribution and a function of the sample size n. The t-value combines effect and sample size.

Summary We wanted to know whether or not two means are identical but have difficulties to decide because we have only inaccurate estimates caused by undersampling. This is a classic discrimination task, just with means instead of single measurements. The t-value, which is easy to calculate from the samples we collected, is nothing else than the estimated d' for this situation. Most importantly, the t-value is a function of the estimated effect size d and the sample size n, namely, a multiplication of the estimated effect size d and the square root of the sample size $n/2$.

[3] Following the convention of SDT, we will always interpret t as being a positive number, unless we specifically say otherwise. Should the computed value be negative, one can always just switch the means in the numerator.

3.2.3 The Type I and II Error

Undersampling may create an error, which may lead to wrong conclusions. For example, we may decide that there is a true mean difference in the height of the North and South trees—even though there is none—because there was a difference in the sample means (False Alarm). Likewise, we may decide that there is no true difference—even though there is one—because the sample mean difference was small (Miss). Following the statistical conventions, we call a False Alarm a Type I error and a Miss a Type II error. How do we cope with these errors? As we have seen in Chap. 2, False alarms and Misses depend on where we set our criterion. The same is true here and the four possible outcomes of a decision are summarized in Fig. 3.4. Commonly, people focus on a special hypothesis called the *null hypothesis*: there is no difference between the population means. In terms of SDT, the null hypothesis claims that, even though an observed difference of sample means occurs, the difference comes from undersampling, i.e., from the noise-alone distribution. The alternative hypothesis, H_a or H_1, is that there is a difference between the two population means. In terms of SDT, the alternative hypothesis states: the observed difference of sample means comes from the signal-and-noise distribution. In Fig. 3.4 we refer to this as "H_0 is False."

As in the yellow submarine example, a large t tells us that the discrimination between population means should be easy, while a small t-value indicates that discrimination should be hard and we may easily arrive at wrong conclusions. It is now straightforward to decide about the null hypothesis. We compute t and then apply a criterion. If the computed t-value is greater than the criterion, we take that as evidence that the estimated difference

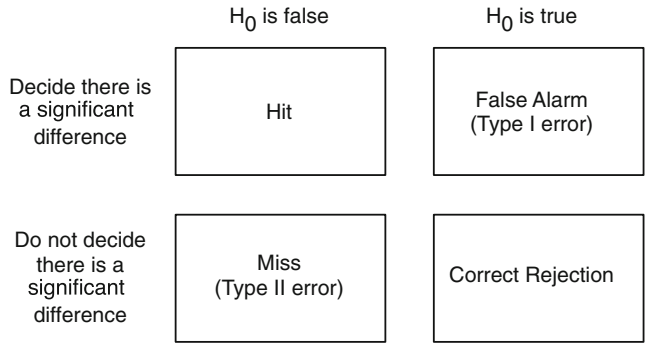


Fig. 3.4 Statistics is about making conclusions about a hypothesis. We have, similar to Chaps. 1 and 2, four outcomes. (1) The null hypothesis is false and we come to the conclusion that there is a difference in the means (Hit), (2) the null hypothesis is false and we have insufficient evidence against the null hypothesis (Miss, or Type II error). (3) The null hypothesis is true and we come to the conclusion that there is a difference in the means (False Alarm, Type I error). (4) The null hypothesis is true and we come to the conclusion we have insufficient evidence against the null hypothesis (Correct Rejection)

of means did not come from the noise-alone distribution: there is a difference between the two means. If the computed t-value is smaller than the criterion, then we do not have confidence that there is a difference between the two means. Maybe there is a difference, maybe not. We do not make any conclusions.

In practice, different fields use different criteria, which reflects their relative comfort levels with making Hits or False Alarms. For example, physics often follows the "5σ rule," which requires $t > 5$ to claim that an experiment has found sufficient evidence that there is a difference between mean values. Compared to other fields, this is a very high criterion; and it partly reflects the fact that physicists often have the ability (and resources) to greatly reduce σ and s by improving their measurement techniques. In particle physics, the Large Hadron Collider produces trillions of samples. Fields such as medicine, psychology, neuroscience, and biology, generally use a criterion that (to a first approximation) follows a "2σ rule." This less stringent criterion partly reflects the circumstances of scientific investigations in these fields. Some topics of interest are inherently noisy, and the population differences are small. Simultaneously, the per-unit cost for medical or biological samples is often much higher than for many situations in physics; and in some situations (e.g., investigations of people with rare diseases) a large sample size is simply impossible to acquire.

SDT also tells us that any chosen criterion trades off Hits and False Alarms, and the 5σ and 2σ rules are no exception. Everything else equal, the 5σ rule will have fewer Hits than the 2σ rule. Likewise, everything else equal, the 5σ rule will have fewer False Alarms than the 2σ rule.

Rather than setting a criterion in terms of standard deviation σ, in many fields (including medicine, psychology, neuroscience, and biology), scientists want to keep the Type I error smaller than a certain value, e.g., 0.05. It should be clear why one wants to limit this kind of error: it would cause people to believe there is an effect when there really is not. For example, one might conclude that a treatment helps patients with a disease, but the treatment is actually ineffective, and thus an alternative drug is not used. Such errors can lead to deaths. From a philosophical perspective, scientists are skeptical and their default position is that there is no difference: a treatment does not work, an intervention does not improve education, or men and women have similar attributes. Scientists will deviate from this default skepticism only if there is sufficient evidence that the default position is wrong.

Summary A Type I error occurs when there is no difference in the means, i.e. the null hypothesis is true, but we decide there is one. The Type I error is a False Alarm in terms of SDT. To decide about the null hypothesis, we compute t, which reflects the discriminability in terms of SDT, and then apply a criterion.

3.2.4 Type I Error: The p-Value is Related to a Criterion

Here, we show how the criterion determines the Type I error rate. Let us consider what the sampling distribution of the difference of sample means looks like when the Null hypothesis H_0 is true, i.e., $\mu_{North} - \mu_{South} = 0$. The distribution is centered on zero with a standard error that we estimated from our data. Suppose we set the criterion to $t = 2.0$, which is often called the critical value (cv) and written $t_{cv} = 2.0$. If our data produces a t-value larger than $t_{cv} = 2.0$, we will decide that there is a difference between the sample means-even though there is none, i.e, a Type I error. The probability of such a t-value is the area under the curve beyond $t_{cv} = 2.0$ (see Fig. 3.5a). This kind of test is called a "one-tailed t test". Calculating this area (assuming large sample sizes) gives 0.0228. Thus, if you use the criterion $t_{cv} = 2.0$, you will make a Type I error with a probability of only

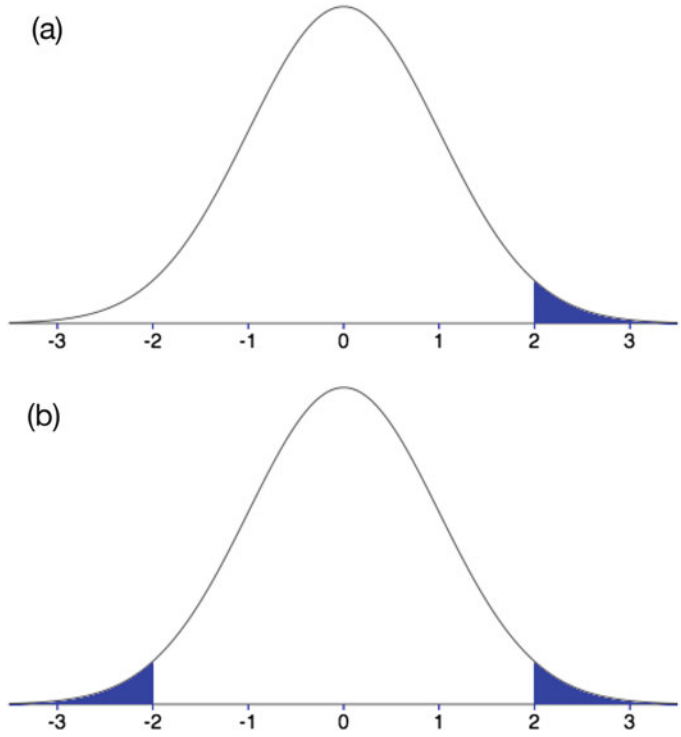


Fig. 3.5 The relation between criterion critical values and Type I error rates. The curve shows the noise alone distribution, i.e., when the Null hypothesis is true. The distribution is centered at zero and the variance is estimated from the data. (**a**) With a critical value of $t_{cv} = 2.0$, the Type I error rate is the area under the curve larger than t_{cv}. This test is called a one-sample t-test. (**b**) With a critical value of $t_{cv} = \pm 2.0$, the Type I error rate is the area under the curve for scores more extreme than ± 2.0

0.0228. If you followed the physics approach and used a 5σ rule, then $t_{cv} = 5$ and the Type I error rate is 0.000000287.

There is great flexibility in this approach. For example, you might suspect that Northern and Southern trees have different heights, but have no guess as to which would be larger. In this scenario, you might use a criterion $t_{cv} = \pm 2.0$, where a t-value more extreme (further from zero) than 2.0 would be taken as evidence that there is a difference in population means. With this approach, the Type I error rate would be 0.0456, which is twice as large as for the one-tailed test (see Fig. 3.5b). This test is called a "two-tailed t test".

In the above example, we set a criterion and computed the Type I error for this criterion. In statistics, usually, it is the other way around. We fix a Type I error rate and compute the corresponding criterion t-value. For example, if we accept a 5% Type I error rate, the corresponding criterion t-value is $t_{cv} = \pm 1.96$ for the two-tailed t test if n is large.[4] Rather than specify a certain t-value as the criterion, people compute the area under the curve beyond the t-value computed from the data. This area is called the p-value (see Fig. 3.6). Hence, we compute the t-value from the data and then the p-value. The p-value tells how likely it is that, if the Null hypothesis is true, we obtain our t-value or an even larger one. If the p-value is smaller than 0.05, we call the effect significant. Hence, to control the Type I error rate one simply requires that the computed p-value be less than the desired Type I error rate.

As mentioned, within medicine, psychology, neuroscience, and biology, a common desired rate is 0.05. For large sample sizes and for situations where one considers both positive and negative t-values (two-tailed t-test), a $p = 0.05$ corresponds to $t = \pm 1.96$. Thus, setting the Type I error rate to 0.05 corresponds to setting a criterion of $t_{cv} = \pm 1.96$. This relationship is why these fields follow an (approximate) 2σ rule. Whereas the t-value can be computed by hand, we need a statistics program to compute the p-value.

Summary If the t-value is larger than a certain value (which depends on the sample size n), we conclude that there is a significant effect.

3.2.5 Type II Error: Hits, Misses

In general, SDT tells us that for a given d', setting a criterion not only determines the Type I error rate, it also establishes the rate of Hits, Misses, and Correct Rejections. Indeed, it is easy to see that using a criterion that sets the Type I error rate to 0.05 also determines the Correct Rejection rate (the rate of concluding there is insufficient evidence for an effect when there really is no effect) to be $1.0 - 0.05 = 0.95$. As shown by the blue

[4] For small sample sizes, the t_{cv} criterion is larger because the sampling distributions are not quite Gaussian shaped. Statistical software that computes the p-value automatically adjusts for the deviation of sampling distributions from a Gaussian shape.

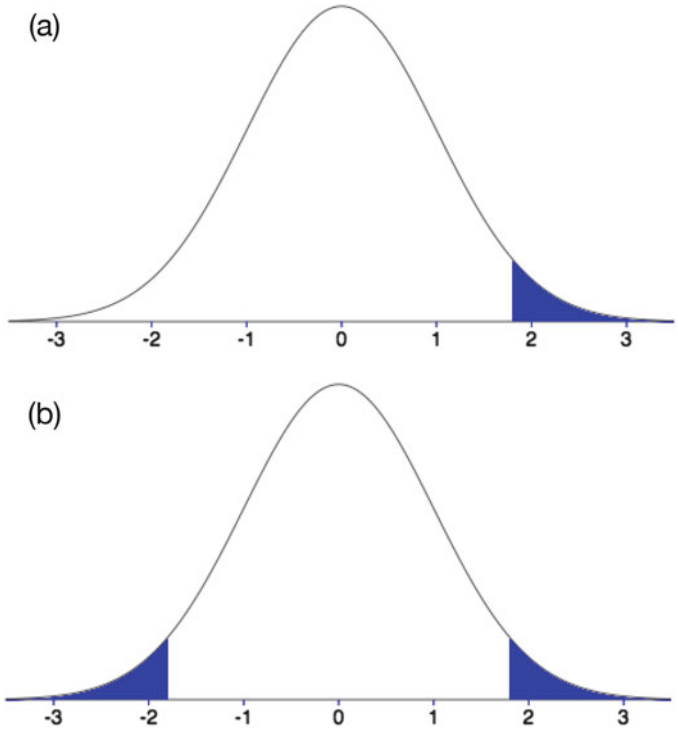


Fig. 3.6 The relation between t test statistics and p-values. (**a**) For a one-tailed test with $t = 1.8$, the p-value is the area under the curve larger than 1.8, $p = 0.0359$. (**b**) For a two-tailed test with $t = 1.8$, we compute the area more extreme than ± 1.8 in both tails. Here the p-value is 0.0718

area in Fig. 3.7, for a given criterion the area under this alternative sampling distribution to one side or the other corresponds to the probability of taking samples that produce a Hit (exceed the criterion and conclude evidence for a difference of population means) and Type II error (not satisfy the criterion and not conclude evidence for a difference of population means).

Hence, it seems that it would likewise be easy to compute the Type II error rate. However, this is not the case. When computing the Type I error, we know that the sampling distribution corresponding to the Null hypothesis is centered at one value, namely, 0. Hence, there is only one Null hypotheses. However, there are infinity many alternative hypotheses (see Implication 2e). But perhaps, we are only interested in substantial differences between the means of the North and South trees when the North trees are at least 1.2 m larger than the South trees. In this case, we know the minimal separation between the population distributions and can ask the question how large the sample size n must be to reach a significant result at least 80% of the time.

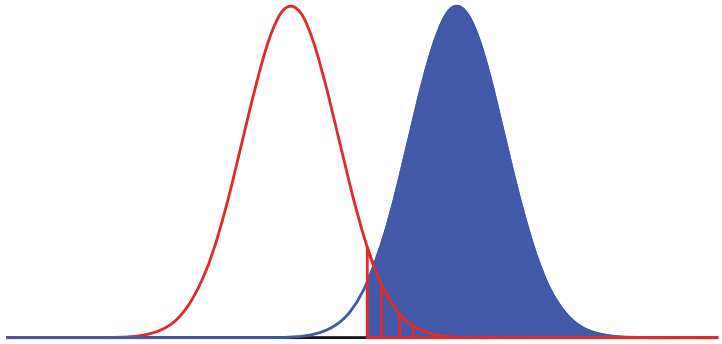


Fig. 3.7 Sampling distributions and the calculation of the Hit rate. The criterion corresponds to the lower end of the red hatched region and the lower end of the blue filled region. Any t-value that falls above this criterion will lead to a decision to reject the H_0 and conclude there is a difference in population means. The area in blue is the probability that the H_a sampling distribution will produce a t-value from this region. The Type I error rate is the red hatched area under the red curve

The Type II error plays an important role in terms of power. Power is the probability to obtain a significant result when indeed there is an effect, i.e., the Alternative hypothesis is true. Power is just another term for the Hit rate. The Hit rate is 1—Type II error rate. Power will be crucial in Part III and further explained in Chap. 7.

3.3 Summary

The above considerations are fundamental for understanding statistics. For this reason, we spell out the main steps once more and highlight the most important points. Even if you did not go through the above subsections, the main ideas should be understandable here.

We were interested in whether the *mean* height of oak trees of the North rim of the Alps is the same as for the South rim. The question is easy to answer. We just measure all trees, compute the two means, and know whether or not there is difference. If we miss a few trees, we obtain estimates, which are likely not too different from the true means. The fewer trees we measure, the larger is the *sampling error*, i.e., the more likely it is that our two sample means differ substantially from the two true means. As we have shown, this sampling error can be quantified by the standard error $s_{\overline{x}}$, which depends on the population standard deviation σ and the sample size n. If σ is small, we need only to sample a few trees to get a good estimate of the mean. For example, if $\sigma = 0$ we need only to sample one tree from each population because all trees have the same height in each population. If σ is large, we need to sample many trees to get a good estimate of the mean.

Let us now collect a sample of n trees from both the North and the South rim of the Alps with the samples being much smaller than the number of all trees in the two populations. We compute the mean heights for both samples. Because of undersampling,

almost surely there is some difference between the two sample means. However, we cannot know whether the observed difference indicates that there is a real difference in the means of the populations or whether the observed difference occurs because of undersampling while the population means are actually identical. If the true means are identical but we conclude, based on the sample mean difference, that the true means are different, we are making a Type I error (Fig. 3.4). Scientists generally want to avoid making a Type I error because their default position is that there is no effect until the data suggest otherwise. No decision making process can avoid sometimes making a Type I error, but we can control how often we make such an error. The important value is the t-value, which we can easily compute by hand:

$$t = \frac{(\overline{x}_{North} - \overline{x}_{South})}{s}\sqrt{\frac{n}{2}} = d\sqrt{\frac{n}{2}} \tag{3.12}$$

We compute the sample means $\overline{x}_{North}$ and $\overline{x}_{South}$ from the n trees x_i we collected, we estimate the standard deviation s of the trees (see grey box), and multiply by a function (the square root) of the sample size $n/2$. The right hand side shows that the t-value is nothing else as an estimate of the effect size d multiplied with a function of the sample size. The t-value tells us how easily we can discriminate whether or not a difference in the sample means comes from a real difference of the population means. The situation is exactly as in Chap. 2. The t-value is nothing else as a d' where, instead of dividing by the standard deviation, we divide by the standard error, which is a measure of the sampling error, taking both sources of noise, population variance and undersampling, into account. A large t-value means we can easily discriminate between means and a small t-value suggests the decision is hard. Note that a large t-value can occur because the estimated effect size d is large, n is large, or both are large.

Assume that there is no effect, i.e., the mean height of the North and South trees is identical ($\delta = 0$), then the p-value tells us how likely it is that a random sample would produce a t-value at least as big as the t-value we just computed. Thus, if we are happy with a 5% Type I error rate and the p-value is smaller than 0.05, we call our mean difference "significant".

The p-value is fully determined by the t-value and is computed by statistics programs. Most importantly, the t-value combines an estimate of the effect size d with the sample size ($\sqrt{\frac{n}{2}}$), which is why the t-value, and hence the p-value, confounds effect and sample size and, therefore, represents only partial information! This insight will be important to understand several implications, which we present after the following example.

3.4 An Example

Computing the p-value is simple, as the following short example will show. Understanding the implications of the t-test is more complicated.

Let us assume that we collected the heights of five trees from the North and five trees from the South. The data are presented in the first column of Fig. 3.8. The computations for the two-tailed t-test are also presented in the figure. For the given sample sizes and the computed t-value, our statistical software program tells us that the corresponding p-value is 0.045. Since this p-value is smaller than 0.05, we conclude that the data indicates a significant difference between the mean heights of Northern and Southern trees.[5]

Results from tests like this are often summarized in a table as presented in Table 3.1. The p-value is in the column marked "Sig. (2-tailed)." In the table, degrees of freedom (df) are mentioned. The degrees of freedom are important for the computation of the p-value because the shape of the sampling distribution is slightly different from a Gaussian for small sample sizes. In addition, one can compute the df from the sample size and vice-versa. In the case of the t-test, $df = n_1 + n_2 - 2 = 5 + 5 - 2 = 8$.

As mentioned, significance does not tell you too much about your results. It is important to look and report the effect size. Cohen proposed guidelines for effects sizes for a t-test, which are shown in Table 3.2.

Take Home Messages

- Since the p-value is determined by the t-value, it confounds effect size (d) and sample size (n). The original idea behind the t-test was to provide tools to understand to what extent a significant result is a matter of random sampling, given a certain effect size d. Nowadays, the p-value is often mistaken as a measure of effect size, which was never intended and is simply wrong!
- Partial information: proper conclusions can only be based on both the estimated population effect size, d, and the sample size, n. Hence, it is important to report both values, to take both values into account for conclusions, and to understand whether a significant result is driven by the estimated effect size d, the sample size, or both.

[5] Alternatively, one could identify a critical value criterion, $t_{cv} = \pm 2.306$ and note that t is farther from zero than this critical value.

North	$(x_{i,North} - \bar{x}_{North})^2$
20.80	(20.80 - 18.00)^2 = 7.840
17.81	0.036
17.92	0.006
18.30	0.090
15.17	8.009
$\bar{x}_{North} = \sum_{i=1}^{n} \frac{x_{i,North}}{n} = 18.00$	$s^2_{North} = \frac{\sum_{i=1}^{n} (x_{i,North} - \bar{x}_{North})^2}{n-1} = 3.995$
South	$(x_{i,South} - \bar{x}_{South})^2$
16.91	3.648
15.28	0.078
13.70	1.690
16.81	3.276
12.30	7.290
$\bar{x}_{South} = \sum_{i=1}^{n} \frac{x_{i,South}}{n} = 15.00$	$s^2_{South} = \frac{\sum_{i=1}^{n} (x_{i,South} - \bar{x}_{South})^2}{n-1} = 3.996$

$$df = (5-1) + (5-1) = 4+4 = 8$$

$$s_p^2 = \frac{s^2_{North}(n-1) + s^2_{South}(n-1)}{(n-1)+(n-1)} = \frac{3.995(4) + 3.996(4)}{4+4} = 3.996$$

$$s_{\bar{x}_{North}-\bar{x}_{South}} = \sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{n}\right)} = \sqrt{3.996\left(\frac{1}{5} + \frac{1}{5}\right)} = 1.264$$

$$t = \frac{\bar{x}_{North} - \bar{x}_{South}}{s_{\bar{x}_{North}-\bar{x}_{South}}} = \frac{18.00 - 15.00}{1.264} = 2.373$$

Fig. 3.8 The computations for the independent samples t-test start with computing the means for each data column ($\bar{x}_{North}$ and $\bar{x}_{South}$ shown at the bottoms of the first column). Using these, the variances can be computed (s^2_{North} and s^2_{South} at the bottoms of the second column). The degrees of freedom (df) for each column is computed by taking the number of scores in the column minus one. The pooled variance (s_p^2) is then computed by taking a weighted sum of the two variance terms (weighting by the degrees of freedom). We then substitute the pooled variance into the formula for the standard error $s_{\bar{x}_{North}-\bar{x}_{South}}$ and use this in the denominator of our t formula. The numerator is simply the difference between the means we calculated in our first step

Table 3.1 Output from a typical statistical software package

	t	df	Sig. (2-tailed)	Cohen' s d
Tree height	2.373	8	0.045	1.5 (large effect)

The columns labeled *t*, *df*, and *Sig. (two-tailed)* yield the t-value, its corresponding degrees of freedom, and $p-value$ respectively. The *df* value reported here is the sum of df_N and df_S (i.e., $4 + 4 = 8$)

Table 3.2 Effect size guidelines for d according to Cohen

	Small	Medium	Large
Effect Size	0.2	0.5	0.8

3.5 Implications, Comments and Paradoxes

For the following implications, Eq. 3.12 is crucial, because it tells us that the t-value and, thus the p-value, are determined by the estimated d and the sample size n.

Implications 1 Sample Size

Implication 1a According to Eq. 3.12, if the estimated $d \neq 0$, then there is always an n for which the t-test is significant. Hence, even very small effect sizes can produce a significant result when the sample size is sufficiently large. Hence, not only large effect sizes lead to significant results as one might expect, any non-zero effect size leads to significant results when n is large enough.[6]

Implication 1b If the estimated $d \neq 0$ (and $d < 4.31$), then there are sample sizes $n < m$, for which the t-test is not significant for n but is significant for m.[7] This pattern may seem paradoxical if you read it as: there is no effect for n but there is an effect for m. However, this is not the correct reading. We can only conclude that for m we have sufficient evidence for a significant result but insufficient evidence for n. From a null result (when we do not reject the null hypothesis) we cannot draw any conclusions (see Implication 3). We will see in Part III that this seeming paradox points to a core problem of hypothesis testing.

Implication 1c. Provocative Question Isn't there always a difference between two conditions, even if it is just very tiny? It seems that, except for a few cases, the difference between population means $\mu_1 - \mu_2$ is never really zero. How likely is it that the North and South tree means are both exactly 20.2567891119 m? Hence, we can always find a sample size n such that the experiment is significant. Why then do experiments at all?

Implications 2 Effect Size

Implication 2a As mentioned, the p-value is not a measure of the population effect size δ and, for each $d \neq 0$, there is a n for which there is a significant outcome. Thus, small effects can be significant. According to a study, consuming fish oil daily may significantly prolong your life. However, it may prolong your life by only 2 min. Do you bother?

Implication 2b By itself, the p-value does not tell us about the effect size. For example, when the sample size increases (everything else equal), the p-value decreases because the variance of the sampling distribution becomes smaller (see Fig. 3.3). Thus, if the effect size d is exactly the same, the p-value changes with sample size.

[6] We can describe the situation also as following. If there is a real effect $d \neq 0$, e.g., between the tree means, we can find a sample size n, for which we obtain a significant result in almost all cases (or with a certain high probability).

[7] If $d > 4.31$ you do not need to compute statistics because the difference is so large. In this case, even $n = 2$ leads to a significant result.

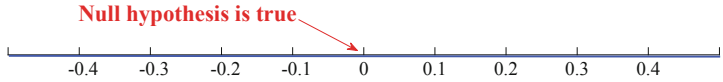


Fig. 3.9 In terms of effect size, the null hypothesis is represented by exactly one point, the null point. All other, infinitely many, points belong to the H_1 hypothesis

Implication 2c The p-values of two experiments A and B may be exactly the same. However, one cannot draw conclusions from this fact. For example, it may be that experiment A has a large effect size d and a small sample size and vice versa for experiment B. Hence, one can never compare the p-values of two experiments when the sample size is not the same. Likewise, a lower p-value in experiment A than B does not imply a larger effect size. The sample size might just be higher.

Implication 2d A study with a small sample size leading to a small p-value indicates a higher estimated effect size than a study with a larger sample size and the same p-value.

Implication 2e Just to reiterate and to visualize the basic situation. The worst case scenario is when the null hypothesis is true, i.e., there is no difference between means, and we conclude there is one: making a Type I error. In this case $\mu_1 - \mu_2 = 0$. If the null hypothesis is not true, $\mu_1 - \mu_2$ is not 0 and can be a value from $-\infty$ to $+\infty$ in principle. All of these values are part of the alternative hypothesis that there is a difference between the North and South trees. Hence, when we are worrying about the Type error I and the null hypothesis, we are worrying about only one single point embedded in infinitely many other points (see Fig. 3.9).

Implications 3 Null results

Implication 3a Absence of proof is not proof of absence: one can never conclude that there is *no* effect in an experiment ($d = 0$) when there was no significant result. A non-significant p-value indicates either that there is no difference or a real difference that is too small to reach significance for the given sample size n.

Implication 3b A difference of significance is not the same as a significant difference. Consider a study measuring the effect of a cream containing Aloe Vera on skin eczema. Patients with eczema are randomly assigned to two groups: one receiving the cream with Aloe Vera and one receiving a placebo cream. After 4 weeks, the size of the eczema is measured again. There was a significant reduction in the Aloe Vera group but not in the placebo group (Fig. 3.10). It might be tempting to conclude that the study demonstrates that Aloe Vera cures eczema. However, there is a reduction in the placebo group too—just smaller (which may be due to self-healing). In fact, when we compute the difference in eczema reduction for each participant in both groups and compute a two-tailed t-test between the two groups, the difference is not significant.

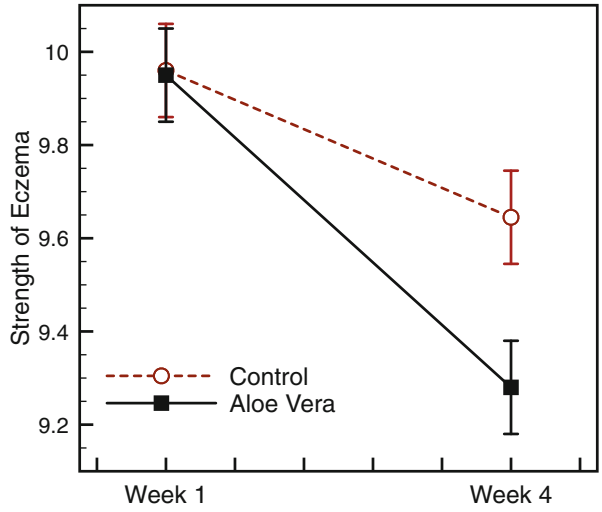


Fig. 3.10 An Aloe Vera cream was given to an experimental group and a placebo to a control group to study whether Aloe Vera can reduce the effects of eczema. The strength of eczema was tested at the beginning of the experiment and after 4 weeks. The plot shows the mean value of eczema strength for each group at each measurement. The error bars indicate the standard error for each mean value. Eczema strength significantly reduced in the experimental group but not in the control group. Can we conclude that Aloe Vera reduces eczema? We cannot because eczema strength reduced also in the control group potentially because of self-healing. Actually, there was no significant effect when the improvements in both groups were compared with a two-tailed t-test. A difference of significance is not the same as a significant difference. The graph shows the mean values and corresponding standard errors (see grey box "Some definitions")

One might argue that with a larger sample size, the difference between the two groups might become significant. This may indeed be true. However, also the effect in the placebo group may become significant. What should we conclude? Contrary to intuition, this situation does not constitute a problem because we can ask whether or not there is a stronger effect in the Aloe Vera than the placebo condition (thus, discounting for the self-healing).

Importantly, the example shows that it often makes very little sense to compare statements such as "there was an effect in condition A but not in condition B". Such conclusions are ubiquitous in science and should be treated with great care (see also Chap. 7). The classic example is as in the Aloe Vera case to compare an intervention condition, where a significant results is aimed to occur, with a control condition, where a Null result is aimed for.

Implications 4 Truth, Noise and Variability

Implication 4a Why do we compute statistics? Often it is implicitly assumed that statistics "clears off" the noise that is inevitable in complex systems. In the submarine example

measurements are corrupted by changes in the water, such as fish, algae, or the device itself, which randomly fluctuate. This kind of noise is called measurement noise. All these noise sources compromise the true signal from the rock and the true signal when there is no rock. The situation can be described with the following equation:

$$x_j = \mu + \epsilon_j$$

where x_j is the outcome in the jth measurement, μ is the true signal, and ϵ_j is the noise, which changes from trial to trial. Usually it is assumed that ϵ_j is Gaussian distributed with a mean of zero. Thus, basing decisions on one trial is not a good idea. As mentioned, averaging across many measurements can clear off the noise. That's why it is better to compare mean values than single measurements. As we have seen, the larger the n, the better the measurement of the mean.

This kind of model is appropriate in many fields, such as physics. However, in biology, medicine and many other fields, the situation is often quite different. For example, we might determine the strength of headaches before and after a pain killer was consumed. We find that the drug decreases headache strength *on average*. However, as with most drugs, there may be people who do not benefit from the drug at all. In addition, some people receive more benefit than others, i.e., for some people headaches may almost disappear while for others there is only a small (or even opposite) effect. These person-specific effects might hold for all days and situations.

Person-specific effects can be described with the following equation:

$$x_{ij} = \mu + v_i + \epsilon_{ij}$$

where x_{ij} is one measurement, e.g., person i taking the pain killer at day j. μ is the mean value of the entire population, e.g., to what extent the drug decreases headaches on average. v_i is the sensitivity of the person i for the pain killer. As mentioned, some people always benefit strongly from the drug, while for others there is no effect, and for even others headaches always increase when taking the pain killer. Hence, v_i determines how much one person differs from other persons—and the mean μ. ϵ_{ij} is measurement noise and reflects, for example, to what extent the pain killer leads to different effects from day to day in the very same person. In some way, ϵ_{ij} indicates unsystematic variability, whereas v_i captures systematic variability. Just as another example. A person may have a higher blood pressure than someone else, and this difference is reflected by the inter-participant variability v_i. At the same time, blood pressure varies greatly from minute to minute in the very same person, and this difference is reflected by ϵ_{ij}, the intra-participant variability.

In many experiments, one cannot easily disentangle v_i and ϵ_{ij}. Both terms contribute to the estimated standard deviation of the population distribution, s. From a mathematical point it does not matter whether there is strong inter-participant variability or strong measurement noise. However, for interpreting the statistical analysis, the distinction is crucial. Assume there is a strong beneficial effect of the pain killer for half of the

population whereas there is a smaller detrimental effect for the other half of the population. On average, the drug has a positive effect and this effect may turn out to be significant. Importantly, whereas the drug is beneficial *on average*, this is not true individually. For half of the population, the effect is detrimental and it is not a good idea to use the pain killer. Hence, when v_i is not zero, significant results do not allow conclusions on the individual level. A study may show that carrots are good for eye sight on average. Whether this is true for you is unclear. Carrots may actually deteriorate your vision, even though they help the vision of other people. These considerations do not imply that such studies are wrong, they just show the limitations of studies where $v_i \neq 0$ for some i. For an international comparison of blood pressure values, average values are good. However, it is usually not a good idea to compare yourself to such a large group, whatever is being measured. Such a sample is not only heterogeneous across the regions but also contains people of different ages. A body mass index of 27 may be an issue for children below 5 years but not necessarily for people older than 70 years. Hence, it depends very much on the research question to what extent a mean comparison makes sense. It is a matter of interpreting statistics, not of computing statistics.

Implication 4b The above considerations have philosophical implications. Usually, we assume that something is either the case or it is not the case. Either gravity acts on all matter in the entire universe, or it does not. Either oxygen is necessary for humans, or it is not. All of these facts hold true for each individual, i.e., for each element in the universe, for all humans, etc. If a fact has been proven by methods involving statistics, this conclusion is not necessarily justified when v_i is different from 0 because the results hold true only on average, not necessarily for all individuals.

Implication 4c The variability vs. noise problem becomes even more serious when the study contains a non-homogeneous sample differing systematically in a feature that is not explicitly considered. For example, based on how often they go to the doctor, it seems that shorter students are ill more often than taller students. However, this fact has nothing to do with body size. It is simply the case that female students are shorter than male students on average *and* see the gynecologist much more often than male students see the urologist. However, females see the gynecologist mainly for preventive medical checkups and are by no means more often ill than male students. Since students generally see doctors very infrequently, visits to the gynecologist weigh strongly in the statistics. It is obvious how mis-interpretations can occur even in such simple examples. In more complex situations such mis-interpretations are less easy to spot. By the way, one should question whether it is good idea to make conclusions about illness frequency based on doctor visits.

Implication 4d One can also consider the variability vs. noise problem the other way around. When you are planning an experiment, you need to specify whom to include. To be representative, it is good to sample from the entire population, e.g., from all people in a country or even world wide. However, with this procedure, you may include

a heterogeneous population, which makes conclusions difficult. Should you include astronauts or coma patients? What about ill people? The large portion of people with too high blood pressure? The more subpopulations you exclude, the less representative is your sample. Eventually, your sample may include only you.

Implication 4e As a last point. Effects often depend on dosage, i.e., different people may respond differently to different dosages. A pain killer may have beneficial effects for some people in a low dosage but be detrimental for a higher dosage. Hence, there is not only systematic inter-person variability but also systematic intra-person variability in addition to the unsystematic noise ϵ_{ij}. In many experiments, there are many sources involved, i.e., effects depend on dosage, inter-individual differences, and noise—limiting conclusions strongly. As we will see, dosage dependent effects are best described by correlations (Chap. 8) rather than by t-tests.

Implications 5a The Statistics Paradox and the Dangers of Cohort Studies

For large effect sizes, as they occur for example in physics, we often do not need to compute statistics. Likewise, the hypothesis that elephants are on average larger than ants does not need statistics because any living elephant is larger than any ant, δ is extremely large. The original idea of statistics was to determine whether a "bit noisy effect" really exists and to determine the sample sizes n needed to show that indeed the effect is real. We may say that statistics was developed for medium effect sizes and medium sample sizes. In the past it was usually impossible to obtain significant results with small effect sizes because data were scarce and handling large sample sizes was cumbersome. Hence, n was usually small and only experiments with large effect sizes produced significant results. This has changed largely because data collection has become cheap, and it is possible to combine and handle millions of samples as, for example, in genetics. For this reason, nowadays statistics is widely used not only for medium effects but also for very small effect sizes. However, this development is not free of danger. First of all, large sample sizes should not be confused with large effect sizes (Implication 2a). Second, conclusions are often very difficult to draw, particularly, in so called cohort studies. In cohort studies, for example, patients are compared with controls, or vegetarians are compared with meat eaters. The two groups are defined by a given label.

Here is an example. Starting in 1948, blood pressure was measured for 5209 participants in the small city of Framingham in Massachusetts. Participant age is plotted in Fig. 3.11 on the x-axis and the systolic blood pressure on the y-axis. Data were split depending on the level of education. First, there is a clear effect of age. Second, more education is associated with lower blood pressure. Using statistical techniques described in Chap. 6, the effect of education turned out to be significant. Does this mean that prolonged education *causes* lower blood pressure? Likely not. Maybe people with more education smoke less. Maybe, maybe not. They may smoke fewer cigarettes per day (dosage dependent). Maybe, maybe not. They may have started smoking later or quit

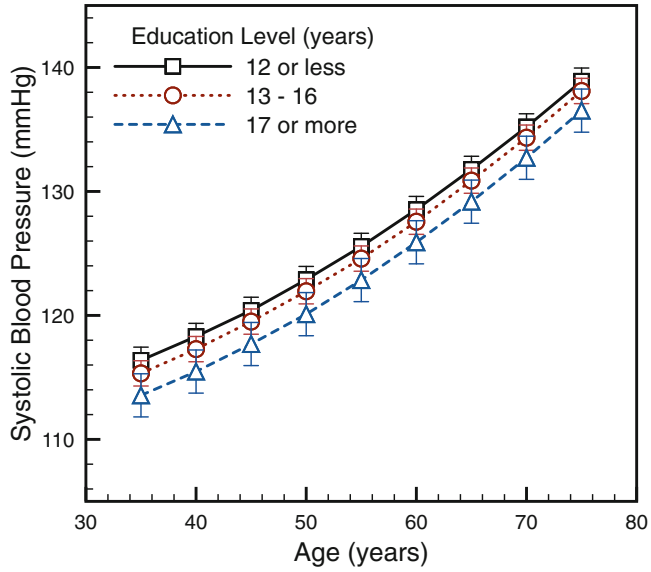


Fig. 3.11 Mean systolic blood pressure for three populations of people from the city of Framingham in Massachusetts, USA, as a function of age. The error bars indicate the standard error for each mean. The three populations differ in the number of years of education. Blood pressure increases with age. In addition, blood pressure is lowest for the group with the most years of education and is highest for the group with the fewest years of education. What can we conclude? As we argue below: not too much. Data are re-plotted from Loucks et al. [1]

earlier. Nutrition may play a role. Sports. The work environment. Genetics. Maybe there is a small subgroup, who works under very unhealthy conditions, which alone causes the higher blood pressure. There are so many potential factors, many of which are unknown today and will be discovered in the future: tangerines in your diet may lower blood pressure. Moreover, combinations of factors may play a role. Perhaps nutrition only plays a role when people do no sports.

The difference in blood pressure between the different education groups is only about 2 mm Hg. To put this effect size into perspective, measure your blood pressure and repeat 5 min later. You will see that 2 mm Hg is very small compared to your intra-person variance (ϵ_{ij}) and very low compared to the large range of inter-person variability (v_i). In addition, blood pressure strongly changes during activity. Maybe there is only a difference when the blood pressure is measured during rest. Maybe, maybe not. The main problem with these, so called, cohort studies is that there are too many factors that are causally relevant, but cannot be controlled for. To control for all these effects and the combinations, sample sizes may need to be larger than the number of people on the planet. In addition, is it really worth investigating 2 mm Hg? If you want to lower your blood pressure, a little bit of sport might do a good job and is much cheaper than paying thousands of dollars for education.

Implications 5b. Small Effects Sizes As shown, studies with small effect sizes require extra care. However, small effect size are not always problematic. First, it is good to reduce the side effects of a drug that is consumed by millions of people, even if it is only by 1%. Second, many important discoveries started off with small effects; but subsequent investigations refined the methods and produced bigger effects.

Implications 5c. Conclusions Importantly, both small and large sample sizes can be problematic. It is well known that *small* sample sizes are a problem because of undersampling. It is less well understood that *large* sample sizes may be as problematic when effect sizes are small because even tiny differences may become significant. In particular, cohort studies with small effects sizes and large sample sizes are often useless because small correlations between the investigated factor and unrelated factors can create significant results. For this reason, it is important to look at both the effect size and the sample size. Whereas the sample size n is usually mentioned, this is not always true for effect sizes. For the t-test, the effect size is often expressed as Cohen's d (see also Chap. 4). In the following chapters, we will introduce effect sizes for other tests.

How to Read Statistics? For different samples, the estimate of the effect d' may vary strongly. The larger the sample size n the less variance is there and the better is the estimate. Hence, first, look whether n is sufficiently large. If so, decide whether the effect size is appropriate for your research question. Tiny effect sizes are only in some cases important and may come from confounding, unidentifiable factors. In Part III, we will see that the combination of sample size and effect size can give interesting insights into the "believability" of a study. For example, we will ask how likely it is that four experiments, each with a small sample and effect size, all lead to significant results with p-values just below 0.05.

Take Home Messages

1. Even small effect sizes lead to significant results when the sample size is sufficiently large.
2. Do not compare the p-value of two experiments if n is not identical: a smaller p does not imply more significance.
3. Statistical significance is not practical significance.
4. Absence of proof is not proof of absence: avoid conclusions from a Null result.
5. Do not pit a significant experiment against a non-significant control experiment.
6. Cohort studies with small effects are usually useless.
7. A statement like "X is true" can only be true for sure if inter-subject variability is zero.

Reference

1. Loucks EB, Abrahamowicz M, Xiao Y, Lynch JW. Associations of education with 30 year life course blood pressure trajectories: Framingham Offspring Study. BMC Public Health. 2011;28(11):139. https://doi.org/10.1186/1471-2458-11-139.

Variations on the *t*-Test

4

Contents

What You Will Learn in This Chapter

In Chap. 3, we introduced the basic concept of statistics within the framework of SDT. Here, we present a classic introduction to hypothesis testing and present variations of the *t*-test.

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_4

4.1 A Bit of Terminology

Type of Experiment

- Experimental study: samples are randomly assigned to two groups. For example, patients are *randomly* assigned to an experimental group that takes a potentially potent drug and a control group that receives a placebo.
- Cohort study: the groups are defined by predefined labels, such as patients vs. controls, vegetarians vs. meat eaters, astronauts vs. earthlings. Cohort studies are common and useful, however, they also face severe problems as seen in Chap. 3, Implication 5a.

Type of Variables and Metrics In plots, usually, the x-axis represents the independent variable, the y-axis the dependent variable. For both variable types there are four main types of measurement scales.

- Nominal: there is no order. For example, blood pressure is determined for people from four different countries. On the x-axis, you can plot *any* order of the countries. As another example for a nominal scale: therapy A vs. B.
- Ordinal: just ranks. For example, a general has a higher rank than a lieutenant but the rank of the general is not, say, two times higher than the rank of the lieutenant. On the x-axis, you can plot the ranks in an *ascending* order. The distance between points on the x-axis has no meaning.
- Interval: values can be added or subtracted but not meaningfully divided or multiplied. A 30 °C day is 15 °C hotter than a 15 °C day but it is not twice as hot because 0 °C does not mean the absence of heat. For this reason, physicists use the Kelvin temperature scale, which sets 0 K as absolute zero.
- Ratio: values can be added, subtracted, multiplied and divided, in particular, ratios make sense. The classic example is a measurement of weight (e.g., kilograms). The value zero defines the origin of the scale and refers to "none" of whatever the variable describes, be it length, weight or whatever.

Type of Tests

- Parametric test: a test where a model of the data distribution is presupposed. For example in Chap. 3, we assumed that the population tree heights are Gaussian distributed. Parametric distributions can typically be described by a small number of parameters (e.g., the mean and the standard deviation for Gaussian distributions).
- A non-parametric test: a test that does not assume any specific distribution. Some non-parametric equivalents of the t-test are discussed below.

4.2 The Standard Approach: Null Hypothesis Testing

In Chap. 3, we explained statistics within the framework of SDT. Here, we describe the classic null hypothesis testing approach with the example of the two-sample t-test.

The steps for making a statistical decision for a two-sample t-test are:

1. State your alternative hypothesis, called H_1, such as Therapy A is better (or different) than Therapy B.
2. Assume the null Hypothesis H_0 is true: there is no difference between Therapy A and B.
3. From your data, compute the standard error:

$$s_{\overline{x}_A - \overline{x}_B} = s\sqrt{2/n}$$

4. Compute the test statistic as in Chap. 3:

$$t = \frac{\overline{x}_A - \overline{x}_B}{s_{\overline{x}_A - \overline{x}_B}}$$

   and the corresponding p-value.
5. Make your decision. If $p \leq 0.05$, reject the H_0 hypothesis and accept H_1: call the effect significant. If $p > 0.05$, you cannot make a statement; in particular do not conclude that H_0 is true.

The above approach has the nice characteristic of setting a limit on the probability of making a Type I error (false positive). Suppose that the null hypothesis is actually true; meaning that we actually draw our samples from the noise-alone distribution. If we now draw many samples from the noise-alone distribution, we will find, on average, that p is less than 0.05 only 5% of the time. We could be more stringent and require $p < 0.01$, in which case p is less than 0.01 only 1% of the time. Of course, there is always a trade-off; the more stringent we are, the more Type II errors (misses) we make when the samples are actually from the alternative hypothesis distribution.

4.3 Other t-Tests

4.3.1 One-Sample t-Test

Sometimes, one wants to compare a single mean with a fixed value. This test is called a one-sample t-test. For example, sometimes a researcher wants to show that a therapy increases IQ, which is 100 on average. We assume that without the therapy the distribution has a mean score of $\mu_0 = 100$. Hence, if the null hypothesis is true and the therapy has no effect, we get the standardized distribution of the IQ in the population. The sampling

distribution of the mean has a standard deviation of:

$$s_{\overline{x}} = \frac{s}{\sqrt{n}} \tag{4.1}$$

which is the standard error of the mean. We compute a t-value as:

$$t = \frac{\bar{x} - \mu_0}{s_{\overline{x}}} \tag{4.2}$$

and the degrees of freedom is:

$$df = n - 1 \tag{4.3}$$

With this information we can compute a p-value and make decisions just as for the two-sample t-test.

4.3.2 Dependent Samples t-Test

Often there are two samples of data, but they are related in some specified way. For example, a researcher wants to test whether a therapy increases the level of red blood cells by comparing cell concentration before and after the therapy in the *same* participants. As another example, one might measure preferences for action movies among couples in a relationship. The key characteristic is that every score measured in one sample can be uniquely tied to a score in the other sample. Hence, we can create a *difference score* for each pair. Thus, the two measures of the red blood cell concentration before (x) and after (y) therapy for a given patient would produce a single difference score for that patient:

$$d = y - x \tag{4.4}$$

We now have a single sample of a set of difference scores, and we can run a one-sample t-test on those difference scores just as above. The standard error of the difference scores is:

$$s_{\overline{d}} = \frac{s_d}{\sqrt{n}} \tag{4.5}$$

where s_d is the standard deviation of the sample difference scores d (pay attention to the context of the discussion so as to not confuse this variable with Cohen's d). Like before, we can compare the sample mean to a hypothesized difference of population means (being sure to subtract the population means in the same way that was done for individual scores):

$$t = \frac{\overline{d} - (\mu_y - \mu_x)}{s_{\overline{d}}} \tag{4.6}$$

We compute the p-value in the same way as before by using:

$$df = n - 1 \tag{4.7}$$

This test is also called a repeated measures t-test, a paired t-test, or a within-subjects t-test.

4.3.3 One-Tailed and Two-Tailed Tests

In the above examples, we implicitly considered whether the means of two groups are different from each other ($\mu_A \neq \mu_B$), i.e., we did not specify whether Therapy A is better than Therapy B. These t-tests are called *two-tailed* t-tests, because a large t-value could happen in either tail of the null sampling distribution. One could also pose that if there is a difference then Therapy A is better than Therapy B ($\mu_A > \mu_B$). Alternatively, one could pose that if there is a difference then Therapy B is better than Therapy A ($\mu_B > \mu_A$). In these cases, the t-value can only be in one tail (see Chap. 3, Fig. 3.5). Thus, for a one-tailed test you only have one criterion to satisfy and hence a smaller criterion is required to maintain the desired false positive rate of 0.05. For this reason, the one-tailed test has a higher power than a corresponding two-tailed test.

However, the use of the one-tailed t-test is controversial. In our tree example, the North trees could be either larger or smaller than the South trees. Thus, a two-tailed t-test is more appropriate unless one has a good argument that the North trees are larger than the South trees. One cannot use a one-tailed test when the two-tailed test has led to a non-significant result (see Sect. 11.3.5)! One also cannot decide to use a one-tailed test just because you observe that the data produces one mean larger than the other. Decisions about whether to use a one-tailed or a two-tailed test must be made based on theoretical justification; it cannot be determined by the data or the outcome.

4.4 Assumptions and Violations of the t-Test

As traditionally taught, the main point of the t-test is to control the Type I error rate (false alarm rate). Deviations from the below assumptions almost always alter the corresponding Type I error rate; sometimes in a big way and sometimes in a small way.

4.4.1 The Data Need to be Independent and Identically Distributed

Sampled data need to be Independent and Identically distributed (IID). This is a requirement for many statistical tests. For example, you want to test whether a pain killer reduces not only headaches but also body temperature. You can collect a sample of participants and measure their body temperature before and after the intake of the drug and compute a

paired t-test. It is important that a person only participates one time. If you would like to make a hypothesis about the general population, you cannot do the experiment 10 times on 10 days only on yourself because this data is not independent. Maybe you are the only person on the planet for whom the painkiller works.

Here is another example. You are measuring visual acuity at eight locations in the visual field for three patients. Thus, you have 24 data points but they are not independent, so you cannot subject the 24 data points to a t-test. You could average the eight data points for each patient and compute a t-test. Hence, your sample size is only 3, not 24.

Data need to be identically distributed, i.e., they need to come from the same population probability distribution. For example, the height of different types of plants cannot be mixed in a sample. Even if both distributions are Gaussians, the variances may be very different for oaks and edelweiss. If you, for example, measure the heights in a sample of plants collected at both the North and South rims, there might be large differences just because you included more oaks and fewer edelweiss in the North than South sample.

4.4.2 Population Distributions are Gaussian Distributed

The t-test requires that the population distributions are Gaussians[1] or sample sizes are large (often a value of $n = 30$ suffices). However, the t-test is rather robust with respect to populations that are not too different from a Gaussian shape. By robust, we mean that the Type I error rate is close to what is intended (e.g., 5% when using $p < 0.05$ to reject H_0). As long as the distribution is unimodal,[2] even a high amount of skew has only a little effect on the Type I error rate of the t-test (a skewed distribution is not symmetric and has a longer tail on one side of the distribution than the other).

4.4.3 Ratio Scale Dependent Variable

Since the t-test compares means, it requires the dependent variable to be on a ratio scale of measurement. Computing a mean does not make sense for nominal data. Computing variance (or standard deviation) does not make sense for nominal or ordinal data. Since the t-test uses both the sample mean and the sample standard deviation, neither nominal nor ordinal data should be analyzed with a t-test.

There are different opinions about whether a t-test should be used for data on an interval scale. Technically, the properties of the t-test require ratio scale data, but in many cases the t-test behaves rather reasonably for interval data.

[1] Whether the Gaussian assumption is met can be tested by the Kolomogorov-Smirnov test.

[2] A unimodal distribution has only one peak. For example, the Gaussian has only one peak. Bimodal distributions have two peaks.

Table 4.1 Type I error rates for 10,000 simulated t-tests with different population standard deviations and sample sizes

	$n_1 = n_2 = 5$		$n_1 = 5,\ n_2 = 25$	
	$\sigma_2 = 1$	$\sigma_2 = 5$	$\sigma_2 = 1$	$\sigma_2 = 5$
$\sigma_1 = 1$	0.050	0.074	0.052	0.000
$\sigma_1 = 5$	0.073	0.051	0.383	0.047

4.4.4 Equal Population Variances

The standard two-sample t-test assumes that each population has the same variance. Unequal standard deviations, especially combined with unequal sample sizes, can dramatically affect the Type I error rate. Table 4.1 reports the Type I error rate for 10,000 simulated t-tests where the null hypothesis was actually true. For each simulated test, a computer program generated "data" from population distributions and ran the t-test on that generated data. Across different simulations, the *population* standard deviations were either equal (e.g., $\sigma_1 = \sigma_2 = 1$) or unequal (e.g., $\sigma_1 = 5$, $\sigma_2 = 1$) and the sample sizes were either equal (e.g., $n_1 = n_2 = 5$) or unequal (e.g., $n_1 = 5$, $n_2 = 25$).

Table 4.1 demonstrates that if the sample sizes are equal then a difference in the population standard deviations somewhat increases the Type I error rate. Around 7% of the samples rejected the null hypothesis. However, if the samples sizes are unequal and the variances are different, then the Type I error rate is much smaller or larger. When the small sample is paired with the small population standard deviation, then the Type I error rate is much smaller than the intended criterion, 0.05. In this particular set of simulations, not a single t-test rejected the null hypothesis. On the other hand, if the small sample is paired with the large population standard deviation then the Type I error is nearly 40%, which is nearly eight times larger than the intended 5% criterion! The problem is that the default t-test pools the standard deviation from each sample to produce a single estimate of the population standard deviation. If the small sample is paired with the small population standard deviation, then the pooled estimate is too large and the test is unlikely to reject the null. If the small sample is paired with the large population standard deviation, then the pooled estimate is too small and the test is too likely to reject the null.

These problems can be addressed by using a variation of the t-test called the Welch test. However, there is a cost; if the population standard deviations are actually equal, then the Welch test has smaller power than the standard t-test (it is less likely to reject the null hypothesis when there really is a difference).

4.4.5 Fixed Sample Size

Before the experiment, one needs to fix the sample sizes for both groups. One cannot change the sample sizes during the ongoing experiment. This requirement is more difficult

Table 4.2 Parametric tests and their corresponding non-parametric tests

Parametric	Non-parametric
One sample t-test	Sign test
Two-sample t-test	Wilcoxon rank sum test
Repeated measures t-test	Man-Whitney U-test

to satisfy than you might suppose. We discuss a type of violation and its impact in Sect. 10.4.[3]

4.5 The Non-parametric Approach

If your data are not Gaussian distributed you might consider using a non-parametric test. For each t-test described above, there is a non-parametric test, as summarized in Table 4.2.

Non-parametric tests have less power because they cannot exploit a model, i.e., non-parametric tests usually need larger sample sizes for significant results.

The calculations for non-parametric tests look rather different than the calculations of a t-test, however, the non-parametric tests follow the same basic principles of SDT.

Take Home Messages

1. For a t-test, make sure your data are iid distributed and the y-axis is a ratio-scale.
2. Data should be Gaussian distributed or n should be large.

4.6 The Essentials of Statistical Tests

Let us return to the t-test. We had a research question about *mean* differences of trees. Then, we assumed a statistical model, namely, that trees are Gaussian distributed. From the model we derived the equation for the t-value, which is called a *test statistic* and allowed us to compute the p-value and thereby control the Type I error rate. This principle can be applied to many statistical questions. For example, we may ask questions about whether the *variances* of two population distributions differ, the shapes of the population distributions differ (χ^2 test), or the ratio of two means is different from 1 (z-test). More complex tests compute, for example, the means depending on other variables and even

[3] One can fix the sample size n and apply additional criteria such as: the total sample comprises 20 participants, however, if a participant has reduced visual acuity, as determined by an eye test before the experiment, this person can be excluded at this stage and can be replaced by another participant.

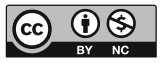
more complex tests assume much more complicated models, for example, hierarchical probability distributions.

The principle for all tests is always the same and all cases can be understood in the framework of SDT following the rationale that explains the t-test. The only difference is that a different statistical model than the t-distribution is used. How exactly the test statistics are computed for the various tests is of less importance for understanding statistics because these calculations are done by the computer. For all parametric tests, the p-value confounds effect and sample size.

4.7 What Comes Next?

It is always a good idea to keep our experimental design as simple as possible so you can apply a t-test or a corresponding non-parametric test. However, maximum simplicity is not always possible. For example, we may want to study more than two tree populations, and then a t-test cannot be applied. More variables, e.g., more tree populations, come with a multiple testing problem, which we describe in the next part of this book. The multiple testing problem can be addressed with either statistical methods or clever experimental designs (Chap. 7). We will portray the most common methods because they include an approach that is not evident in the t-test. Although there exist other tests, we do not explain them because this book is about the essentials of statistics and not a compendium.

In Part I of this book, we have laid out many fundamental terms of statistics, as they are needed for statistics users. Readers who are not interested in the specific tests of Part II can proceed directly to Part III, where these key terms are used to explain why we currently have a science and statistics crisis.



Part II

The Multiple Testing Problem

The Multiple Testing Problem

5

Contents

What You Will Learn in This Chapter

In Part I, we focused on the most basic statistical comparison, namely, the comparison of two means. We described the t-test, which has high power and, hence is a good test and should be used whenever possible. However, sometimes more than two means need to be compared; e.g., if we want to compare the population of trees in three regions of the world. In this case, a multiple testing problem arises that increases the risk of making a Type I error.

In this chapter, we will introduce the multiple testing problem and present Bonferroni corrections as one (rather suboptimal) way to cope with it.

5.1 Independent Tests

To understand the multiple testing problem, consider the following situation. If we compute one t-test, we know that if the null hypothesis is actually true then there is a Type I error rate of $\alpha = 0.05$. Another way of saying this is that we do not produce a Type I error (False Alarm) in $1 - 0.05$ of the cases when the null is true. If we compute two *independent* t-tests, the chance of not making any False Alarm is $0.95^2 = 0.9$. For 12 comparisons: $0.95^{12} = 0.54$. So the risk of making at least one False Alarm with 12 comparisons is $1 - 0.54 = 0.46$. Hence, it becomes more and more likely to produce False

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_5

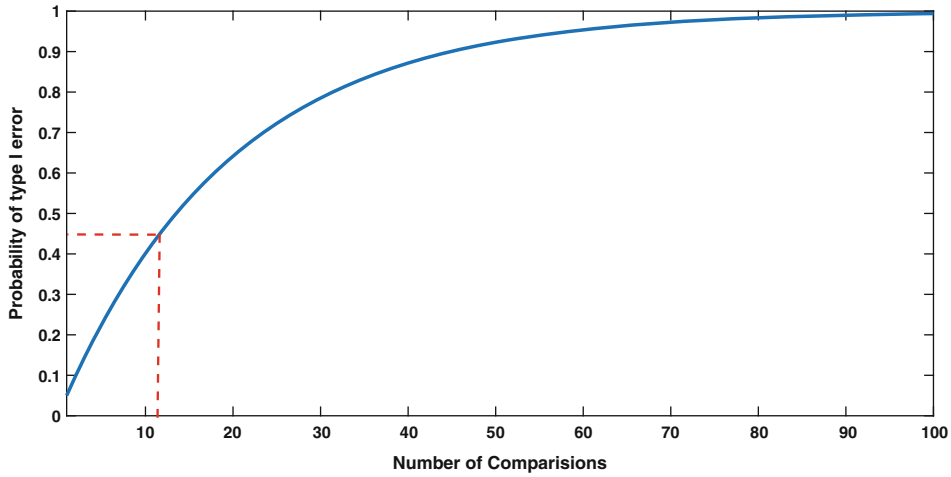


Fig. 5.1 The Type I error rate (False Alarm rate) strongly depends on the number of comparisons. For example with 12 comparisons (red dashed lines), the probability of making at least one False Alarm is 0.46, i.e., much increased compared to 0.05 with only 1 comparison

Alarms when the number of comparisons increases (Fig. 5.1). In general, the probability of making at least one Type I error for m independent tests is:

$$1 - (1 - \alpha)^m \tag{5.1}$$

or $1 - (1 - 0.05)^m$ for $\alpha = 0.05$.

Bonferroni Corrections One classic way to account for the increase in Type I error is to reduce the required significance level. If we want to set the Type I error rate for m independent tests to be equal to 0.05, we set Eq. 5.1 to be equal to 0.05 and solve for α:

$$\alpha = 1 - (0.95)^{\frac{1}{m}} \approx \frac{0.05}{m} \tag{5.2}$$

To have a False Alarm rate of 0.05 across all m tests, you need

$$p < \frac{0.05}{m} \tag{5.3}$$

Hence, with multiple tests we need a smaller p-value for any given test to reach significance. Signal Detection Theory tells us that a more conservative criterion always involves a trade off in Hits and False Alarms. Indeed, power (Hits) strongly decreases when using Bonferroni correction.

Statisticians are not in general agreement about whether, or when, Bonferroni (or other similar) corrections are appropriate. Obviously, you should not treat m as the total number of hypothesis tests you will perform over a scientific career. Indeed, if you run hypothesis tests on very different topics, then it seems appropriate to have a separate Type I error rate for each topic and no correction would be necessary.

A variation of the multiple testing situation is the following. You collected a sample with a certain hypothesis, which turned out to *not* be significant. You decide to test further hypotheses. For example, maybe you find no difference in memory for men and women on some task. You then decide to test whether younger women perform differently than older women and whether younger men perform differently than older men. For each of these hypotheses there is a risk of a False Alarm and you need to correct for it. Hence, asking too many questions can be problematic. Although these tests are not independent, a Bonferroni correction might effectively control the Type I error rate.

5.2 Dependent Tests

Equation 5.1 holds when all tests are independent. When you use one data set to try to answer many questions, the tests may not be independent because the data is being used multiple times. Although a Bonferroni correction might work to restrict Type I error, it may be overly conservative; but the impact depends on the nature of the dependencies.

For example, suppose we sample from gold fish in a pond and are interested whether larger tails predict larger hearts. By accident we sampled fish that suggest there is such a relationship whereas in fact the population does not have such a relationship: we picked a sample that produced a False Alarm. Now, we test a second hypothesis from the same sample, namely, that larger tails predict larger lungs. Suppose there is a perfect correlation between heart and lung size; our second analysis will produce another False Alarm.

Assume you are asking 10 questions about the fish in the pond. If you are unlucky you got the wrong sample and 10 wrong answers to your questions. In general, whether or not data are correlated is usually unknown, which is one more reason to abstain from asking more than one question about a sample.

5.3 How Many Scientific Results Are Wrong?

As mentioned, the Type I error is usually set to 5%. One might expect that hence 5% of all scientific results, where classic statistics is used, are wrong. However, this is not true. The statement would be true if the effect size is 0 ($\delta = 0$) for all experiments conducted. However, scientists usually aim for real effects, so for many experiments it is likely that there is actually a true effect and, thus, no chance of making a Type I error. Assume scientists conduct only experiments where there is a real effect. In this case there are no Type I errors, since the null hypothesis is wrong for all experiments. Hence, the number

of wrong scientific results depends on the incidence rate (see Chap. 1) of no effect. This number is largely unknown and, thus, we do not know how many results are False Alarms (or misses).

Take Home Messages

1. You can only ask one question for a set of data. Otherwise you need to account for multiple comparisons.
2. Keep your designs as simple as possible.
3. If you cannot keep your design simple and have more than one group comparison, read the next chapter.

ANOVA

6

Contents

What You Will Learn in This Chapter

In Chap. 3, we examined how to compare the means of two groups. In this chapter, we will examine how to compare means of more than two groups.

6.1 One-Way Independent Measures ANOVA

Suppose we wish to examine the effects of geographic region on tree heights. We might sample trees from near the equator, from the 49th parallel, and from the 60th parallel. We want to know whether the mean tree heights from all three regions are the same (Fig. 6.1). Since we have three regions we cannot use the t-test because the t-test only works if we are comparing two groups.

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_6

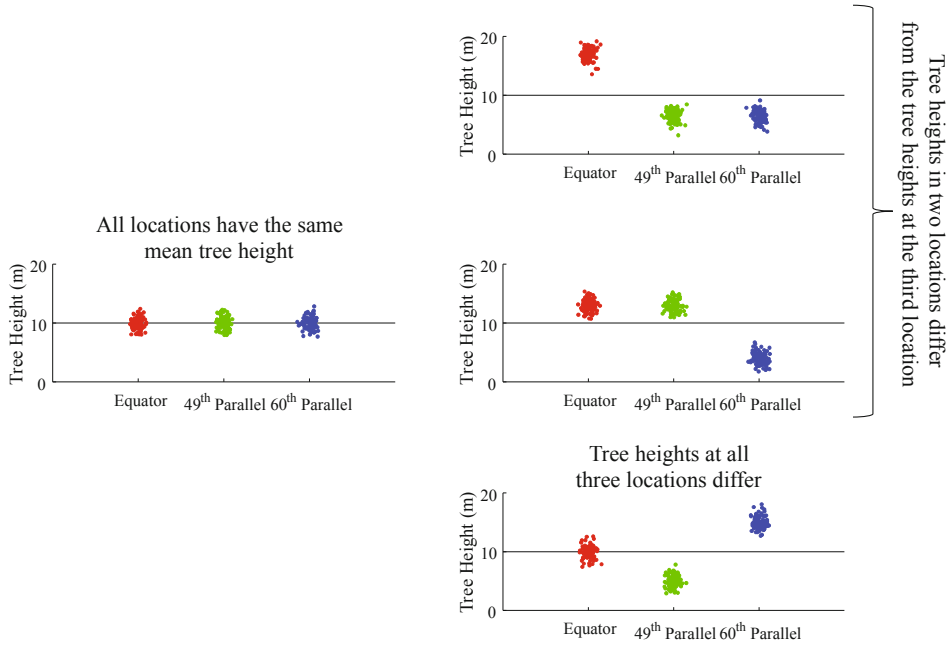


Fig. 6.1 We will examine the tree heights at three different latitudes. Left: the mean heights at all three latitudes is the same, shown by the horizontal line. Right: At least one latitude has a mean height that differs from the other two. The horizontal line shows the mean tree height for all three groups, called the " Grand Mean"

In principle we could compute three t-tests to compare all possible pairs of means (equator vs 49, equator vs 60, and 49 vs 60). However in this case, as shown in Chap. 5, we would face the multiple testing problem with the unpleasant side effect of increasing our Type I error rate as the number of comparisons increases. Situations like this are a case for an analysis of variance (ANOVA), which uses a clever trick to avoid the multiple testing problem.

6.2 Logic of the ANOVA

Terms

There are many terms for a 1-way ANOVA with m-groups:

- way = factor
- group = treatment = level

The logic of the ANOVA is simple. We simplify our alternative hypothesis by asking whether or not at least one of the tree populations is larger than the others. Hence, we are stating *one* hypothesis instead of three by lumping all alternative hypotheses together:

Null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

Lumped alternative hypotheses

$$H_1 : \mu_1 = \mu_2 \neq \mu_3$$
$$H_1 : \mu_1 \neq \mu_2 = \mu_3$$
$$H_1 : \mu_1 \neq \mu_3 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$$

The ANOVA assumes, similarly to the t-test, that all groups have the same population variances σ. If the null hypothesis is true, then the *population* means are equal for all three groups of trees. Any observed differences in the *sample* means come from the variance σ alone, which is due to random differences in tree heights (noise), but not to systematic differences in tree heights with geographic region (Fig. 6.1). It turns out that when the null hypothesis is true, the variability between means can be used to estimate σ (by multiplying by the sample sizes). An ANOVA compares this between means estimate to a direct estimate that is computed within each group.

Now assume that the mean tree heights in the three geographic regions are in fact different. In this case, the individual tree heights depend on both the variance within a group σ and the variability between the group means. In this case, the estimate of σ based on the variability between means tends to be larger than σ. In contrast, the estimate of σ based on the variability within each group tends to be similar to the true value. The ANOVA divides the two estimated variances and obtains an F-value:

$$F = \frac{\text{Variance estimate based on variability between group means}}{\text{Variance estimate based on variability within groups}}$$

Formally, this equation is expressed as:

$$F = \frac{\frac{\sum_{j=1}^{k} n_j (M_j - M_G)^2}{k-1}}{\sum_{j=1}^{k} \frac{\sum_{i=1}^{n_j} (x_{ij} - M_j)^2}{n_j - 1}}$$

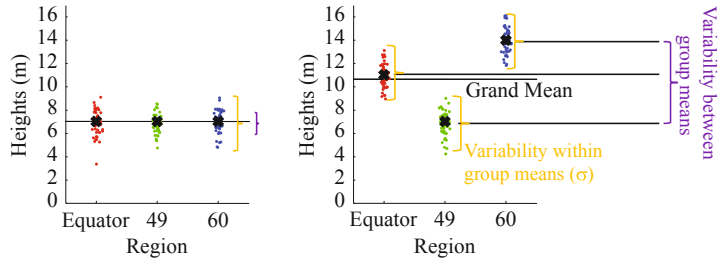


Fig. 6.2 Logic of the ANOVA. Left: the null hypothesis is true and, hence, all population means are identical. In this case, all the variability is within-group variability, which we consider to be the noise. The ANOVA assumes that variability in the data is the same for all three populations of trees. Right: the null hypothesis is false. Here, we show an extreme case where the variability of the means is larger than the variability of the data about the means. In this case, most of the variability in the data is explained by the effect of the three different regions on the tree heights. Usually the situation is somewhere in between these extremes. The null hypothesis for an ANOVA is that any observed differences are only due to the noise. The purpose of the ANOVA is to distinguish variability caused by the independent variable from variability of the data points around their individual treatment means

where k is the number of groups (three tree populations), n_j is the number of scores in group j (the number of trees within each sampled geographic region), M_j is the mean for group j (mean of geographic region sample j), M_G is the grand mean of all scores pooled together, and x_{ij} is the ith score for group j (the height of a single tree). To make it easier to distinguish the means from individual scores we use the symbols M_j and M_G rather than the traditional symbol for a sample mean $\bar{x}$. The multiplication by n_j in the numerator weights the deviations of the group means around the grand mean by the number of trees in each group so that the numbers of scores contributing to the variance estimates are equated between the numerator and denominator.

Consider two extreme examples. First, the null hypothesis is true (as in Fig. 6.2 left). In this case, the variance estimates for both the numerator and the denominator are similar and will produce an F-value that is close to 1. Next, let us consider an example of an alternative hypothesis where the differences between tree heights in the three geographic regions are large and σ is very small, i.e., the tree heights differ largely between the three populations but are almost the same within each population (as in Fig. 6.2 right). The variability in the measurements is mostly determined by the group differences and the F-value is large.

Just as in the t-test, a criterion is chosen for statistical significance to set the Type I error rate to a desired rate (e.g., $\alpha = 0.05$). When F exceeds the criterion, we conclude that there is a significant difference (i.e., we reject the null hypothesis of equality between the group means).

The tree example is a one-way ANOVA, where there is one factor (tree location) with three groups (regions) within the factor. The groups are also called levels and the factors are also called ways. There can be as many levels as you wish within a factor, e.g. many more regions, from which to sample trees. A special case is a one-way independent measures ANOVA with two levels, which compares two means as does the t-test. In fact, there is a close relationship between the two tests and in this case it holds that: $F = t^2$. The p-value here will be the same for the ANOVA and the two-tailed t-test. Hence, the ANOVA is a generalization of the t-test.

As with the t-test, the degrees of freedom play an important role in computing the p-value. For a one-way independent measures ANOVA with k levels, there are two types of degrees of freedom df_1 and df_2, respectively. In general, $df_1 = k - 1$ and $df_2 = n - k$ where n is the total number of sampled scores pooled over all groups, e.g., all trees in the three groups. The total of the degrees of freedom is $df_1 + df_2 = n - 1$.

6.3 What the ANOVA Does and Does Not Tell You: Post-Hoc Tests

Assume our ANOVA found a significant result. What does it tell us? We reject the null hypothesis that all means are equal:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

thereby accepting the alternative hypothesis, which can mean that any of the following are true:

$$H_1 : \mu_1 = \mu_2 \neq \mu_3$$
$$H_1 : \mu_1 \neq \mu_2 = \mu_3$$
$$H_1 : \mu_1 \neq \mu_3 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$$

By rejecting the H_0, we accept one of the alternative hypotheses—but we do not know which one. This is the price of avoiding multiple testing by lumping the four alternative hypotheses into one.

Here, the ANOVA offers a second trick. If we rejected the null hypothesis, it is appropriate to compare pairs of means with what are called "post-hoc tests," which, roughly speaking, corresponds to computing pairwise comparisons. Contrary to the multiple testing situations discussed in Chap. 5, these multiple comparisons do not inflate the Type I error rate because they are only conducted if the ANOVA finds a main effect.

There are many post-hoc tests in the statistical literature. Commonly used post-hoc tests include: Scheffé, Tukey, and REGW-Q. The process is best described with an example, which is provided at the end of this chapter.

6.4 Assumptions

The assumptions of the ANOVA are similar to the assumptions for the t-test described in Chap. 4.

1. Independent samples.
2. Gaussian distributed populations.
3. The independent variable is discrete, while the dependent variable is continuous.
4. Homogeneity of variance: All groups have the same variance.
5. The sample size needs to be determined before the experiment.

6.5 Example Calculations for a One-Way Independent Measures ANOVA

6.5.1 Computation of the ANOVA

Suppose there is a sword fighting tournament with three different types of swords: light sabers, Hattori Hanzo katanas, and elvish daggers (see Fig. 6.3). We are asking whether there are differences in the number of wins across swords. Hence, our null hypothesis is that there is no difference. The data and the computation of the F-value are shown in Fig. 6.3.

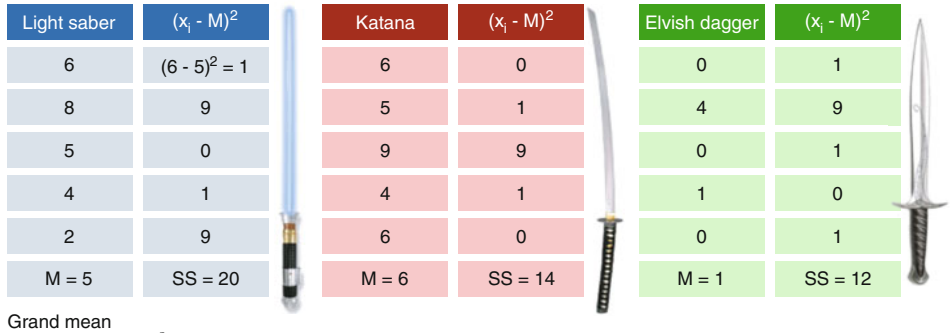
Our final[1] F-value is 9.14. This means that the variability of the group means around the grand mean is 9.14 times the variability of the data points around their individual group means. Hence, much of the variability comes from differences in the means, much less comes from variability within each population. An F-value of 9.14 leads to a p-value of $0.0039 < 0.05$ and we conclude that our results are significant, i.e., we reject the null hypothesis that all three sword types yield equal mean numbers of wins ($F(2, 12) = 9.14$, $p = 0.0039$). Furthermore, we can conclude that at least one sword type yields a different number of wins than the other sword types. We can now use one of the various post-hoc tests to find out which sword(s) is/are superior.

[1] If the data are analyzed by a statistics program, you will get $F = 9.13$. The difference is due to rounding of MS_{Within} in Fig. 6.3.

Light saber	$(x_i - M)^2$
6	$(6 - 5)^2 = 1$
8	9
5	0
4	1
2	9
M = 5	SS = 20

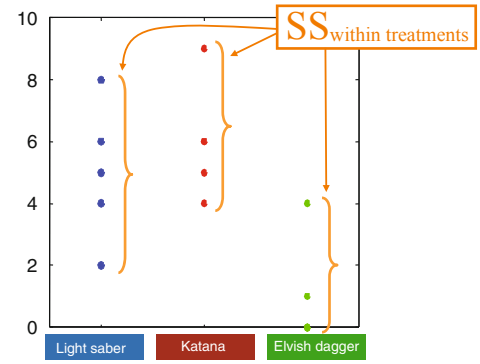
Katana	$(x_i - M)^2$
6	0
5	1
9	9
4	1
6	0
M = 6	SS = 14

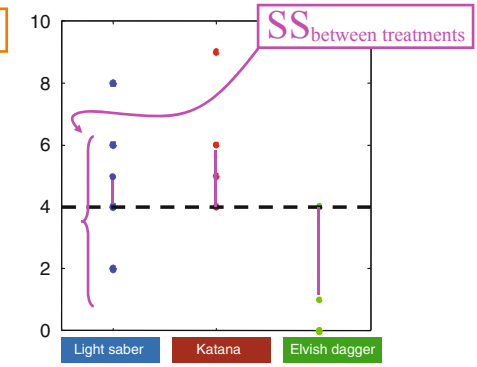
Elvish dagger	$(x_i - M)^2$
0	1
4	9
0	1
1	0
0	1
M = 1	SS = 12



Grand mean

$$M_G = \frac{\sum_{k=1}^{3}\sum_{i=1}^{n_k} x_{ik}}{N} = \frac{6+8+5+4+2+6+5+9+4+6+0+4+0+1+0}{15} = 4$$





$$SS_{within} = \sum_{k=1}^{3} SS_k = 20 + 14 + 12 = 46$$

$$SS_{between} = \sum_{k=1}^{3} n_k (M_k - M_G)^2 = 5(5-4)^2 + 5(6-4)^2\,5(1-4)^2 = 70$$

$$MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{46}{12} = 3.83$$

$$MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{70}{2} = 35$$

$$F = \frac{MS_{between}}{MS_{within}} = \frac{35}{3.83} = 9.14$$

Source	SS	df	MS	F	p
Between treatments	70	2	35	9.14	0.0039
Within treatments	46	12	3.83		

Fig. 6.3 Example calculations for a one-way independent measures ANOVA. Each of the three swords is used by five independent fighters, making up all together 15 fighters. Hence, we have a 1×3 ANOVA. The opponents in the fights are not from the 15 test fighters. The upper left panel shows how many fights were won with the swords. The table below shows the data by numbers. First, we compute the mean for each sword type. For example, with the light sabers five wins occurred on average. Next, we compute the variability for each sword, also called the sum of squares inside a

6.5.2 Post-Hoc Tests

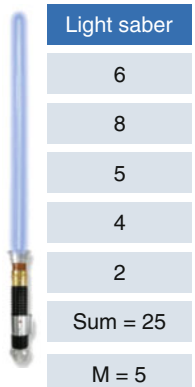
Various procedures exist for performing post-hoc tests, but we will focus here on the Scheffé test in order to illustrate some general principles.

The idea behind the Scheffé test is to perform multiple comparisons by computing pairwise ANOVA' s (e.g., light sabers vs. katanas, light sabers vs. elvish daggers, and katanas vs. elvish daggers). One assumption of the ANOVA is that all populations have equal variances. If this is true, then the best estimate of the variability within each population is the pooled estimate from the overall ANOVA calculated by MS_{within} (i.e., 3.83 in this case). The Scheffé test also uses $df_{between}$ from the overall ANOVA, and the calculations for performing this test are illustrated in Fig. 6.4.

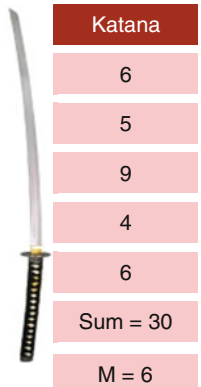
The p-value for each of the comparisons is computed using the degrees of freedom from the original ANOVA (i.e., $df_{between} = 2$ and $df_{within} = 12$). This yields the results in Table 6.1 for our post-hoc tests. Only the second and third comparisons are below our critical threshold of $\alpha = 0.05$, thus, we can conclude that the light sabers differ from the elvish daggers ($F(2, 12) = 5.22$, $p = 0.023$), and that katanas also differ from elvish daggers ($F(2, 12) = 8.15$, $p = 0.006$), but that we failed to find a significant difference between light sabers and katanas ($F(2, 12) = 0.33$, $p = 0.728$).

A common way to illustrate these differences is to plot a graph showing the mean number of wins for the three sword types with error bars denoting standard errors around each mean, and lines connecting the significantly different sword types with asterisks above them (Fig. 6.5).

Fig. 6.3 (continued) treatment. For the computation, we subtract each data point from the mean and square the value. To compute the variance within groups, we add the three sums of squares. In this case, we arrive at a value of 46. The next main step is to compute the variability between means. For this, we compute first the Grand Mean M_G, which is simply the mean number of wins over all 15 sword fights. In this example, the Grand Mean is 4. Next, for each sword, we subtract the means from the Grand Mean, square the values and multiply with the number of fights for each sword type (five in this example). We arrive at a value of 70. Next we divide our two sum of squares values by the degrees of freedom df_1 and df_2 in order to arrive at variances. We had three types of swords, hence, $df_1 = 3 - 1$, so we divide 70 by 2. We had 15 fights, hence, $df_2 = 12$, so we divide 46 by 12 (MS means mean square). For the test statistic, we divide the two values of 35 and 3.83 and arrive at $F = 9.14$. As with the t-value, the F-value can easily be computed by hand. For the p-value, we use statistics software, which delivers $p = 0.0039$. The output of software packages summarize the computation in panels similar to the one shown here. In publications, a result like this is presented as ($F(2, 12) = 9.14$, $p = 0.0039$)



Light saber
6
8
5
4
2
Sum = 25
M = 5



Katana
6
5
9
4
6
Sum = 30
M = 6

Elvish dagger
0
4
0
1
0
Sum = 5
M = 1

$$df_{between} = 2$$
$$MS_{within} = 3.83$$

From the original ANOVA table

$$SS_{between} = \sum_{k \in comp} n_k \left(M_i - G_{comp}\right)^2$$

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

$$F = \frac{MS_{between}}{MS_{within}}$$

Light saber versus Katana

$$G_{comp} = \frac{25 + 30}{10} = 5.5$$

$$SS_{between} = 5\left(5 - 5.5\right)^2 + 5\left(6 - 5.5\right)^2 = 2.5$$

$$MS_{between} = \frac{2.5}{2} = 1.25$$

$$F = \frac{1.25}{3.83} = 0.3264$$

Light saber versus Elvish dagger

$$G_{comp} = \frac{25 + 5}{10} = 3$$

$$SS_{between} = 5\left(5 - 3\right)^2 + 5\left(1 - 3\right)^2 = 40.0$$

$$MS_{between} = \frac{40}{2} = 20$$

$$F = \frac{20}{3.83} = 5.2219$$

Katana versus Elvish dagger

$$G_{comp} = \frac{30 + 5}{10} = 3.5$$

$$SS_{between} = 5\left(6 - 3.5\right)^2 + 5\left(1 - 3.5\right)^2 = 62.5$$

$$MS_{between} = \frac{62.5}{2} = 31.25$$

$$F = \frac{31.25}{3.83} = 8.1593$$

Fig. 6.4 Calculations for Scheffé post-hoc comparisons on our sword example. First we calculate the grand mean for each comparison (G_{Comp}), which consists of the mean of the pooled scores for the pairs of groups being considered. Next, using this mean we compute the sum of squared deviations between group means and the grand mean for the two groups being ($SS_{Between}$). The $MS_{between}$ scores are then obtained by dividing by $df_{between}$ from the overall ANOVA (i.e., two in this case). Finally, the F-ratio for each comparison is computed by dividing by the MS_{within} term from the overall ANOVA (i.e., 3.83 in this example)

Table 6.1 Post-hoc Scheffé test results for our three comparisons

Comparison	Result
1 vs. 2[a]	$F(2,12) = 0.33$, $p = 0.728$
1 vs. 3[b]	$F(2,12) = 5.22$, $p = 0.023$
2 vs. 3[c]	$F(2,12) = 8.16$, $p = 0.006$

[a]Light sabers versus katanas
[b]Light sabers versus elvish daggers
[c]Katanas versus elvish daggers

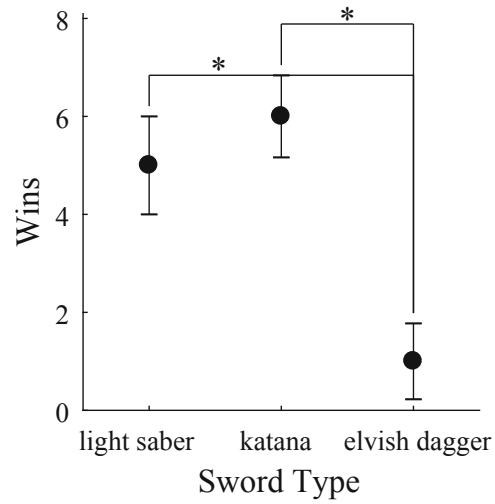


Fig. 6.5 Mean numbers of wins for the three sword types plotted with standard errors and lines with asterisks above them connecting sword types that are significantly different

6.6 Effect Size

As with the t-test, the p-value from an ANOVA confounds the effect size and the sample size. It is always important to look at the effect size, which for an ANOVA is denoted by η^2. The effect size η^2 tells you the proportion of the total variability in the dependent variable that is explained by the variability of the independent variable. The calculation is:

$$\eta^2 = \frac{SS_{between}}{SS_{total}}$$

with

$$SS_{between} = \sum_{j=1}^{k} n_j(\bar{x}_j - M_G)^2 \tag{6.1}$$

$$SS_{total} = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - M_G)^2 \tag{6.2}$$

Table 6.2 Effect size guidelines according to Cohen

	Small	Medium	Large
Effect size	0.01	0.09	0.25

where M_G is the grand mean (i.e., average over all data points). This ratio tells you the proportion of the total variability in the data explained by variability due to the treatment means. For the above sword example the effect size is $\eta^2 = 0.60$, which is a large effect according to Cohen, who provided guidelines for effect sizes (Table 6.2).

6.7 Two-Way Independent Measures ANOVA

The one-way independent measures ANOVA generalizes nicely to cases with more than one factor. Here, we will discuss the simplest of such cases, the two-factor design.

Suppose that you and your friends acquire super powers in a science experiment and you are preparing to start your life of fighting crime. You and your super hero friends do not want your enemies to hurt your loved ones so you need costumes to conceal your identity. Furthermore, sometimes you might fight crime during the day, and other times you might fight crime at night. You want to know which costume material (spandex, cotton, or leather) will be the best for crime fighting as measured by the number of evil villains a hero can catch while wearing costumes made from each material, and you want to know if there is an effect of time of day on which material is best. You assign each friend to a costume material and time of day condition and count the number of evil villains each hero catches. You have different friends in each group. In this case, there are three separate hypotheses that we can make about the data:

1. H_0: There is no effect of time of day on the number of villains caught.
   H_1: The number of villains caught during the day are different from the number of villains caught at night.
2. H_0: There is no effect of costume material on the number of villains caught.
   H_1: At least one costume material yields different numbers of villains caught than the other costume materials.
3. H_0: The effect of time of day on the number of villains caught does not depend on costume material.
   H_1: The effect of time of day on the number of villains caught does depend on costume material.

The first two null hypotheses relate to what are called *main effects*. The two main hypotheses are exactly the same as computing two one-way ANOVAs. The third hypothesis is a new type of hypothesis and pertains to the *interaction* between the two factors, costume and day time. To measure the main effect of costume material, we take the average number of villains caught in the spandex group, averaging over both day and

night conditions, and compare this with the same averages for the cotton and leather costume conditions. To measure the main effect of time of day, we look at the average number of villains caught for the day condition, averaging over the spandex, cotton, and leather costume material conditions, and compare this with the same average for the night condition.

For the interaction, we consider all groups separately, looking at the number of villains caught for spandex, cotton and leather costume groups separately as a function of day- and night-time crime-fighting conditions. If there is a significant interaction, then the effects of time of day on the number of villains caught will depend on which costume material we are looking at. Conversely, the effect of costume material on the number of villains caught will depend on which time of day our friends are fighting crime at.

Testing these three null hypotheses requires three separate F-statistics. Each F-statistic will use the same denominator as in the one-way ANOVA (i.e., the pooled variance of the data about the treatment means, or MS_{within} as shown in Fig. 6.3), but the numerators ($MS_{between}$) will be specific for the particular hypotheses tested.

Figure 6.6 shows example raw data and the means being compared for the three hypotheses being tested (see margins). When pooling over time of day it looks like costume material has very little effect on crime fighting abilities. When pooling over costume material, it looks like time of day has also little effect on crime fighting abilities. It is only when we consider each mean individually that we can see the true effects of time

	Spandex	Cotton	Leather	Time of Day Means
Day	18	10	3	9.7
	10	8	5	
	16	12	1	
	12	6	7	
	14	14	9	
Day Means	14	10	5	
Night	5	6	15	10
	7	14	13	
	3	10	17	
	9	8	11	
	1	12	19	
Night Means	5	10	15	
Costume Means	9.5	10	10	Grand Mean = 9.8

Fig. 6.6 Number of villains caught for each super hero. As well as means for main effects (time of day and costume type), and individual cell means (spandex during the day, spandex at night, cotton during the day, etc.). The grand mean is the mean overall data points

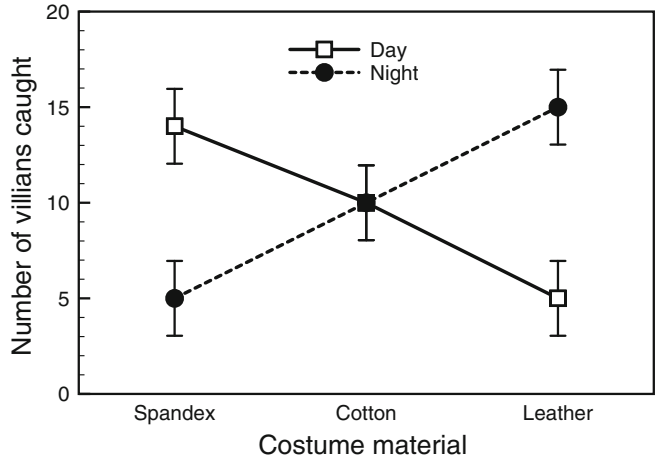


Fig. 6.7 Interaction plot for the number of villains caught as a function of costume material and time of day. Here we can see that the effect of costume material on the number of villains caught depends critically on the time of day

of day and costume material on how many villains our friends are catching: there is an interaction between costume material and time of day in relating the number of villains caught (Fig. 6.7). This interaction is such that spandex is better during the day and leather better at night, with cotton always being somewhere in the middle.

This example illustrates the value of a two-factor design. Had we done only separate one-way ANOVAs examining the relationships between costume material and number of villains caught, or time of day and number of villains caught, we would have found little or no effects. Including both variables reveals the true nature of both effects, showing the effect of one to depend on the level of the other. Figure 6.8 demonstrates three possible outcome patterns that isolate just one significant effect (Main effect of A, Main effect of B, Interaction) without any of the other effects. It is also possible to have combinations of main and interaction effects.

Another virtue of a two-factor design relative to a one-factor design is that variability that would otherwise be included in the error term (i.e., MS_{within}) is now partly explained by variability due to another factor, thereby reducing MS_{within} and increasing the power for detecting effects when present.

Thus, it may seem that the more factors we add the better we will understand the data and obtain significant results. However, this is not true because we lose power for each factor we are adding due to the fact that we have fewer scores contributing to each mean. Typically, larger samples are needed when the number of factors increases.

Importantly, if we find a significant interaction, the main effect varies depending on the other factor. Thus, we should usually refrain from making conclusions about the main effect if there is an interaction.

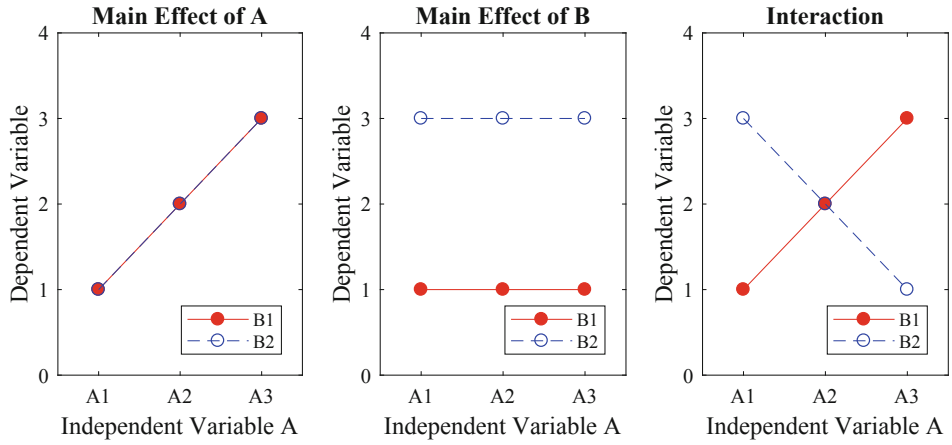


Fig. 6.8 The results of a two-factor ANOVA may reveal three general types of significant results: a main effect of variable A, a main effect of variable B, and an interaction between A and B. In the following example, we have a 2×3 ANOVA, where A1, A2, A3 may indicate the superhero's costume materials (Spandex, Cotton, Leather) and B1 and B2 times of day (night and day, respectively). The dependent variable would be villains caught. Left. Main effect of A. The costume material matters. More villains are caught while wearing leather than cotton. Time of day plays no role. As many villains are caught during the day as during the night. For this reason B1 and B2 are on top of each other. Center. Main effect of B. The costume material does not matter but the time of day does. More villains are caught during the day. Right. Interaction as in Fig. 6.7. In the case of a significant interaction, significant main effects are typically not examined because the more relevant comparison involves looking at the effect of one variable within the levels of the other variable

The one-way ANOVA avoids the multiple testing problem. However, a multi-way ANOVA reintroduces a kind of multiple testing problem. For example, consider a 2×2 ANOVA, with a significance criterion of 0.05. A truly null data set (where all four population means are equal to each other) has a 14% chance of producing at least one $p < 0.05$ among the two main effects and the interaction. If you use ANOVA to *explore* your data set by identifying significant results, you should understand that such an approach has a higher Type I error rate than you might have intended.

A typical statistical software package outputs the results of a two-way ANOVA as in Table 6.3.

6.8 Repeated Measures ANOVA

The ANOVA we have discussed up to now is a straightforward extension of the independent samples t-test. There also exists a generalization of the dependent samples t-test called the *repeated measures* ANOVA. You can use this kind of ANOVA when, for

Table 6.3 Typical statistical software outputs for a two-way ANOVA

Source	SS	df	MS	F	p	η^2
Costume material	1.67	2	0.83	0.083	0.920	0.0069
Time of day	0.83	1	0.83	0.083	0.775	0.0035
Costume × time	451.67	2	225.83	22.58	0.000003	0.6530
Error	240.00	24	10.00			

The columns show the source of variability (Source), the sums of squares (SS), the degrees of freedom (df), the mean squares (MS), the F-ratios (F), the p-values (p—sometimes also labeled "Sig."), and the effect sizes (η^2). The row labeled "Error" holds the computations for the within subjects variability, while the remaining rows show between subjects variability for the main effects and interactions

Table 6.4 Typical statistical software outputs for a repeated measures ANOVA

Source	SS	df	MS	F	p	η^2
Between times	70	2	35	70	0.00000009	0.94
Within times	40	12				
Between subjects	36	4				
Error	110	14				

Here the example is for patient symptoms measured before, during, and after treatment (i.e., at different times). The first row (*Between times*) shows the effect of time of measurement on symptoms. The within times row shows the variability due to subjects within each time condition. It is broken down into consistent trends for each individual subject (*Between subjects*) and random error caused by things like drug diffusion rates (*Error*). The columns show the source of variability (Source), the sums of squares (SS), the degrees of freedom (df), the mean squares (MS), the F-ratios (F), the p-values (p—sometimes also labeled "Sig."), and the effect sizes (η^2). The row labeled " Error" holds the computations for the error term which is used in the denominator of the F-ratio. In the independent measures ANOVA this term is the within treatments term. Here, however, we remove from the within treatments term the variability due to subjects, so we now simply call the remaining variability " error variability." The remaining rows show between subjects variability for the main effects and interactions. To summarize these results we would say that there is a significant effect of time of measurement on symptoms $F(2, 14) = 70$, $p = 0.00000009$. Here, we have taken the degrees of freedom from the between times and error rows, and have taken the F- and p-values from the between times row

example, measuring some aspect of patient health before, during, and after some treatment program. In this case, the very same patients undergo three measurements. A repeated measures ANOVA has higher power than the independent measures ANOVA because it compares the differences within patients first before comparing across the patients, thus, reducing variability in the data. Example output from a repeated measures ANOVA is provided in Table 6.4.

Take Home Messages

1. With an ANOVA you can avoid the multiple testing problem—to some extent.
2. More factors may improve or deteriorate power.

Experimental Design: Model Fits, Power, and Complex Designs

7

Contents

What You Will Learn in This Chapter

An ANOVA is one option to cope with the multiple testing problem. A much simpler way to cope with multiple testing is to avoid it by clever experimental design. Even if you need to measure many variables, there is no need to subject all of them to a statistical test. As we will show in this chapter, by collapsing many data into one meaningful variable or simply by omitting data, you may increase your statistical power. Simple and simplified designs are also easier to interpret, which can be a problem in many complex designs. In this chapter, we also show how to compute the power of an experiment, which is for example important to determine the sample size of your experiment.

7.1 Model Fits

When comparing two means, a *t*-test has a high power and is straightforward to interpret. Experiments with more group comparisons suffer from the multiple testing problem. The more comparisons we compute, i.e., the more groups or levels there are, the lower is the power. Experiments with more groups are also more complex to analyse because interactions can occur, which are not present in simple *t*-tests (Chap. 6). Hence,

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_7

simple experimental designs are usually preferred. However, complexity is sometimes unavoidable. A classic example is a learning experiment, where performance needs to be measured at many time points. For example, participants train on a simple visual task for 10 blocks containing 80 trials each. For each block, we determine the percentage of correct responses (Fig. 7.1a) and look for increases in performance across blocks. How can we quantify the learning success and compute statistics? The null hypothesis is that no learning occurs, i.e., performance is identical for all 10 blocks. Intuitively, one might think of using a repeated measures ANOVA with 10 levels, one for each block. However, this is not a good idea because, first, ANOVAs are about nominal variables, i.e., the order of the blocks plays no role. Second, were performance increasing for the first five blocks and then decreasing, the ANOVA would indicate a significant result when performance for block 5 is significantly different from block 1. However, a result like this is not about learning but a strange concatenation of learning and unlearning. Third, we would lose quite some power. What to do? Here, we show it is by no means necessary to subject all data to statistical analysis.

As shown in Fig. 7.1b for the learning experiment, one approach is to discard all blocks except for the first and the last one (the intermediate blocks are relevant to the experiment because they promote learning, but they are not relevant to the statistical analysis). The null hypothesis is that performance in these two blocks does not differ. We can use a repeated measures t-test to test this hypothesis. However, learning data are often noisy and thus we are losing power with this procedure. To obtain less noisy data, we may average the first and last two blocks and subject the two averages to a repeated measures t-test (Fig. 7.1c).

In both cases, we are discarding a large amount of data and thus do not take full advantage of our data. We might do better by fitting a model to the data. We may know, for example, from previous experiments that learning is reflected by a linear increase in performance, which we can model by the equation $mx + b$, where m is the slope of the learning curve, b is the y-intercept, and x is the block number. We can use a computer program to compute the optimal parameters for m and b, for each observer individually. Since we are only interested in the slope, we can discard b. Our null hypothesis is: $m = 0$. Hence, for each observer we obtain one m-value. If 12 observers joined the experiment, we compute a one-sample t-test with these 12 values of m and see whether they are significantly different from 0.

There is great flexibility in this approach. For example, if learning is not linear but follows an exponential rather than a linear function, then we can fit an exponential function, which also contains a "slope" parameter. When we are interested in cyclic processes, such as changes in temperature across a day or the numbers of insects across a year, we can fit a sine function. In general, we can fit any type of function to our data and extract one or a few parameters. We thus take full advantage of the data and do not lose power. It is the choice of the experimenter how to proceed. However, the experimenter must make the choice before the experiment is conducted. One cannot decide after having looked at the data and then try many possibilities until finding a significant result (see Sect. 11.3.5).

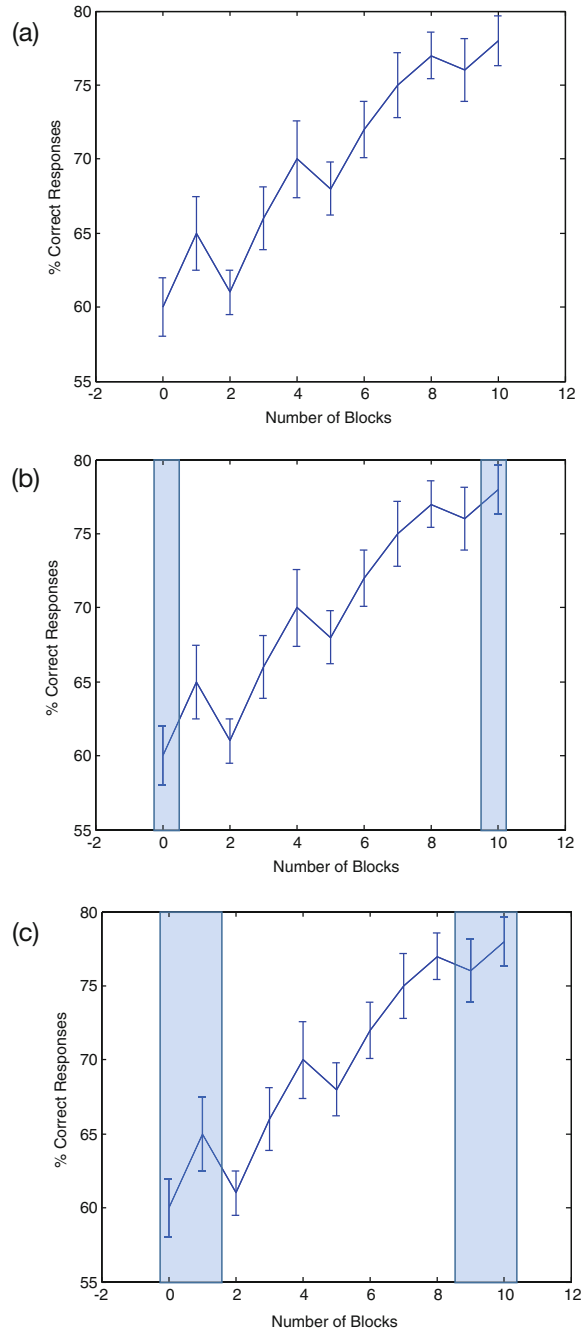


Fig. 7.1 Analyzing learning data. (**a**) Performance improves with number of blocks. (**b**) A statistical analysis might just compare the first and last blocks. (**c**) Alternatively, the analysis might average the first two and last two blocks and then compare the averages

The above example shows how to simplify statistics by reducing the number of variables. As shown, there is no need to subject all your data in its original form to statistical analysis. There are no general rules on how to simplify your analysis because each experiment is different. However, it is always a good idea to think about what is the main question your experiment is aimed to answer. Then, you decide what variables address the question best and how you can compute statistics. The simpler the design and the fewer variables, the better.

7.2 Power and Sample Size

7.2.1 Optimizing the Design

It often takes a lot of effort and resources to run an experiment. Thus, it is usually worthwhile to estimate whether the experiment is likely to succeed and to identify sample sizes that provide a high probability of success (if there is actually an effect to detect). Success generally means producing a large t-value and obtaining a significant result. We can do this in a couple of ways.

First, try to increase the population effect size $\delta = \frac{\mu_1 - \mu_2}{\sigma}$. You can do this by considering situations where you anticipate the difference between population means to be large. For example, you may try to find the optimal stimuli for a visual experiment or the most discriminative tests for a clinical study.

In addition, try to reduce σ. It is the ratio of the population mean differences and the standard deviation that determines δ, and thus t and p. You may try to make your measuring devices less noisy, for example, by calibrating every day. You may try to homogenize the sample, for example, by testing patients at the same time every day, making their coffee consumption comparable, using the same experimenter etc. You may think about excluding certain patients, for example, by imposing age limits to not confuse deficits of a disease with age effects. However, such stratifications limit the generality of your research (see Chap. 3, Implications 4). There are many ways to reduce σ and it is always a good idea to think about it.

Second, increase the sample size, n. Even if δ happens to be small, a large enough sample will produce a large t-value. With a large enough sample size it will be possible to discriminate even small differences between means (signal-and-noise) from a situation where there is actually no difference between means (noise-alone). Note, for this approach to be meaningful, you have to be confident that a small effect size matters. There is no point in running a large sample study to detect a trivially small effect (see Chap. 3, Implications 1 and 2).

7.2.2 Computing Power

Even when $\delta \neq 0$, experiments do not always produce significant results because of undersampling (Chap. 3). Here, we show how likely it is that for a given $\delta \neq 0$ and a given sample size n a significant result occurs. Vice versa, we show how large n needs to be to produce a significant result with a certain probability.

We estimate an experiment's success probability by computing power. Power is the Hit rate. It is the probability of selecting a random sample that allows you to correctly reject the null hypothesis. It supposes that the null hypothesis is false, meaning that there is a non-zero effect. As noted in Chap. 3, computing power requires a specific population standardized effect size. Where this specific population effect size comes from is situation-specific. Sometimes it can be estimated from other studies that have previously investigated the same (or a similar) phenomenon. Sometimes it can be derived from computational models that predict performance for a novel situation. Instead of predicting an effect size, it is sometimes worthwhile to identify a value that is deemed to be interesting or of practical importance.

Once a population effect size is specified, we turn to computer programs to actually calculate power (there is no simple formula). Figure 7.2 shows the output of a free program called G*Power. Here, we selected t-test from the *Test family* and a *Statistical test* of a difference between two independent samples for means. For *Type of power analysis* we selected "Post hoc." Under *Input parameters* we selected for a two-tailed test, entered an estimated population effect size $d = 0.55$, chose our Type I error rate to be $\alpha = 0.05$, and entered planned sample sizes of $n_1 = n_2 = 40$. The program provides graphs at the top and *Output parameters* on the bottom right. The graphs sketch the sampling distributions (see Fig. 3.7) that should be produced by the null (red curve) and the specific alternative hypothesis (blue curve). The shaded blue area is labeled β to indicate the Type II error rate. This is the probability for a non-significant result if $\delta = 0.55$. Power is the complement of the Type II error rate. As indicated, for the provided input parameters, the computed power is 0.68. This means that there is a probability of 0.68 that under the specified conditions you will obtain a significant result.

Suppose that we were unsatisfied with the 0.68 probability and wanted to identify sample sizes that would have a 90% chance of rejecting the null hypothesis. From the *Type of power analysis* menu, we select "A priori" and in the revised *Input parameters* we change the Power value from 0.68 to 0.9. Figure 7.3 shows the program output for the new situation. In the *Output parameters* panel, we see that the sample sizes needed to have a power of 0.9 for a two-tailed, two-sample t-test, when the population effect size is $\delta = 0.55$ are $n_1 = n_2 = 71$.

In general, for a given population effect size, one can identify the smallest sample sizes so that an experiment has the specified power. Calculating such sample sizes is an important part of experimental design. It usually makes little sense to run an experiment without knowing that it has a reasonable probability of success, i.e., reasonable power. Unfortunately, many scientists run experiments without doing a power analysis because

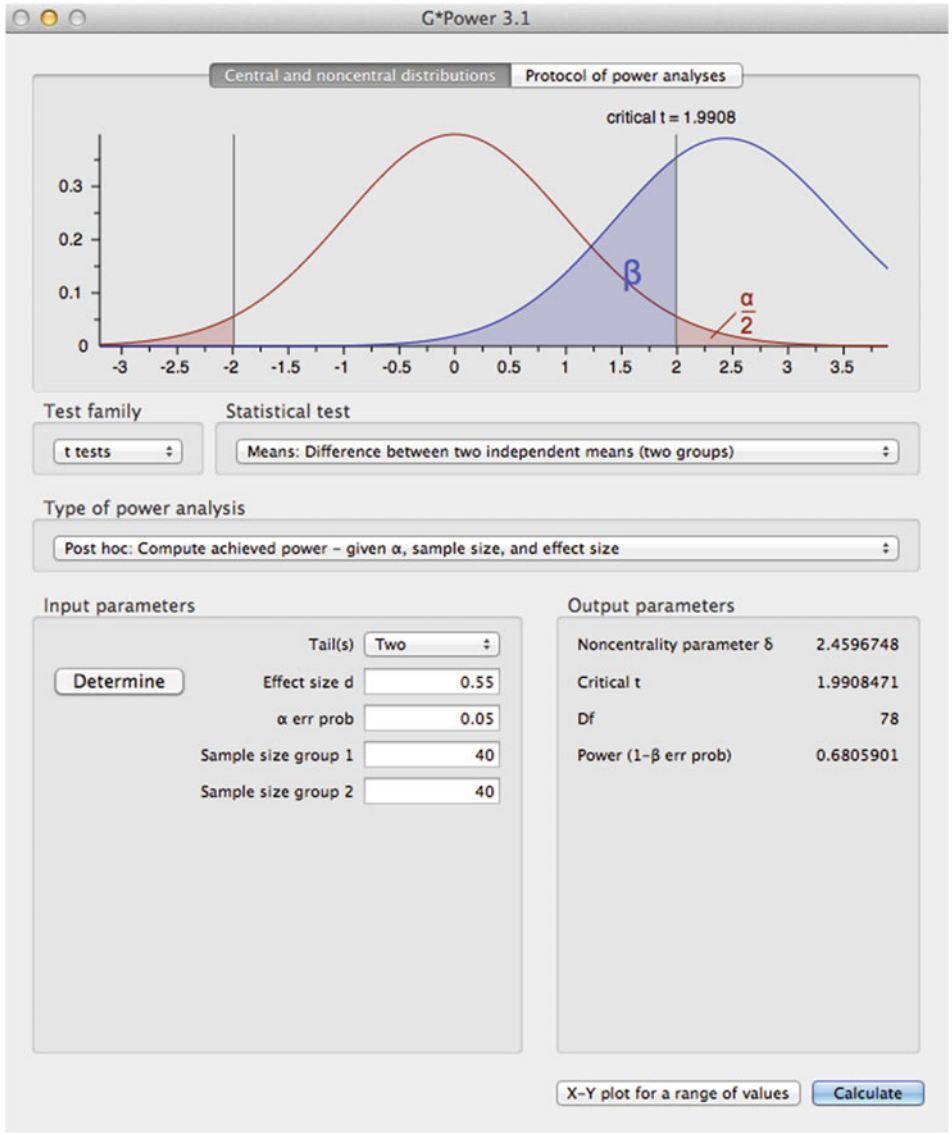


Fig. 7.2 Output from the G*Power program to compute power for a t-test with specified sample sizes. In this case, the effect size (0.55) and the sample sizes ($n_1 = n_2 = 40$) are known and we are searching for power, i.e., how likely it is that we obtain a significant result with this effect and sample size for an independent t-test and $\alpha = 0.05$. The output parameters are the noncentrality parameter δ, which is not the same as the population effect size and we ignore it here, the critical t-value, the degrees of freedom Df and, most importantly, the power

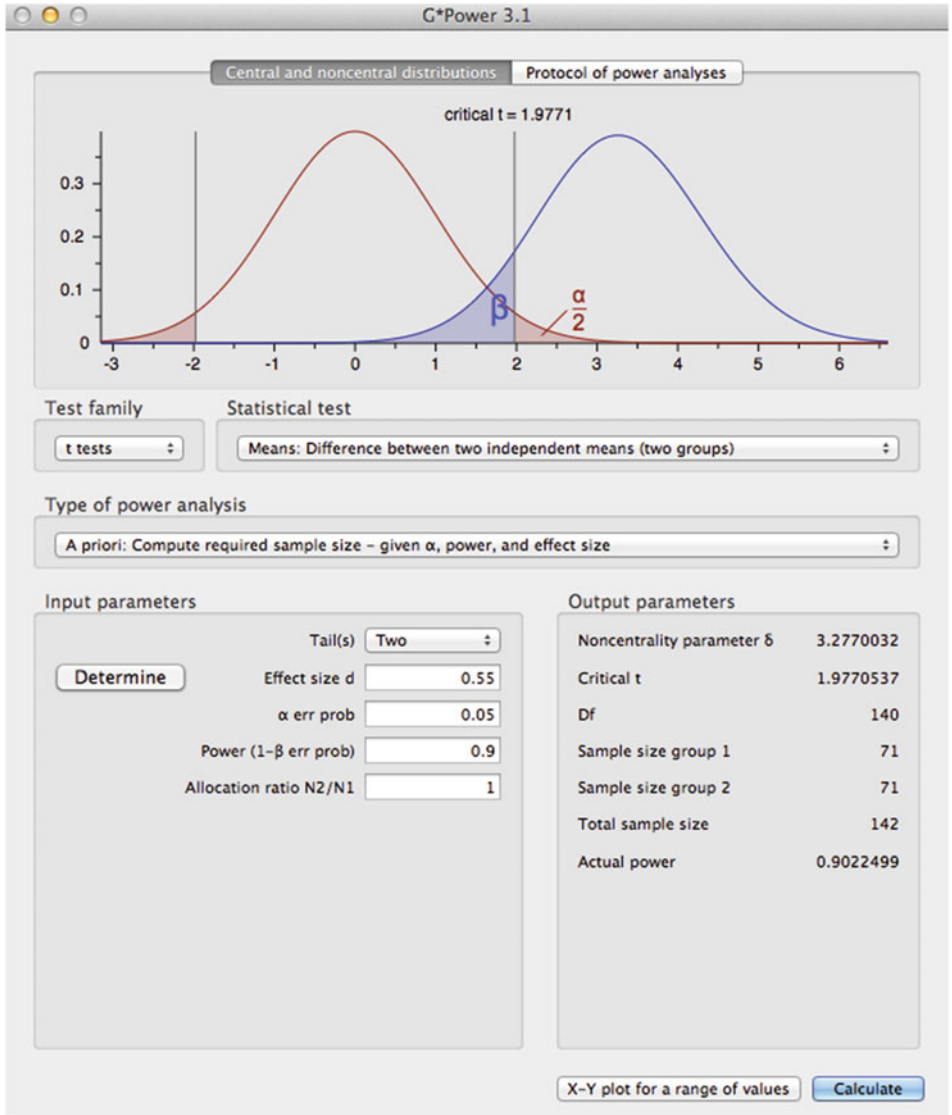


Fig. 7.3 Output from the G*Power program to compute the necessary sample sizes so that an experiment has a *t*-test with at least 90% power. In this case, the effect size (0.55) is known or desired and we are searching for the sample size to obtain a significant result with a probability of 0.9

they do not have a particular effect size in mind. Such investigations might turn out to be valuable, but whether they work or not is largely a matter of luck. If you cannot perform a meaningful power analysis (with a justified effect size), then the best you can do is "hope" that your experiment produces a significant outcome. If the experiment fails to produce a significant outcome, you can hardly be disappointed because you never really had any (quantitative) reason to expect that your sample was large enough to show an effect. Many times scientists are doing exploratory work even when they think they are doing confirmatory work. Confirmatory work almost always is built on knowledge about an effect size that can be used to design an experiment with high power.

7.3 Power Challenges for Complex Designs

Ideally, a power analysis is done before gathering any data; this is called a priori power. However, it is also possible to estimate power in a post hoc fashion by using the sample sizes and the estimated effect size from the data. For simple cases (e.g., a two-sample t-test) the post-hoc power analysis does not tell you anything beyond the test for significance. If you use G*Power to calculate power for different combinations of t and sample sizes you will discover that, if your t-test gives you $p > 0.05$ then your power calculation will be less than 0.5. Likewise, if your t-test gives you $p < 0.05$, then your power calculation will be greater than 0.5. If your t-test gives you $p = 0.05$, then your power calculation will give approximately 0.5. Here, we show that post hoc power calculations can be more useful for complicated statistical analyses that involve multiple tests on a set of data.

We saw above how to use G*Power to compute power for simple experimental designs. This program, and similar alternatives, tends to focus on just one statistical test at a time. In practice, scientists often use a combination of statistical tests to argue for a theoretical interpretation. Estimating power for a combination of statistical tests often requires generating simulated data sets that correspond to the experiment's design and sample sizes. This simulated data is then analyzed in the same way that the experimental data will be analyzed. By repeating this process thousands of times, one can simply count how often the full set of statistical outcomes matches the outcomes needed to support a theoretical claim. This simulation approach allows a researcher to consider "success probability", which generalizes the concept of power.

We will see that complex experimental designs with multiple tests can struggle to have high power. Even if individual tests have reasonable power, it can be the case that the full set of tests has low power.

To demonstrate this generalization, it may be helpful to consider a concrete example. The example is purposely complicated because the complications highlight important characteristics of power analyses. A prominent study published in 2017 reported empirical evidence that performance on a memory task was related to breathing through the nose. The motivation for the study was that nasal breathing can entrain the hippocampus of

the brain, which is related to memory processing. In contrast, oral breathing does not entrain the hippocampus and so should not influence memory performance. Subjects were asked to breath either orally (through the mouth) or nasally (through the nose) while viewing pictures during a memory encoding phase, and then during a retrieval test subjects identified pictures they had seen before. During both the encoding and retrieval phases the pictures were presented at random times so that sometimes the picture was presented while the subject was inhaling and sometimes the picture was presented while the subject was exhaling. The main conclusion was that identification accuracy was better for pictures that were presented to nasal breathers during inspiration (breathing in). This was true for encoding pictures and for retrieving pictures. In contrast, oral breathers showed no significant effect of inward versus outward breathing.

The study and its analysis is rather complicated, so it is useful to characterize all the hypothesis tests. For convenience, we also list the relevant statistics from the study. All tests compared memory performance of subjects.

1. Nasal breathers ($n_1 = 11$) showed a significant ($F(1, 10) = 6.18$, $p = 0.03$) main effect of breathing phase (inhale or exhale) on memory performance.
2. Nasal breathers showed enhanced memory for pictures that had been retrieved while inhaling compared to pictures that had been retrieved while exhaling ($t(10) = 2.85$, $p = 0.017$).
3. Oral breathers ($n_2 = 11$) did not show enhanced memory for pictures that had been retrieved while inhaling compared to pictures that had been retrieved while exhaling ($t(10) = -1.07$, $p = 0.31$).
4. There was no significant difference between nasal and oral breathers overall ($F(1, 20) = 1.15$, $p = 0.29$).
5. There was a significant interaction of breathing phase (inhale and exhale) with breath route (nasal and oral) when pictures were labeled by how they were encoded (inhale or exhale) ($F(1, 20) = 4.51$, $p = 0.046$).
6. There was also a significant interaction of breathing phase (inhale or exhale) with breath route (nasal and oral) when pictures were labeled by how they were retrieved (inhale or exhale) ($F(1, 20) = 7.06$, $p = 0.015$).

If you are confused, then take comfort in knowing that you are not alone. This study and its analysis is very complicated, which makes it difficult for a reader to connect the reported statistics to the theoretical conclusions. Moreover, some of the comparisons seem inappropriate. For example, the authors of the study used tests 2 and 3 to demonstrate a difference of significance for the nasal and oral breathers (comparing retrieval during inhaling versus exhaling). We noted in Chap. 3 (Implication 3b) that a difference of significance is not the same as a significant difference. Likewise, the authors of the study took the null result in test 4 as indicating "no difference" in performance of nasal and oral breathers overall. We saw in Chap. 3 (Implication 3a) that absence of proof is not proof of absence.

Table 7.1 Estimated success probabilities for the findings of a study relating memory performance to breathing orally or nasally

Test	Probability of success
Nasal: main effect of breath phase	0.690
Nasal retrieval: effect of breath phase	0.655
Oral retrieval: null effect of breath phase	0.809
Nasal vs. oral breathers: null main effect	0.820
During encoding: interaction for breath phase and route	0.604
During retrieval: interaction for breath phase and route	0.708
All tests	0.216

For the moment let us set aside our concerns about the appropriateness of the tests. Success for this study required four significant outcomes and two non-significant outcomes. If any of these outcomes were unsuccessful, it would call into doubt some of the conclusions made by the authors. As it turns out, the data supported every one of these necessary outcomes. We will show that with so many outcomes that must be satisfied by a single data set, such full success should be rare even if the effects are real and close to the values estimated by the experimental data. To estimate the probability of such a level of success, a statistical software program, R, was used to generate 100,000 simulated experiments with the reported sample sizes, means, standard deviations, and correlations (for within-subject aspects of the experiment). Table 7.1 shows how often each test produced the desired outcome. The success probability for any given hypothesis test varies between 0.60 and 0.82. For each significant test, the success probability of that specific test corresponds to power. For tests 3 and 4 listed above, a successful outcome was a non-significant result, and the table lists the probability of *not* rejecting the null hypothesis.

However, the probability of *every* test being successful for a given simulation is much lower than the probability for an individual test being successful because the data needs to have just the right properties to deliver a significant result for certain tests and to deliver a non-significant result for other tests. Based on the simulations, the joint probability that all of the tests would be successful in a single experiment is only 0.216. This low probability suggests that, simply due to random sampling, a direct replication of the study with similar sample sizes would have a rather small probability of producing the same pattern of outcomes.

A researcher replicating this study would want to pick sample sizes that give a high probability of success. Larger samples increase the power of a test, so that a study with just one test is more likely to find an effect if it exists. However, when the theoretical claims are based on both significant and non-significant tests, there are limits to the maximum probability of success because with large sample sizes small effects generate significant results (even for the studies where the authors hope for a null finding). The limit for this study can be investigated with additional simulated experiments that vary the

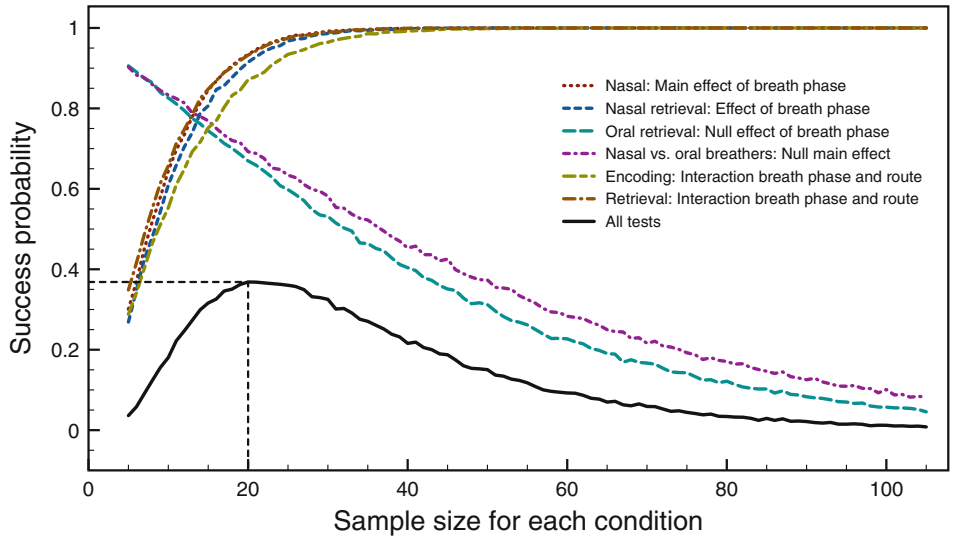


Fig. 7.4 Each colored line shows the estimated probability of success as a function of sample size for a test from a study investigating the effects of breathing on memory performance. The solid black curve shows the estimated success probability for all of the tests. The dashed black lines mark the sample size with the highest possible success probability for all of the tests combined. Each value is based on 10,000 simulated experiments

sample size for each condition. The colored lines in Fig. 7.4 plot the estimated probability of success for each of the six tests as a function of sample size (assuming the same sample size for each condition). For the four tests where success corresponds to producing a significant result, the probability of success increases with sample size and converges on the maximum value of 1 at around a sample size of 40. For the two tests where success corresponds to producing a non-significant result, the probability of success decreases with sample size (because some random samples show significant differences). The dashed black lines in Fig. 7.4 show that considering all six tests together (the black line), the maximum possible success probability is 0.37 with around $n_1 = n_2 = 20$ subjects in each condition.

This success probability analysis suggests that a better investigation of breathing and memory performance needs a different experimental design. Simpler designs are generally better because the more requirements you impose on a set of data (e.g., to produce many significant or non-significant outcomes) the lower the probability that any particular dataset will produce the required set of outcomes. Given the low estimated probability of success for this study, one might wonder how the original authors were so fortunate as to pick random samples that happened to reject/not reject results in exactly the pattern they needed to support their theoretical claims. We address this issue in Chap. 10 by considering how statistics should be interpreted across replications.

Take Home Messages

1. Keep your design simple: consider compressing raw data into intermediate variables, which then are subjected to statistical analysis.
2. Compute a power analysis before you do your experiment to check whether there is a real chance that it may show an existing effect.
3. Keep your design simple: if a theory presupposes both significant and null results your power may be strongly reduced.

Correlation

8

Contents

What You Will Learn in This Chapter

In Chaps. 3 and 6 we investigated the effect of latitude on tree height by measuring trees at 2 and 3 locations, respectively, and testing for differences in mean heights. As we will see, a better way to answer this question involves testing tree heights at even more locations of latitude. Computing an ANOVA is not a good idea for this situation because the ANOVA does not take the ratio scale properties of the latitude into account. The ANOVA treats each location as nominal (see Chap. 7). Correlations allow us to include the ratio scale aspect of the information and thereby summarize the effect of latitude into one value, r.

8.1 Covariance and Correlations

Let us first visualize correlations. If there were a perfect negative correlation, then an increase in one variable corresponds to a consistent decrease in another variable, for example, tree height decreases as latitude increases. If we plot latitude on the x-axis and tree height on the y-axis, the points fall on a straight line as in Fig. 8.1a (perfect negative

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_8

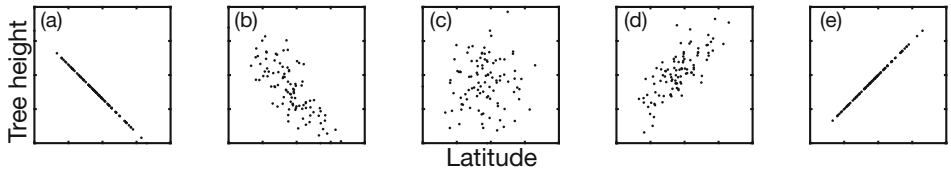


Fig. 8.1 Tree heights versus latitude for five different scenarios, (**a**)–(**e**). Each data point shows the height of one tree at one location. Correlations measure the linear relationship between the two variables

correlation). On the other hand, if there is no relationship between the variables, the data looks like a diffuse cloud of points as in Fig. 8.1c (no correlation). If tree height increases as latitude increases, there is a perfect positive correlation (Fig. 8.1e). Usually, we find cases in between the three basic scenarios (Fig. 8.1b, d).

This linear relationship is captured by the covariance equation:

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{X}) \times (y_i - \overline{Y})}{n - 1} \tag{8.1}$$

where, for example, the latitude data are x_i, the tree heights are y_i, and the $\overline{X}$ and $\overline{Y}$ are the respective mean values, i.e., the mean latitude and the mean tree height, respectively. The data consists of n pairs of latitudes and tree heights. The covariance generalizes the concept of variance because $\mathrm{cov}(x, x)$ is the variance of x.

A disadvantage of covariance is that it depends on the scale. For example, if you measure tree height in meters the covariance is smaller than if you measure it in centimeters. For this reason, we normalize the covariance by the standard deviation of x and y and arrive at the correlation:

$$r = \frac{\mathrm{cov}(x, y)}{s_x s_y} \tag{8.2}$$

This type of correlation is called Pearson's correlation. Correlation values range between -1.0 and $+1.0$, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and $+1$ indicates a perfect positive correlation (see Fig. 8.1a, c, and e). Values between these limits indicate intermediate strengths of the relationship between the variables.

8.2 Hypothesis Testing with Correlations

Figure 8.2 shows a sample ($n = 50$) of tree height data from many latitudes. Each point corresponds to a single tree. Obviously, there is not a perfect correlation, but the correlation seems to be different from zero. We use hypothesis testing to look for a significant

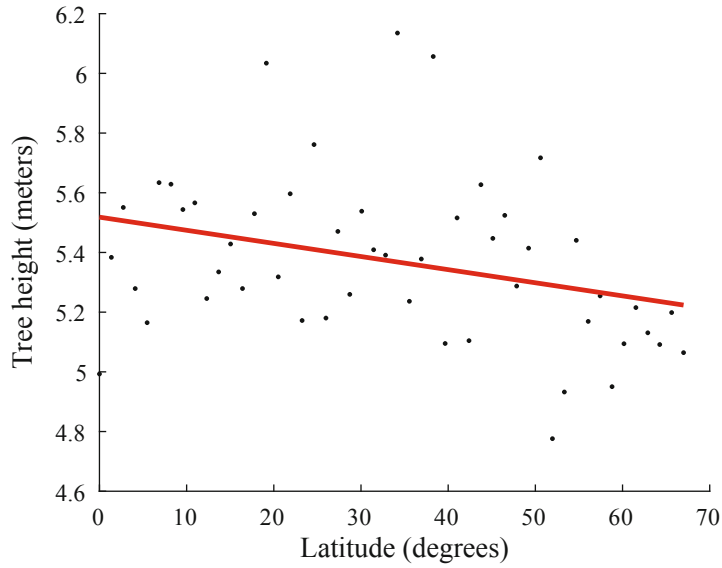


Fig. 8.2 Tree heights versus latitude for a sample of 50 trees. The correlation is $r = -0.312$. The red line is a best fitting straight line

correlation. Our null hypothesis is:

$$H_0 : \rho = 0$$

where ρ corresponds to the population correlation.

We do not need to go into the details, but if the null hypothesis is true, then the standard deviation of the sampling distribution of a sample correlation is:

$$s_r = \sqrt{\frac{1 - r^2}{n - 2}} \tag{8.3}$$

and the appropriate test statistic is a t value computed as:

$$t = \frac{r - 0}{s_r} \tag{8.4}$$

with degrees of freedom $df = n - 2$. The typical statistical software output for the data in Fig. 8.2 would look something like that shown in Table 8.1.

Table 8.1 Typical statistical software outputs for a correlation

r	t	df	p
−0.312	−2.28	48	0.027

Since the p value is less than 0.05, we conclude there is a significant correlation. The fact that the r-value is negative indicates that taller trees are found at lower latitudes

8.3 Interpreting Correlations

Assume we found a significant correlation between variables x and y, what does it tell us? First, it does not tell us that x causes y. This can be simply understood by noting that

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{X}) \times (y_i - \overline{Y})}{n-1} = \frac{\sum_{i=1}^{n}(y_i - \overline{Y}) \times (x_i - \overline{X})}{n-1} = \text{cov}(y, x) \tag{8.5}$$

which, if interpreted improperly, would suggest that x causes y and that y causes x. A significant correlation can occur for four reasons:

1. x causes y
2. y causes x
3. some intermediate variable z causes x and y
4. the correlation is spurious

An example for an intermediate variable (reason 3): it is not the latitude that determines tree heights. Rather factors related to latitude directly influence tree heights, such as water supply. Spurious correlations (reason 4) can occur by random. For example, for years 2000–2009 the correlation is $r = 0.947$ between US per capita consumption of cheese and the number of people who died by becoming tangled in their bedsheets. If scientists find such a high correlation in an experiment, they open a bottle of champagne! Spurious correlations are inevitable if you look across large enough sets of data.

It is important to note that because correlations only measure linear relationships, a non-significant correlation does not mean there is no relationship (or causation) between x and y. For example, air temperature systematically changes with time of day in a sinusoidal fashion (it goes up and down during the day-night cycle), but a correlation between time of day and temperature might produce $r \approx 0$.

It is always a good idea to look at a graph of data in addition to computing a correlation. Data of very different types can give rise to the same r-value (Fig. 8.3), so knowing

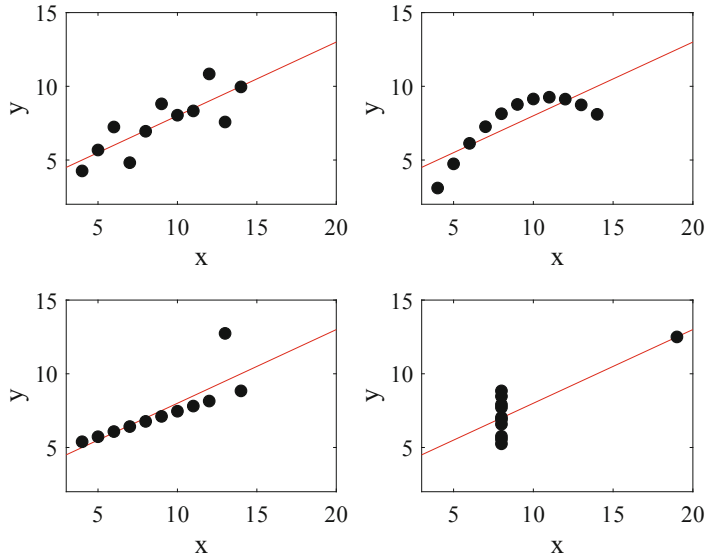


Fig. 8.3 Anscomb's quartet. Each data set has the same r-value ($r = 0.816$) despite looking very different when plotted

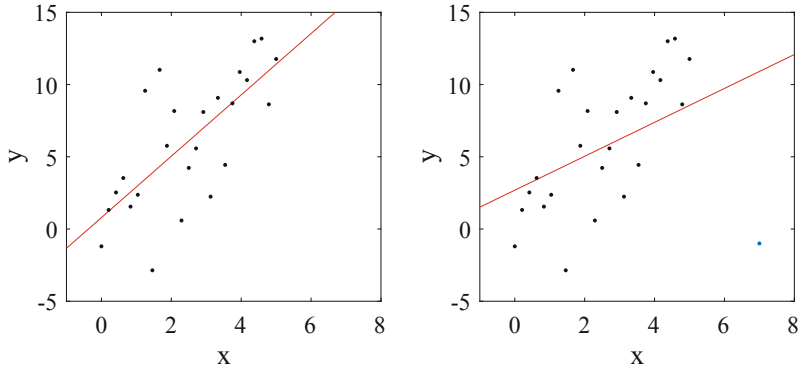


Fig. 8.4 Outliers can have a substantial impact on correlation. Left: original data set with $r = 0.71$. Right: a single outlier has been added (blue point at the bottom right) causing a big decrease in the correlation ($r = 0.44$)

only the correlation value provides only partial information about the data set. Moreover, correlations are very sensitive to outliers (Fig. 8.4), and a single data point added or removed from a data set can dramatically change the correlation value.

Table 8.2 Effect size guidelines for $|r|$ according to Cohen

	Small	Medium	Large
Effect size	0.1	0.3	0.5

8.4 Effect Sizes

Correlation is often used as a measure of effect size that indicates how much one variable is related to another variable. In particular, the square of a correlation, r^2, indicates the proportion of variability in one score (e.g., tree height) that can be explained by variability in the other score (e.g., latitude). This is the same kind of information provided by η^2, which we covered in Chap. 6. According to Cohen, an r-value of less than 0.1 is considered a small effect and the very same is true for values lower than -0.1 (Table 8.2).

8.5 Comparison to Model Fitting, ANOVA and t-Test

In Chap. 7 we fit a linear model to the learning data and focused on the slope, which is similar to computing a correlation because the correlation is a measure of linear relationships. A hypothesis test for a non-zero slope gives the same result as a hypothesis test for a non-zero correlation.

As mentioned in Chap. 7, it is not a good idea to use an ANOVA when the independent variable is on a ratio scale because the ANOVA treats the independent variable as being on a nominal scale. By taking full advantage of the ratio scale an analysis based on correlation has higher power than an ANOVA.

One could also use the t-test by splitting the data into, for example, smaller and larger than median latitudes, i.e., half the data go into a North group, the other half into a South group. In general, such approaches are not as good as an analysis based on the correlation because they (again) do not include the ratio scale nature of the independent variable. For example, in Fig. 8.5 the data from Fig. 8.2 are split into lower and higher latitude regions. The t-test does not produce a significant result. Thus, if we analyze the data with these subsets, we fail to note the significant difference found by looking at the correlation in the original data set (Table 8.1).

In some way, a correlation may be seen as a generalization of the ANOVA and the t-test.

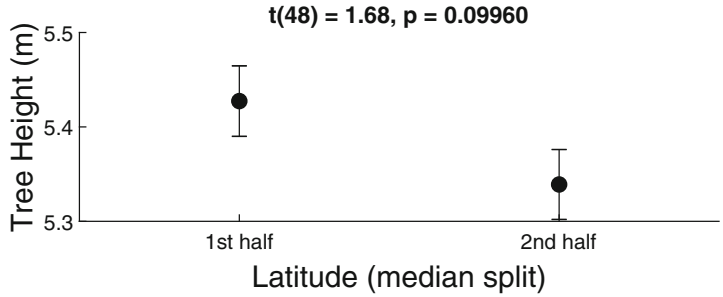


Fig. 8.5 Data produced from a median split of the data in Fig. 8.2. A t-test investigating differences between the means is not significantly different

8.6 Assumptions and Caveats

Hypothesis tests for correlations hold several assumptions.

1. As always, data need to be independent and identically distributed.
2. The y-variable is Gaussian distributed when conditioned on any given x-value. That is, if we were to take all the y-values at a single x-value and make a histogram of them, the histogram would be Gaussian distributed.
3. Both variables are interval or ratio scaled.
4. Sample size is fixed before the experiment.

If data are on an ordinal scale, correlations can be computed with the Spearman's ρ, which uses ranks (ordinal scale) rather than the ratio scale. Spearman correlations are the non-parametric equivalent of the parametric Pearson correlations.

8.7 Regression

In this subsection, we quickly sketch the relationship between correlations and regressions. The hasty reader may skip it. Regression will play no role in the following chapters.

A correlation tells us about how tightly packed the data are around the best fitting line. For example a correlation of 1.0 tells us that all data points are perfectly on the line. However, what is this best fitting line? Regression gives us the equation of that best fitting line, which has one parameter for the slope (m) and one for the y-intercept (b; i.e., where the line hits the y-axis). The slope of the regression line is the standard deviation in the y-direction divided by the standard deviation in the x-direction, weighted by the r value from Eq. 8.2:

$$m = r\frac{s_y}{s_x} \tag{8.6}$$

Table 8.3 Typical statistical software outputs for a regression

Parameter	Coefficient value	t	p
Intercept (constant)	12.146	4.079	0.00017
Slope (latitude)	−0.147	−2.275	0.027

This means for every standard deviation we walk in the x-direction, we step up by the standard deviation in the y-direction multiplied by the r-value.

The intercept b is:

$$b = \bar{y} - m\bar{x}$$

For the tree height data presented in Fig. 8.2, the slope is $m = -0.1473$ and the intercept is $b = 12.1461$. This means that at a latitude of zero degrees, the average tree height is 12.1461 m, and that for every degree of latitude that we go North of that, we increase in tree height by −0.1473 m (in other words, tree heights go down as we increase our latitude). These results are typically summarized in statistical software as shown in Table 8.3

Here, in addition to the regression line slope and intercept, the statistical software also outputs a t- and p-value for the slope and intercept, the so-called regression coefficients. These statistics test the null hypothesis that the slope and intercept are equal to zero. In this example, the p-values are smaller than 0.05, and so both are significantly different from zero. In such a situation, the corresponding correlation (r-value) is typically significantly different from zero. An intercept that is not significantly different from zero means that the regression line roughly crosses the point (0, 0) on the graph.

Take Home Messages

1. Correlations are the preferred choice if both the x- and y-axis are ratio or interval scaled.
2. Causation and correlation should never be confused.
3. Very different sets of data can lead to the same r.

Part III

Meta-analysis and the Science Crisis

Meta-analysis

9

Contents

What You Will Learn in This Chapter

In Part III of this book, we will show that combining data from multiple experiments can provide completely new insights. For example, whereas the statistical output of each experiment itself might make perfect sense, sometimes the combination of data across experiments indicates problems. How likely is it that four experiments with a small effect and a small sample size all lead to significant results? We will show it is often very unlikely. As a simple consequence, if experiments always produce significant results, the data seem to good to be true. We will show how common, but misguided, scientific practice leads to too-good-to-be-true data, how this practice inflates the Type I error rate, and has led to a serious science crisis affecting most fields where statistics plays a key role. In this respect, Part III generalizes the Implications from Chap. 3. At the end, we will discuss potential solutions.

In this chapter, we extend the standardized effects size from Chap. 2 and show how to combine data across experiments to compute meta-statistics.

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_9

9.1 Standardized Effect Sizes

As we noted in Part I of this book, much of statistics involves discriminating signal and noise from noise alone. For a standard two sample t-test, the signal to noise ratio is called Cohen's d, which is estimated from data as (see Chap. 3):

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s}.$$

Cohen's d tells you how easily you can discriminate different means. The mean difference is in the numerator. A bigger difference is easier to detect than a smaller one, but we also need to take the noise into account. A bigger standard deviation makes it more difficult to detect a difference of means (see Chap. 2). When $n_1 = n_2 = n$, the t-value for a two-sample t-test is just:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_{\overline{x}_1 - \overline{x}_2}} = \frac{\overline{x}_1 - \overline{x}_2}{s\sqrt{\frac{2}{n}}} = \frac{d}{\sqrt{\frac{2}{n}}} = d\sqrt{\frac{n}{2}}$$

So, a t-value simply weights Cohen's d by (a function of) sample size(s). As mentioned in Chap. 3, it is always good to check out the effect size. Unfortunately, many studies report just the p-value, which confuses effect size and sample size. Based on the above equation, we can compute Cohen's d from the reported t-value and the sample sizes:

$$d = t\sqrt{\frac{2}{n}}$$

An important property of Cohen's d is that its magnitude is independent of the sample size, which is evident from d being an estimate of a fixed (unknown) population value.[1]

In Chap. 3, we have shown that we can estimate δ by d. However, d is only a good estimator when the sample size is large. For rather small samples, d tends to systematically overestimate the population effect size δ. This overestimation can be corrected by using Hedges' g instead of d:

$$g = \left(1 - \frac{3}{4\,(2n-2)-1}\right) d$$

For nearly all practical purposes Hedges' g can be considered to be the same as Cohen's d. We introduced it here because we will use Hedges' g to compute meta-analyses. The

[1] Note that although the sample size n appears in this particular formula, it basically just compensates for t increasing with larger sample size.

Appendix to this chapter includes formulas for when $n_1 \neq n_2$ and for other types of experimental designs.

9.2 Meta-analysis

Suppose we run the same (or very similar) experiments multiple times. It seems that we should be able to pool together the data across experiments to draw even stronger conclusions and reach a higher power. Indeed, such pooling is known as meta-analysis. It turns out that the standardized effect sizes are quite useful for such meta-analyses.

Table 9.1 summarizes statistical values of five studies that concluded that handling money reduces distress over social exclusion. Each study used a two-sample t-test, and the column labeled g provides the value of Hedges' g, which is just an estimate of the effect size.

To pool the effect sizes across studies, it is necessary to take the sample sizes into account. An experiment with 46 subjects in each group counts a bit more than an experiment with 36 subjects in each group. The final column in Table 9.1 shows the weighted effect size, $w \times g$, for each experiment (see the Appendix for the calculation of w). The pooled effect size is computed by summing the weighted effect sizes and dividing by the sum of the weights:

$$g^* = \frac{\sum_{i=1}^{5} w_i g_i}{\sum_{i=1}^{5} w_i} = 0.632.$$

This meta-analytic effect size is the best estimate of the effect size based on these five experiments. Whether it is appropriate to pool standardized effect sizes in this way largely depends on theoretical interpretations of the effects. If your theoretical perspective suggests that these experiments all measure essentially the same effect, then this kind of pooling is appropriate, and you get a better estimated effect size by doing such pooling. On the other hand, it would not make much sense to pool together radically different experiments that measured different effects.

Meta-analyses can become quite complicated when experiments vary in structure (e.g., published analyses may involve t-tests, ANOVAs, or correlations). Despite these

Table 9.1 Data from five experiments used for a meta-analysis

n	t	g	$w \times g$
36	3.01	0.702	12.15
36	2.08	0.485	8.66
36	2.54	0.592	10.43
46	3.08	0.637	14.17
46	3.49	0.722	15.83

difficulties, meta-analysis can be a convenient way to combine data across experiments and thereby get better estimates of effects.

Take Home Messages

1. Pooling effect sizes across experiments produces better estimates.
2. Combining data across experiments increases power.

Appendix

Standardized Effect Sizes Beyond the Simple Case

When samples sizes are different ($n_1 \neq n_2$), the t-value of a two-sample t-test is:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_{\overline{x}_1 - \overline{x}_2}} = \frac{\overline{x}_1 - \overline{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{d}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = d\sqrt{\frac{n_1 n_2}{n_1 + n_2}}.$$

If a published study does not report the means and standard deviations from the samples, it is possible to compute Cohen's d from the reported t-value and sample sizes:

$$d = t\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

For Hedges' g, the calculation with unequal sample sizes is:

$$g = \left(1 - \frac{3}{4\,(n_1 + n_2 - 2) - 1}\right) d$$

There are similar standardized effect sizes and corrections for other experimental designs. For example, for a one-sample t-test with a null hypothesis of the population mean being equal to the value a, Cohen's d is calculated as

$$d = \frac{\overline{x} - a}{s}$$

which, again, represents signal in the numerator (deviation from the value specified by the null hypothesis) and noise in the denominator (the sample standard deviation). An unbiased version of Cohen's d for the one-sample case is Hedges' g:

$$g = \left(1 - \frac{3}{4\,(n - 1) - 1}\right) d$$

For repeated measures t-tests, the appropriate standardized effect size depends on how it will be used. Sometimes, a scientist wants an effect size relative to the difference scores that are calculated for each subject. For that use, the one-sample d or g is appropriate. Other times, scientists want to find an effect size that is equivalent to what it would be for a two-sample independent t-test. In that situation it is necessary to compensate for the correlation between scores. When computed from the reported t value of a dependent sample, the formula is:

$$d = \frac{t}{\sqrt{n}}\sqrt{2\,(1-r)}$$

Unfortunately, most papers do not report the correlation between scores for a dependent sample. For our purposes, the basic idea of a standardized effect size is more important than the specific calculation. However, you should be aware that formulas you may find on the Internet sometimes include unstated assumptions such as equal sample sizes for an independent t-test or $r = 0.5$ for a dependent t-test.

Extended Example of the Meta-analysis

Table 9.2 fills in some intermediate terms that are not present in Table 9.1.

To pool the effect size across studies, we weight each g value by its inverse variance. The calculation of the inverse variance involves multiple steps. For an independent two-sample t-test, the formula for the variance of Cohen's d is

$$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2\,(n_1 + n_2)}$$

and the variance for Hedges' g includes the square of the correction term used earlier:

$$v_g = \left(1 - \frac{3}{4\,(n_1 + n_2 - 2) - 1}\right)^2 v_d$$

Table 9.2 Detailed meta-analysis, including additional computations, of the data shown in Table 9.1

n_1	n_2	t	g	v_g	w	wg
36	36	3.01	0.702	0.058	17.3	12.15
36	36	2.08	0.485	0.056	17.9	8.66
36	36	2.54	0.592	0.057	17.6	10.43
46	46	3.08	0.637	0.045	22.2	14.17
46	46	3.49	0.722	0.046	21.9	15.83

which is shown in a separate column in Table 9.2. To do the meta-analysis, each standardized effect size is multiplied by its inverse variance:

$$w = \frac{1}{v_g}$$

which is shown in a column in Table 9.2 next to a column listing the product of wg for each experiment. The pooled effect size is computed by summing the products and dividing by the sum of the weights:

$$g^* = \frac{\sum_{i=1}^{5} w_i g_i}{\sum_{i=1}^{5} w_i} = 0.632.$$

Understanding Replication

10

Contents

What You Will Learn in This Chapter

This chapter uses the power analyses from Chap. 7 and the meta-analytic methods from Chap. 9 to identify improper statistical analyses in published findings. The basic idea is simple. Unless the power is very high, we know that even real effects will not always produce significant outcomes simply due to random sampling. If power is only moderate but all studies are significant, the reported results seem too good to be true. Our considerations have a crucial implication: replication cannot be the final arbiter for science when hypothesis testing is used, unless experimental power is very high. Chapter 11 shows how such results can be produced even when scientists are trying to do everything properly.

10.1 The Replication Crisis

Across all sciences, replication is considered to be the "gold standard" for demonstrating important findings. Should a colleague happen to doubt the veracity of your empirical claim a surefire way to shut him down is to demonstrate that the effect can be consistently reproduced. The demonstration is especially effective if an independent lab replicates the effect. Along similar lines, if an independent lab reports that an effect cannot be replicated,

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_10

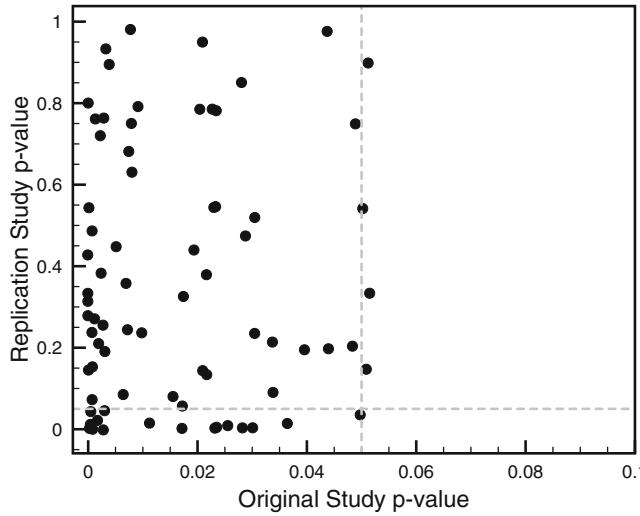


Fig. 10.1 Each point corresponds to a pair of p-values for an original study and its replication. While almost all original studies produced $p < 0.05$, very few replication studies produced such a small p-value. The figure is reproduced from Open Science Collaboration [1]. Please note the highly different scales of the x- and y-axes. The x-axis shows values in the range from 0.0 to 0.1, while the y-axis goes from 0.0 to 1.0. There is no obvious relationship between the p-values of the original and reproduction studies. A good result would have found that the p-values of the replication studies were smaller than 0.05, i.e., all black dots should be below the dashed horizontal line

there tends to be vigorous discussion about whether the correct procedures were followed and what the results mean. Successful replication is highly valued, and is taken as strong support for a scientific claim.

Unfortunately, many fields do not seem to be doing well with regard to replication. A group of psychologists, called the Open Science Collaboration [1], conducted replications of 97 studies that were published in three top journals. In their 2015 report, only 36% of the replication studies produced results consistent with the original studies. Each point in Fig. 10.1 plots the reported p-value for a replication study against the p-value for the original study. The dashed vertical line indicates the 0.05 criterion for the original studies, and almost all original studies reported a p-value below this criterion. This is not surprising because usually only significant results are published. The dashed horizontal line indicates the 0.05 for the replication studies, and almost all studies reported p-values above this criterion. The shocking result is that there hardly seems to be any relationship between the original study p-value and the replication study p-value. For example, some original studies reported p-values much smaller than 0.01 but the replication results yield p-values close to 1.0. Even worse, many of the replication studies had *larger* sample sizes than the original studies, and so should have produced smaller p-values, as we emphasized in Chap. 3 (Implication 2d).

Replication problems are not limited to psychology. In 2012, researchers at the biotech firm Amgen reported that they were unable to reproduce findings in 47 out of 53 landmark papers involving cancer research. There is an on-going effort by academic researchers to run replication studies similar to what was done by psychologists. The early results of that effort do not seem better than the replication results in psychology. For many people the lack of replication success in these studies indicates extremely serious problems that are sometimes referred to as a "replication crisis."

We agree that the problems are serious. However, we propose that rather than looking at new replication studies and wonder why they do not succeed, it is easier to look at the original published results and show that they never made sense.

Consider the following two phenomena that have been studied with multiple experiments.

- Phenomenon A: Nine of the ten experiments produced significant results, so it has a replication success rate of 0.9.
- Phenomenon B: Ten of the nineteen experiments produced significant results, so it has a replication success rate of 0.53.

If you follow the view that successful replication is a good guide for veracity, then the experimental outcomes definitely favor phenomenon A over phenomenon B. Neither phenomenon shows perfect replication, but we know from Chaps. 3 and 7 that not every experiment should work. Even so, phenomenon B only replicates about half the time, so we might even wonder whether the effect is real.

The problem with this interpretation is that phenomena A and B correspond to real investigations. Phenomenon A refers to what is known as precognition: the ability of people to get information from the future and use it in the present. A paper published in a top journal in 2011 reported that nine out of ten studies produced significant evidence for precognition. Despite the reported findings, very few scientists believe that precognition is a real effect; largely because its existence would undermine the very successful theory of general relativity. Thus, we are left to conclude that a high replication rate is not always sufficient to cause people to believe in the veracity of the effect.

Likewise, phenomenon B refers to what is known as the bystander effect: a tendency for people to not provide help to someone if there are other people around who could also provide help. Experiments on the bystander effect are rather difficult to run because one needs to have collaborators who pose as people needing help and other collaborators who pose as people who are around but not providing help. For this reason, these studies tend to use relatively small sample sizes. As a result, it is not uncommon for a study on the bystander effect to not produce a significant result. Even so, pretty much everyone believes that the bystander effect is a real phenomenon. Thus, we are left to conclude that a high replication rate is not always necessary to cause people to believe in the veracity of the effect.

We seem to be left with an odd situation. Scientists cite replication as a gold standard for judging the veracity of effects, but when faced with *actual* sets of experiments, replication seems neither sufficient nor necessary to establish veracity. It makes one wonder why scientists bother running experiments at all!

The way out of this odd situation requires a better understanding of statistics and replication. In the next subsection, we show that experiments should not always replicate, in particular, when effect and sample sizes are small. Replication success should reflect the estimated success probabilities of experiments. We should worry when experiments replicate too often.

10.2 Test for Excess Success (TES)

A set of experiments should succeed at a rate that follows the probability of success. Let us see whether that holds true for the precognition studies.

In the precognition study, each experiment was analyzed with a one-tailed, one-sample t-test. Table 10.1 lists the sample size for each experiment and the standardized effect size (Hedge's g). We use the meta-analytic techniques described in Chap. 9 to compute a pooled estimate of the standardized effect size. Doing so gives $g^* = 0.1855$. Doing the meta-analysis here is appropriate because the author of the studies used a similar analysis to provide partial support for his theoretical claim that precognition exists. Our pooled effect size, g^*, is our best estimate of the effect size, and we can use it to estimate the power of each individual experiment as described in Chap. 7. The last column in Table 10.1 shows the estimated power based on this meta-analytic effect size. Consistent with the observations about power made in Chap. 7, power values rise and fall with sample size. Experiment 9 ($n = 50$) is expected to have the smallest power (0.36) and Experiment 7 ($n = 200$) is expected to have the highest power (0.83). About half of the experiments (those with $n \approx 100$) have power values a bit above one half.

Table 10.1 Statistics for ten experiments that purported to find evidence for precognition

	Sample size (n)	Effect size (g)	Power
Exp. 1	100	0.249	0.578
Exp. 2	150	0.194	0.731
Exp. 3	97	0.248	0.567
Exp. 4	99	0.202	0.575
Exp. 5	100	0.221	0.578
Exp. 6a	150	0.146	0.731
Exp. 6b	150	0.144	0.731
Exp. 7	200	0.092	0.834
Exp. 8	100	0.191	0.578
Exp. 9	50	0.412	0.363

Suppose a scientist decided to replicate this set of ten experiments with the very same sample sizes as in the original report. If we accept the pooled effect size as a good measure of the precognition effect, then the expected number of successful outcomes in a set of ten experiments like these is the sum of the power values across the ten experiments. For example, if the power of each experiment is 1.0, then the number of significant results must be 10, the sum of the power values. For the studies in Table 10.1 the sum of the power values is 6.27. Hence, the experiments should have replicated 6.27 times. That expected degree of success is quite a bit lower than the 9 out of 10 success reported in the original investigation.

How likely is it to get 9 or 10 significant results for this effect? Doing something like a hypothesis test, we can estimate the probability of getting 9 or more successful outcomes from 10 experiments like these. We do not have to get exactly the 9 successes reported in the original report, any 9 out of 10 experiments will do. We compute the success probability by identifying all 11 combinations of experiments that demonstrate 9 or 10 successful outcomes. For each combination, we compute the probability of that particular result by multiplying the power of each successful experiment and the complement of power for each unsuccessful experiment. We then add up all those probabilities to get 0.058. That is, if the effect is real and similar to what was reported, a scientist doing a precise replication of the original ten experiments has only around a 6% chance of having the same degree of success as claimed in the original report. If replication success is supposed to guide our belief in the veracity of experimental results, this low rate seems like a serious problem.

Moreover, the low estimated replication rate begs the question of how the original author was able to produce such a high success rate. Given what we now know (from those studies) about the effect of precognition, it is very strange that those ten experiments were so successful. It is so strange that we can suspect that something went wrong in this set of experiments. We may never know exactly what happened in this set of experiments (even the original researcher might not know), but the burden of proof is on the researcher presenting the results. Perhaps there is a true precognition effect, but these studies do not provide good scientific evidence for it.

What if we apply the same kind of analysis to phenomenon B, where ten out of nineteen studies found statistically significant results for the bystander effect? Following the same basic approach, the pooled standardized effect size is -0.47, where the negative number indicates the presence of the bystander effect. That pooled effect size can be used to estimate the power for each of the nineteen experiments. The power varies from 0.2 to nearly 1.0 because several experiments had as few as 24 participants and one experiment had 2500 participants. Across all nineteen experiments, the sum of the power values is 10.77. Thus, we would expect to see around 11 significant results for experiments like these; and the nineteen experiments actually produced 10 significant results. Thus, the set of experimental results investigating the bystander effect seems believable, because the rate of success matches the estimated magnitudes of the effect and sample sizes of

the experiments. The estimated probability of observing 10 or more significant results for studies like these is calculated to be 0.76.

10.3 Excess Success from Publication Bias

The previous subsection described the Test for Excess Success (TES), which examines whether the reported success rate of a set of experiments agrees with the estimated magnitude of the effect and the sample sizes of the experiments. If there is a big mismatch, then the TES suggests that there is a problem with the set of experiments, a problem with the analyses, or a problem with the theoretical claims based on the data/analyses. This subsection and the next use simulated experiments to show how it might happen that there is too much replication. This subsection considers the impact of publication bias: selective publishing of significant findings and suppression of non-significant findings.

Table 10.2 summarizes statistics from 20 simulated experiments that were each analyzed with a two-sample t-test. Each experiment had a simulated control group, for which there was no effect. For this group, scores were drawn from a normal distribution with a mean of zero and a standard deviation of one. For a simulated experimental group, scores were drawn from a normal distribution with a mean of 0.3 and a standard deviation of one. Hence, the population standardized effect size is $\delta = 0.3$. Sample sizes were the same for the two groups, $n_1 = n_2$. The sample sizes were drawn at random from a uniform distribution between 15 and 50.

The second column in Table 10.2 shows the t-value for each simulated experiment. The bolded t-values indicate statistical significance, as the p-values are less than the 0.05 criterion. There are five significant experiments. How does the success rate of five out twenty do when investigated with the TES? We can treat the simulated data in a way similar to the studies on precognition and the bystander effect. When we pool the effect sizes across all twenty experiments, we get $g^* = 0.303$. This estimated value is very close to the true value of 0.3, which simply demonstrates that meta-analysis works if all experiments are included in the analysis. We can use the pooled effect size to estimate power for each experiment, with the results reported in column 4 of Table 10.2. Summing these power values gives 4.2, and the probability of such experiments producing five or more significant outcomes is 0.42. There is no commonly agreed criterion for an appropriate success probability, but many people get concerned if the probability is less than 0.1. When both significant and non-significant experiments contribute to the analysis, the success rate tends to be consistent with the estimated power values. So far, so good.

Now suppose that a researcher practices a form of publication bias so that only the significant experiments (bolded t-values in Table 10.2) are published and available for further investigation. If we pool only the effect sizes for the five published experiments, we get $g^* = 0.607$, which is double the population effect size. This makes sense because those significant experiments must have a relatively large t-value. Since the effect size is a function of the t-value, these experiments must also have an unusually large estimated

Table 10.2 Statistics from twenty simulated experiments to investigate the effects of publication bias

$n_1 = n_2$	t	Effect size	Power from pooled ES	Power from biased ES
29	0.888	0.230	0.206	
25	1.380	0.384	0.183	
26	1.240	0.339	0.189	
15	0.887	0.315	0.126	
42	0.716	0.155	0.279	
37	1.960	0.451	0.251	
49	−0.447	−0.090	0.318	
17	1.853	0.621	0.138	
36	**2.036**	0.475	0.245	0.718
22	1.775	0.526	0.166	
39	1.263	0.283	0.262	
19	**3.048**	0.968	0.149	0.444
18	**2.065**	0.673	0.143	0.424
26	−1.553	−0.424	0.189	
38	−0.177	−0.040	0.257	
42	**2.803**	0.606	0.279	0.784
21	1.923	0.582	0.160	
40	**2.415**	0.535	0.268	0.764
22	1.786	0.529	0.166	
35	−0.421	−0.100	0.240	

Bolded t values indicate statistical significance ($p < 0.05$)

effect size. Hence, one impact of a publication bias is that the published studies can dramatically overestimate the magnitude of effects. Using the overestimated effect size to compute power for each experiment produces the values in the last column of Table 10.2. These values are dramatically larger than the true power values because they are based on a gross overestimate of the effect size. Nevertheless, the power values sum to 3.13, which indicates that out of five published experiments like these we would expect around three significant results. In reality, all five experiments produced significant results, and the probability that all five experiments would produce a significant result is the product of the power values, which is 0.081. For many people this is such a low probability (e.g., less than 0.1) that they would doubt the validity of the published results.

10.4 Excess Success from Optional Stopping

As mentioned in Chap. 4, a requirement for the t-test is that the sample sizes for the two groups are fixed before the experiment. In practice, however, it is very common for a sample to *not* have a fixed size. Consider the following situation. A scientist gathers data

from two populations and ends up with $n_1 = n_2 = 10$ scores in each sample. The scientist runs a t-test and computes $p = 0.08$. This p-value does not fall below the 0.05 criterion that is used for statistical significance, but it looks promising. Oftentimes researchers in this situation decide to gather ten more scores, so that they now have $n_1 = n_2 = 20$ scores in each sample. Suppose that when the t-test is run on this larger sample it produces $p = 0.04$, which indicates statistical significance. This sounds good: more data gives a better answer. Unfortunately, this kind of procedure can dramatically inflate the Type I error rate. One problem is that this procedure involves multiple tests. Each test has some probability of producing a Type I error. As shown in Chap. 5, with multiple tests the probability of at least one of them making a Type I error is higher than the probability of a single test producing a Type I error.

The more serious problem with this procedure is that data collection is stopped once a desired result has been found. As additional observations are added to the original data set, a conclusion of significance may switch to non-significance, and vice-versa. If the decision to add data is tied to finding a significant result (e.g., no more data is collected once $p < 0.05$), then the data collection process is biased toward producing significant outcomes. This kind of procedure is called "optional stopping," and it increases the Type I error rate. An unscrupulous scientist who started with $n_1 = n_2 = 10$ and added one observation to each data set until getting a significant outcome ($p < 0.05$) or a maximum of $n_1 = n_2 = 50$ would have a Type I error rate over 20%.

It is important to recognize that the problem here is not with *adding* data but with *stopping* data collection because the Type I error rate refers to the *full* procedure. Thus, optional stopping is a problem *even* if the first data set happens to produce a significant result, but the scientist *would have* added more subjects to a non-significant data set. Importantly, if a researcher does not have a specific plan for data collection, then it is impossible to compute the Type I error rate. This is why the standard approach to hypothesis testing assumes a fixed sample size.

The TES is sensitive to a set of studies where researchers followed this kind of improper approach, and it is fruitful to look at simulated experiments to get some intuition on what happens. Table 10.3 summarizes statistics from 20 simulated experiments that were analyzed with a two-sample t-test. For both the control and experimental groups, the sample sizes n_1 and n_2 were the same. Scores were drawn from a normal distribution with a mean of zero and a standard deviation of one. Hence, the population effect size is $\delta = 0$; there is truly no effect here.

To simulate optional stopping, each sample started with $n_1 = n_2 = 15$ scores. A t-test was run on that data and if a significant result was found, the experiment was stopped and reported. If the t-test did not find a significant result, one more data point was sampled for each group and the t-test was repeated. This process continued up to a sample size of $n_1 = n_2 = 100$, where the result was reported.

Since the population effect equals zero, we would expect to get, on average, one significant outcome from twenty simulated experiments (see Chap. 5). The four bolded t-values in Table 10.3 indicate statistical significance, which is a much higher rate (20%)

Table 10.3 Statistics from twenty simulated experiments to investigate the effects of optional stopping

$n_1 = n_2$	t	Effect size	Power from pooled ES	Power from file drawer ES
19	**2.393**	0.760	0.053	0.227
100	0.774	0.109	0.066	
100	1.008	0.142	0.066	
63	**2.088**	0.370	0.060	0.611
100	0.587	0.083	0.066	
100	−1.381	−0.195	0.066	
100	−0.481	−0.068	0.066	
100	0.359	0.051	0.066	
100	−1.777	−0.250	0.066	
100	−0.563	−0.079	0.066	
100	1.013	0.143	0.066	
100	−0.012	−0.002	0.066	
46	**2.084**	0.431	0.057	0.480
100	0.973	0.137	0.066	
100	−0.954	−0.134	0.066	
100	−0.136	−0.019	0.066	
78	**2.052**	0.327	0.062	0.704
100	−0.289	−0.041	0.066	
100	1.579	0.222	0.066	
100	0.194	0.027	0.066	

Bolded t values indicate statistical significance ($p < 0.05$)

than the intended 5%. A simple computation, using the binomial distribution, shows that the probability of getting four or more significant experiments in a set of twenty is 0.016 when each experiment has a 5% chance of producing a significant result. All of the non-significant experiments in Table 10.3 have sample sizes of 100 (the maximum possible sample size) because that is the nature of the optional stopping procedure.

Computing the pooled effect size across all twenty experiments finds $g^* = 0.052$, which is very close to the population effect size of zero. Contrary to the effect of publication bias, optional stopping does not bias estimates of the effect size. Likewise, if we use that estimated effect size to calculate power for each experiment, we get values ranging from 0.053 to 0.066, which are all just above the 0.05 significance criterion because the estimated effect size is just larger than zero. Still, the reported results seem too good to be true. Adding up the power values for all twenty experiments gives just 1.28, so we would expect to find around one significant experiment among twenty experiments like these. The probability of experiments like these producing four or more significant outcomes is calculated from the power values as 0.036. This result (correctly) indicates some kind of problem in the set of experiments: the rate of success is larger than it should be.

The last column of Table 10.3 shows power values based on only the significant experiments in Table 10.3. Here, we suppose that the non-significant experiments were not published (publication bias). In that situation the TES analysis has to work with only the four reported significant experiments. The pooled effect size estimate is $g^* = 0.4$, which is dramatically larger than the true value of zero. As a result of this overestimate of the effect size, the power values for the four significant experiments are also dramatically overestimated. Nevertheless, adding up those four power values indicates that four experiments like these would be expected to produce around two significant outcomes. The probability of all four experiments producing significant outcomes is the product of the power values, which is 0.047. Again, this set of studies (correctly) seems problematic because the success rate is out of line with the estimated effect and the experiment sample sizes.

10.5 Excess Success and Theoretical Claims

The Test for Excess Success is able to identify situations where the reported rate of success does not match the experimental effect and sample sizes. An important point of this analysis is the definition of "success," which is always relative to some theoretical claim. As an example, suppose that a researcher runs ten independent experiments that each investigates a different topic (e.g., the Stroop effect, a memory experiment, differences in EEG alpha synchrony, epigenetic transfer of learned behavior, precognition, and other topics). Suppose that the first four experiments find a significant outcome but the other six experiments do not. Further suppose that the researcher imposes a publication bias and only publishes the four successful experimental results and does not publish the six null results found for the other studies. A TES analysis on the four published studies may (correctly) indicate evidence of publication bias, but this observation is fairly meaningless. The four experiments are unrelated to each other and are unrelated to any overarching theoretical claim. As such, all we can conclude is that there were other unsuccessful experiments that have not been reported, but the existence of such unsuccessful experiments tells us nothing about the veracity of the reported properties of the Stroop effect or performance in the memory experiment.

On the other hand, if the same researcher used the results of the very same four significant experiments to make some theoretical claim (e.g., a unified theory of the Stroop effect, memory, EEG alpha synchrony, and epigenetic transfer), then publication bias potentially undermines that theoretical claim. If a TES analysis indicates that the set of four studies suggests publication bias, then scientists should be skeptical about the corresponding theoretical claims that have been derived by the researcher.

Oftentimes researchers unintentionally make their theoretical conclusions seem too good to be true by having their theory be determined by the significance/non-significance of their tests. In this case the theory becomes nothing more than a coarse summary of what was measured in the experiment. Such a theory is almost certain to be chasing (some)

noise in the experimental results and is almost surely not going to be fully supported by a new set of experiments.

Consistent with Chap. 3, Implication 3a, the conclusion of the TES analysis does not prove that there is no effect across a set of experiments; rather it indicates that the set of experiments does not make a convincing scientific argument.

Take Home Messages

1. If many similar experiments with low effect and sample size all lead to significant results: the data seem too good to be true.
2. Experiments should lead to significant results proportional to their power.
3. Publication bias and optional stopping can lead to strongly inflated Type I error rates.

Reference

1. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349. https://doi.org/10.1126/science.aac4716.

Magnitude of Excess Success

11

Contents

What You Will Learn in This Chapter

Chapter 10 introduced the Test for Excess Success (TES), which detects some forms of bias in statistical analyses across multiple experiments. Although we found study sets where the TES indicated problems, it could be that the vast majority of scientific investigations are fine. This chapter shows, unfortunately, that this is not the case.

11.1 You Probably Have Trouble Detecting Bias

The basic ideas of the TES are fairly simple. An important point of statistics is that failures are necessary. Even if an effect is real, sometimes a researcher should select a random sample that does not show the effect. Only reporting successful outcomes is problematic because it inflates the reported effect size and can indicate the existence of an effect when

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_11

Table 11.1 Summary statistics for three sets of five simulated experiments

Set A				Set B				Set C			
$n_1 = n_2$	t	p	g	$n_1 = n_2$	t	p	g	$n_1 = n_2$	t	p	g
10	2.48	0.03	1.06	21	2.67	0.01	0.81	16	2.10	0.04	0.72
28	2.10	0.04	0.55	27	4.72	< 0.01	1.26	19	2.19	0.04	0.70
10	3.12	0.01	1.34	22	3.66	< 0.01	1.08	25	2.22	0.03	0.62
15	2.25	0.04	0.80	26	2.74	0.01	0.75	14	2.24	0.04	0.82
12	2.34	0.03	0.92	24	2.06	0.05	0.58	23	2.49	0.02	0.72

One set was created with optional stopping. One set was created with publication bias. One set is valid. Which is the valid set?

it does not exist. Too much replication success is a marker that something has gone wrong in the reporting, analysis, theorizing, or data collection process.

To demonstrate the impact of this kind of interpretation, consider the three sets of simulated results in Table 11.1. Each simulated data set was analyzed with a two-sample t-test. Each set of five studies is based on a different type of simulated experiment. For one set of experiments (valid testing) the population effect size is 0.8. The sample sizes, $n_1 = n_2$, were chosen randomly to be between 10 and 30. All five experiments produced significant results that were fully reported.

Another set of studies in Table 11.1 is based on simulated experiments where the population effect size is 0 (no effect). The sample was generated by an optional stopping approach that started with $n_1 = n_2 = 10$ and increased in steps of one up to a maximum size of 30. A total of twenty experiments were simulated and five of them happened to produce a significant result. Those five significant experiments were reported and the 15 non-significant experiments were not reported.

Another set of studies in Table 11.1 is based on simulated experiments where the true effect size is 0.1. The sample size was randomly chosen between 10 and 30. A total of 100 experiments were simulated and five of them happened to produce significant results. Those five significant experiments were reported and the 95 non-significant experiments were not reported.

The task for the reader is to determine which set of experiments in Table 11.1 corresponds to which simulation condition. Just to be clear, one set of experiments has a very large population effect size that was investigated with five proper experiments that were all fully reported. This is a valid set of experiments. A second set of experiments has no effect at all, but used optional stopping and publication bias to only report five significant results. This is an invalid set of experiments. A third set of experiments has a tiny effect size and used many experiments and publication bias to report only five significant results. This is also an invalid experiment set. Which is the valid experiment set in Table 11.1? We suggest the reader look at the statistics and make a judgment before reading the following text.

Did you spot the valid set? If you find it difficult or you are unsure, you may take some comfort in knowing that you are not alone. Even scientists with substantial experience evaluating statistical data often struggle to identify the valid experiment set in Table 11.1. It is worth recognizing the implication of this observation. With publication bias and optional stopping, scientists often do not know how to distinguish between a set of results where there is no effect at all and a set of results where there is a very large effect size.

Doing a meta-analysis that pools the reported effect sizes gives: for Set A $g^* = 0.82$, for Set B $g^* = 0.89$, and for Set C $g^* = 0.70$. Again, for many scientists, this meta-analytic information hardly helps identify the valid experiment set.

The Test for Excess Success is useful here. The calculations are left as an exercise for the reader, but computing power by using the meta-analytic effect size for each set, and multiplying the power values suggests that the probability of all five experiments producing significant results is: for Set A $p = 0.042$, for Set B $p = 0.45$, and for Set C $p = 0.052$. Indeed, it is Set B that is the valid experiment set.

The TES is a formal analysis, but there are rules of thumb that can be quickly used to gauge the validity of an experiment set. One thing to look at is the relationship between the sample size and the effect size. In a valid experiment set these numbers are unrelated (larger sample sizes lead to more precise estimates of effect sizes, but do not affect the magnitude of the effect size). For Set B in Table 11.1, a correlation between the sample size and the effect size gives a modest $r = 0.25$, which reflects random variation in the effect sizes across the experiments. In contrast for Sets A and C, $r = -0.86$ and $r = -0.83$, respectively. This relationship is easily understood when optional stopping is involved: a sample can be large only if the effect size happens to be small (if the estimated effect size were large a small sample would have been significant). One sees similar relationships in other sets of experiments, for example, the studies purporting to find evidence of precognition in Table 10.1 of Chap. 10 have a correlation between sample size and effect size of $r = -0.89$.

Another marker of a problematic data set is having many p-values close to, but always below, the criterion for statistical significance. Experiments run with optional stopping very often produce statistics with a p-value just below the criterion. In contrast, valid experiments with a real effect and appropriate sample sizes generally produce very small p-values; and values close to the criterion should be rare. One can see that Set B in Table 11.1 has almost all very tiny p-values, while the other experiment sets have many p-values between 0.02 and 0.05. Such a distribution of p-values should be a flag that something is odd about the set of experiments.

11.2 How Extensive Are These Problems?

So far we have established that some experiment sets have results that seem too good to be true, and that such findings should undermine our confidence in the validity of the original conclusions. The existence of some problematic experiment sets does not,

Table 11.2 Results of the TES analysis for articles in *Science*

Year	Short title	Success probability
2006	Deliberation-Without-Attention Effect	0.051
2006	Psychological Consequences of Money	0.002
2006	Washing Away Your Sins	0.095
2007	Perception of Goal-Directed Action in Primates	0.031
2008	Lacking Control Increases Illusory Pattern Perception	0.008
2009	Effect of Color on Cognitive Performance	0.002
2009	Monkeys Display Affiliation Toward Imitators	0.037
2009	Race Bias via Televised Nonverbal Behavior	0.027
2010	Incidental Haptic Sensations Influence Decisions	0.017
2010	Optimally Interacting Minds	0.332
2010	Susceptibility to Others' Beliefs in Infants and Adults	0.021
2010	Imagined Consumption Reduces Actual Consumption	0.012
2011	Promoting the Middle East Peace Process	0.210
2011	Writing About Worries Boosts Exam Performance	0.059
2011	Disordered Contexts Promote Stereotyping	0.075
2012	Analytic Thinking Promotes Religious Disbelief	0.051
2012	Stop Signals Provide Inhibition in Honeybee Swarms	0.957
2012	Some Consequences of Having Too Little	0.091

however, indicate that these kinds of problems are pervasive; it could be that such problems are rare. While we might have concerns about the specific studies that seem too good to be true, we would not necessarily worry about the entire field.

A way of examining the extent of these kinds of problems is to systematically analyze a specified set of studies. *Science* is one of the top academic journals; it has over 100,000 subscribers and is a major stepping-stone for any young scientist hoping to land a tenure-track position or be approved for tenure. One might hope that such a journal publishes the best work in any given field, especially given its very low acceptance rate of around 7%. The journal's on-line search tool reported 133 research articles that were classified as psychology or education and were published between 2005 and 2012. We applied the TES analysis to each of the 18 articles that had four or more experiments and provided sufficient information to estimate success probabilities.

Table 11.2 reports the estimated success probabilities for these 18 studies. Surprisingly, 15 out of 18 (83%) of the Science articles reported results that seem too good to be true (i.e., success probability is less than 0.1). The reader will probably recognize several of the short titles in Table 11.2 because many of these findings were described in the popular press and some have been the basis for policy decisions regarding education, charity, and dieting.

One study in Table 11.2 ("Disordered Contexts Promote Stereotyping") deserves special discussion. The lead author on this study was Diederik Stapel, a Dutch social psychologist who was found guilty of publishing fraudulent data. Indeed, the data in his *Science* paper was not gathered in a real experiment but was generated with a spreadsheet by the lead author (the other author was unaware of the fraud). You might think that a fraudster would insure that the data looked believable, but the reported (fake!) findings actually seem too good to be true. Very likely Stapel generated fake data that looked like real data from published experiments; unfortunately, the (presumably real) published data also often seems to be too good to be true.

The pattern of results in *Science* does not seem to be unique. A TES analysis for articles in the journal *Psychological Science* found a similar rate of excess success (36 out of 44, 82%, seem too good to be true). The problems do not seem to be restricted to psychology, as some papers on epigenetics and neuroscience show similar problems.

The overall implication of the analyses in Table 11.2 and other similar studies is that top scientists, editors, and reviewers do not understand what good scientific data looks like when an investigation involves multiple experiments and tests. At best, much of what is considered top experimental work in psychology, and other fields that depend on statistics, will probably prove unreplicable with similar kinds of experiments and analyses.

11.3 What Is Going On?

At this point it might be prudent to step back and consider how science got into its current situation. Certainly there is much pressure for scientists to report successful experimental outcomes (how else to get a faculty position or a grant?), but most (at least many) scientists seem to genuinely care about their field of research and they believe they are reporting valid and important findings that can have a positive impact on society. The implication seems to be that many scientists do not understand how statistical analyses contribute to interpretations of their empirical data. The following discussion is necessarily speculative, but it seems worthwhile to discuss some common confusions.

11.3.1 Misunderstanding Replication

As mentioned in Chap. 10, successful replication is often seen as the "gold standard" for scientific work. What many scientists seem to fail to appreciate, though, is that proper experiment sets show successful replication at a rate that matches experimental power (success probabilities, more generally). An emphasis on replication success blinded scientists from noticing that published experiments with moderate or low power were nevertheless nearly always working. Just due to random sampling, experiments with low power should not always work. Here, are some reasons why low powered studies so frequently deliver significant results even though they should not.

11.3.2 Publication Bias

Scientists may report experimental outcomes that support a certain theoretical perspective and not report experimental outcomes that go against that perspective. If every experiment provided a clear answer to the scientific question, this kind of behavior would be a type of fraud. However, it is often difficult to know whether an experiment has "worked." Given the complexity of many experiments (e.g., a cell culture must grow properly before you can claim to show some inhibitory effect of a suspected chemical), there are many reasons an experiment can fail. Scientists may treat an experiment that fails to show a desired outcome as one that suffers from a methodological flaw rather than one that provides a negative answer to the research question. Some scientists may have many studies that were improperly labeled as "pilot studies" but should actually have been treated as negative answers.

11.3.3 Optional Stopping

Empirically focused sciences constantly look for more data. Such an approach is valuable, but it often conflicts with the characteristics of hypothesis testing. For example, we noted in Chap. 10 how optional stopping inflated the Type I error rate of hypothesis tests. This problem is very difficult to solve within the hypothesis testing framework. For example suppose a scientist notes a marginal ($p = 0.07$) result in Experiment 1 and decides to run a new Experiment 2 to check on the effect. It may sound like the scientist is doing careful work, however, this is not necessarily true. Suppose Experiment 1 produced a significant effect ($p = 0.03$), would the scientist still have run Experiment 2 as a second check? If not, then the scientist is essentially performing optional stopping across experiments, and the Type I error rate for any given experiment (or across experiments) is unknown.

Indeed, the problem with optional stopping is not the actual behavior preformed by the scientist (e.g., the study with a planned sample size gives $p = 0.02$) but with what he would have done if the result turned out differently (e.g., if the study with a planned sample size gives $p = 0.1$, he would have added 20 more subjects). More precisely, if you do not know what you would have done under all possible scenarios, then you cannot know the Type I error rate for your analysis.

11.3.4 Hypothesizing After the Results Are Known (HARKing)

What may be happening for some scientific investigations is that scientists gather data from many experiments and then try to put together a coherent story that binds together the different results. That may sound like good scientific practice because it stays close to the data, but this approach tends to produce theories that are *too* close to the data and end up tracking noise along with any signal. A post hoc story can almost always be created to

justify why an effect appears or disappears. Likewise, findings that do not fit into a story can be labeled as irrelevant and properly (in the mind of the scientist) discarded from the experiment set.

This kind of post hoc reasoning applies for measures within an experiment as well as across experiments. A scientist may run multiple measures, identify one that seems to work across multiple experiments and conclude that this measure is the best. Again, this may seem like good scientific practice (and it can be, if done properly), but it often leads to selection of one measure on the basis of random sampling variation. The other measures may have been just as good (or better), but happened to not show the effect (or maybe they properly showed that the effect was not there).

11.3.5 Flexibility in Analyses

Modern software programs allow scientists to try a wide variety of analyses in search of statistical significance. The data do not show a significant result? Try transforming the data with a logarithm, or take the inverse of the data values and try again. Still no significance? Try removing outliers that are greater than three standard deviations from the mean, or 2.5 standard deviations from the mean. Remove floor effects or ceiling effects (Chap. 2), or data from participants who do not meet some other criterion. If you have multiple measures in your experiment you can combine them in a wide variety of ways (average them, take the max, multiply them, do a principle components analysis). While exploring your data is a perfectly good scientific exercise for an exploration study, it increases your Type I error rate in proper experiments. For this reason, if you tried various analysis for your data and found for an analysis a significant result, you need to replicate the experiment with this analysis and an independent sample.

Standard choices in analysis seem to encourage this kind of flexibility. Recall from Sect. 6.7 that a 2×2 ANOVA will have a 14% chance of producing at least one significant result (a main effect or an interaction) for a truly null data set. You can counteract this property by having a good understanding of precisely which tests are appropriate for your investigation. You would then (properly) ignore the outcomes of other tests reported by the ANOVA.

11.3.6 Misunderstanding Prediction

Scientific arguments seem very convincing when a theory predicts a novel outcome that is then verified by experimental data. Indeed, many of the *Science* articles analyzed in Table 11.2 include a phrase similar to "as predicted by the theory there was a significant difference." Such statements are very strange on two levels. First, even if an effect is real a hypothesis test is not going to produce a significant result every time. Sampling variability means that there will be some data sets that do not show the effect. At best a theory can

only predict the probability of a significant result given a certain sample size. Second, for a theory to predict the probability of success (typically, this is power), the theory must indicate an effect size for a given experimental design and sample size(s). None of the articles analyzed in Table 11.2 included any discussion of theoretically predicted effect sizes or power.

What this means is that the phrase "as predicted by the theory there was a significant difference" is empty of content for those papers. There may be a theory, but it is not the kind of theory that is able to predict the probability of an experiment producing a significant result. (Presumably, if it were that kind of theory, the scientists working with it would have discussed those details.) So, there actually is no prediction at all. The field seems to be in the bizarre situation where theoretical predictions that are not actually predictions seem to work every time. It indicates success at what should be a fundamentally impossible task.

11.3.7 Sloppiness and Selective Double Checking

Mistakes are inevitable in any kind of activity as complicated as science. Data entry errors, calculation errors, copy-and-paste errors can all lead to wrong interpretations. Although scientists are trained to check and re-check everything, these types of errors seem to be very common in published work. A computer program called STATCHECK can analyze published papers to check whether the reported statistics make sense. For example, if a paper reports $t(29) = 2.2$, $p = 0.01$ then there is definitely an error because $t = 2.2$ and $df = 29$ corresponds to $p = 0.036$. In a study of thousands of psychology articles, STATCHECK found that roughly half of published articles have at least one error of this type. Nearly 10% of articles had at least one reporting error that changed whether a reported result was significant.

These kinds of reporting errors may be indicative of a more general degree of sloppiness in research practices. Such a lack of care can mean that even well-intentioned researchers report untrustworthy findings. Worse still, the degree of care in data handling may correspond to whether the reported results match researcher's hopes of expectations. For example, if due to some error in data entry, a t-test finds $t(45) = 1.8$, $p = 0.08$, the researcher may re-check the data entry procedure, find the mistake and re-run the analysis to get $t(45) = 2.3$, $p = 0.03$. On the other hand, if a data entry error leads to a significant outcome such as $t(52) = 2.4$, $p = 0.02$, then the researcher may not re-check the data entry procedure even though doing so might have revealed the error and produced a result like $t(52) = 1.7$, $p = 0.1$.

Because there are so many places in scientific investigations where errors can occur, it is easy for scientists to unintentionally bias their results by selectively re-checking undesirable outcomes and selectively trusting desirable outcomes.

Take Home Messages

1. There is too much replication in many fields including medicine, biology, psychology, and likely many more.
2. It seems that many scientists use techniques they should better avoid: optional stopping, publication bias, HARKing, flexibility in the analysis, and many more.

Suggested Improvements and Challenges

12

Contents

What You Will Learn in This Chapter

The Test for Excess Success highlighted problems with current scientific practice for fields that use hypothesis testing. In addition, errors in statistical reporting (e.g., the reported p-value does not match the reported t-value) are common, results presented at conferences are changed for journal publications, and researchers admit to publication bias and other poor practices. These problems have motivated many researchers to propose counteractions designed to improve scientific practice. In this chapter, we critically evaluate some of these proposals. While there are some positive aspects to many of these proposals, they often have negative characteristics as well, and none of the proposals seem to tackle the more fundamental issues. Along those lines, we do not have a specific proposal that will address all of these problems, but we identify what we think should be the critical long-term goals of science and suggest that scientific practice should reflect those goals.

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_12

12.1 Should Every Experiment Be Published?

Some people suggest that scientists have an obligation to publish every experimental result, regardless of statistical significance. Indeed, non-significant experiments contain information about effects (both null and otherwise) that can only be utilized (through meta-analytic methods) if the data got published. More generally, publishing all the data allows readers to draw proper conclusions about effects and avoid overestimating effect sizes that are caused by publication bias.

However, there are several difficulties with publishing all experimental results. First, can experimenters reliably distinguish an experiment that failed methodologically (e.g., a piece of equipment failed) from those experiments that failed due to random sampling? If these different types of failures cannot be distinguished, then the literature becomes cluttered with experimental results that are flawed for a variety of reasons (of course, this might already be the case, but publishing everything may exacerbate the problem).

It is also a bit hard to see how scientists who publish everything should interpret their finding. Can there even be a conclusions section to a paper that simply adds an incremental amount of data to an existing topic? Moreover, when doing a meta-analysis, how does one decide which findings to include? These are all problems that exist right now, but publishing all the data does not remove them, and may make them worse.

Finally, when does the field decide that enough data has been collected? If a meta-analysis gives $p = 0.08$, should more experiments be run until $p < 0.05$? That approach would be problematic because it is optional stopping, just at the level of experiments added to a meta-analysis rather than at the level of individual subjects added to one experiment. Is there ever a moment when the field reports that an effect exists (see Chap. 3, Implication 1a)? What happens when more data changes the decision?

12.2 Preregistration

Several journals now encourage and promote replication studies, often with a requirement for researchers to preregister their experiment, analysis methods, and data collection plans. Some scientists consider preregistration to be the only viable path to move psychology (and other fields that use statistics) out of what they see as a crisis.

The idea of preregistration is that before actually running an experiment a scientist describes the total experimental plan in a place where the scientist cannot alter the original plan (e.g., the Open Science Framework, or AsPredicted.org). This plan describes the stimuli, tasks, experimental methods, number of samples and how they are sampled, the questions to be investigated, and the data analysis plan. After writing down these details, the experiment is run and any deviation from the preregistered plan is noted (perhaps with justification). Proponents of preregistration note that it prevents researchers from generating theoretical ideas or methods of data analysis after looking at the data

(HARKing). With preregistration, it would be obvious that a researcher stopped data collection early or added observations (perhaps due to optional stopping) or that various measures were combined in a way that is different from what was originally planned. If preregistered documents are in a public place, preregistration might also reduce the occurrence of publication bias because there is a public record about the researcher's intention to run the experiment; along similar lines, journals might agree to publish preregistered experiments prior to data collection.

These attributes all seem like good pragmatic reasons for scientists to practice preregistration. However, deeper consideration raises questions about what should be inferred when a researcher sticks to the preregistered plan. Does success for a pre-registered strategy lend some extra confidence in the results or in the theoretical conclusion? Does it increase belief in the process that produced the preregistered experimental design? A consideration of two extremes suggests that it does not.

Extreme Case 1 Suppose a researcher generates a hypothesis by flipping a coin. For example, a drug may increase or decrease working memory. The coin comes up "heads", so the researcher preregisters the hypothesis that the drug will increase working memory. The experiment is subsequently run and finds the predicted effect. Whether the populations truly differ or not, surely such an experimental outcome does not actually validate the process by which the hypothesis was generated (a coin flip). For the experiment to validate the prediction (not just the hypothesis), there needs to be some justification for the theory/process that generated the prediction. Preregistration does not, and cannot, provide such justification; so preregistration seems rather silly for unjustified experimental designs.

Extreme Case 2 Suppose a researcher generates a hypothesis and has an effect size derived from a quantitative theory that has previously been published in the literature. The researcher preregisters this hypothesis and the corresponding experimental design. The subsequent experiment finds the predicted difference. Such an experimental finding may be interpreted as strong validation of the hypothesis and of the quantitative theory, but it does not seem that preregistration has anything to do with such validation. Since the theory has previously been published, other researchers could follow the steps of the original researcher and derive the very same predicted effect size and thereby conclude that the experimental design was appropriate. In a situation such as this it seems unnecessary to preregister the experimental design because its justification is derived from existing ideas.

Most research situations are neither of these extremes, but scientists often design experiments using a mix of vague ideas, intuition, curiosity, past experimental results, and quantitative theories. It is impossible to gauge the quality of the experimental design for the vague parts; and preregistration does not change that situation. For those parts of the predicted hypotheses (and methods and measures) that are quantitatively derived from existing theory or knowledge, it is possible to gauge the quality of the experiment from

readily available information; and preregistration does not add anything to the quality of the design.

Preregistration does force researchers to commit to making a real prediction and then creating an experiment that properly tests that prediction. This is a laudable goal. But such a goal does not make sense if researchers do not have any hope of achieving it. When researchers design their experiments based on vague ideas, they are doing exploratory work, and it is rather silly to ask such researchers (or even to invite them) to make predictions. If forced to do so, such researchers may generate some predictions, but those predictions will not be meaningful with regard to the process by which they were generated. At best, such studies would provide information about a scientist's intuition, but researchers are generally not interested in whether scientists can generate good guesses. They run confirmatory studies to test aspects of theoretical claims.

At a practical level, many researchers who are motivated to preregister their hypotheses may quickly realize that they cannot do it because their theories are not sufficiently precise. That might be a good discovery for those researchers, and it may lead to better science in the long term. Likewise, preregistration does deal with some types of researcher degrees of freedom, such as optional stopping, dropping unsuccessful conditions, and hypothesizing after the results are known (HARKing). But these are exactly the issues that are handled by good justification for experimental design.

In summary, writing down the justifications for an experimental design may be a good activity for scientists to self-check the quality of their planned experiment. It may also be good to write down all the details and justifications of an experiment because it is easy to forget the justifications later. Moreover, when attempting to be so precise, it may often be the case that scientists recognize that part of their work is exploratory. Recognizing the exploratory parts of research can help guide how scientists interpret and present their empirical findings. However, justification for an experimental design should be part of a regular scientific report about the experiment; so there seems to be no additional advantage to publishing the justification in advance as a preregistration.

12.3 Alternative Statistical Analyses

There is a long history of criticism about hypothesis testing, and such criticisms often include alternative analyses that are claimed to lead to better statistical inference. While not denying that traditional hypothesis testing has problematic issues, and while personally favoring some alternative statistical methods, it is important to understand just what a method does and then determine whether it is appropriate for a particular scientific investigation.

For example, critiques of hypothesis testing sometimes claim that the p-value used in hypothesis testing is meaningless, noisy, or pointless. Depending on the situation, there may be merit to some of these concerns, but the general point cannot be true because the p-value is often based on exactly the same information in a data set (the estimated

signal-to-noise-ratio) as other statistics. For example, when an analysis is based on a two-sample t-test with known sample sizes n_1 and n_2, it is possible to transform the t value to many other statistics. An on-line app to do the conversion is at http://psych.purdue.edu/~gfrancis/EquivalentStatistics/.

This equivalence of information across the statistics suggests that what matters for using a particular statistic is the inference that is being made. If that inference is what you are interested in, then you should use it. Different choices of which statistic to use can give very different answers because they are addressing different questions. Just for illustration and without full explanation, we show a few examples to highlight the idea (it is not necessary to understand what the following terms exactly mean). If $n_1 = n_2 = 250$ and $d = 0.183$, then (if all other requirements are satisfied) the following are all valid inferences from the data:

- $p = 0.04$, which is less than the typical 0.05 criterion. The result is statistically significant.
- $CI_{95} = (0.007, 0.359)$ is a confidence interval for Cohen's d; it is often interpreted as describing some uncertainty about the true population effect size.
- $\Delta AIC = 2.19$, refers to the difference of the Akaike Information Criterion for null and alternative models. The value suggests that a model with different means better predicts future data than a model with a common mean for the two populations.
- $\Delta BIC = -2.03$, refers to the difference of the Bayesian Information Criterion for null and alternative models. The value provides evidence that the null model is true.
- JZS BF$= 0.755$, refers to a Bayes Factor based on a specific Jeffreys-Zellner-Siow prior, which provides weak evidence that the null model is true.

Thus, this data produces a significant result ($p < 0.05$) and favors the alternative model ($\Delta AIC > 0$), but it also provides some evidence that the null model is true ($\Delta BIC < 0$ and JZS BF<1). These conclusions might seem contradictory; but the conclusions are different because the questions are different. If you want to base decisions about effects by using a process that controls the Type I error rate, then the p-value provides the answer you are looking for. If you find the range of a confidence interval useful for representing uncertainty in an estimate of the standardized effect size, then the confidence interval provides what you want. If you want to estimate whether a model based on a common mean or a model with different means better predicts future data, then the ΔAIC value provides an answer. If you want to determine whether the data provides evidence for the null (common mean) or alternative (two different means) model, then the ΔBIC or the JZS BF provides the answer. Note that the ΔAIC, ΔBIC, and BF approaches have an option to *accept* the null hypothesis. Standard null hypothesis testing never accepts the null because absence of proof is not proof of absence (Chap. 3, Implication 3a).

12.4 The Role of Replication

Many scientists consider replication to be the final arbiter of empirical issues. If a finding replicates (perhaps by an independent lab), then the result is considered proven. A failure to replicate raises questions that have to be resolved (one group or the other must have made a mistake). Chapters 9–11 suggest that this view of replication is too simplistic when statistics are used. Simply due to random sampling, even well done studies of real effects will not always produce a successful replication.

Intuitions about the role of replication in science are largely derived from its role in the physical sciences. For example, acceleration of a feather in free-fall is the same as for a hammer, but only in a vacuum where air resistance does not impede their movement. The latter part of the previous sentence is especially important because it emphasizes that the outcome is dependent on the conditions and details of the experiment. For example, to successfully replicate free-fall acceleration in a vacuum it is necessary to have accurate measurements of distance and time; and a photogate timer is superior to an experimenter with a stopwatch. In addition, Newtonian physics posits that it does not matter whether the experiment is performed in the morning or afternoon, by male or female experimenters, or uses a dog treat and a battleship instead of a feather and a hammer.

Replication success in physics is nearly always determined relative to a theory. There is much experimental evidence that Newtonian physics is largely correct and that the type of object is irrelevant to free-fall acceleration, provided one has appropriate conditions and measurement precision. Under such situations, replication failures become especially interesting because they indicate a problem in the experimental set up (perhaps the vacuum has failed) or in the theory (photons have a constant velocity even under the effects of gravity, which leads to Einstein's relativity theory). Science is rife with stories where replication successes provided overwhelming support for a theory (replication as confirmation) and also where replication failures drive theory development.

In contrast to psychology, sociology, biology, and many more disciplines, an important characteristic of replication in the physical sciences is that the experimental outcome is (nearly) deterministic. Great care goes into identifying and reducing sources of noise. For example, a naïve experimental physicist might use the left and right hands to release objects, which would introduce some random difference in the release time. A better free-fall experiment would involve a mechanical device that was calibrated to insure simultaneous release of the two items, thereby largely removing one source of noise. For many phenomena in physics, only the motivation and resources to remove uncertainty limits this kind of careful control.

The situation is rather different for experimental psychology, medicine, and related fields. Some sources of noise can be reduced (e.g., by improving measurement scales or training subjects to perform better) but the limits imposed by the topic often exceed the motivation and resources of the experimenter. More importantly, there is often natural variability across the effect being measured (e.g., some people show an effect while other

people do not), that is, variability is part of the phenomenon rather than being added noise (Chap. 3, Implication 4). As a result, these fields often have no choice but to utilize statistical methods, such as null hypothesis testing, and they will sometimes produce inconsistent outcomes simply due to sampling variability.

For these fields to use replication in the way it is used by physics, studies need to work nearly every time (high power) and/or there needs to be a theory that distinguishes between sampling variability and measurement variability (Chap. 3, Implication 4a).

12.5 A Focus on Mechanisms

As we have seen throughout this book, problems of statistics are ubiquitous in many sciences. These problems are not only problems of bad statistical practice as outlined in the last chapters. The problems are often of conceptual nature. As we have seen in Chap. 3 Implications 4, science that is mainly based on statistics can often not disentangle true variability and noise and, thus, it remains an open question whether a significant effect is true in general or just holds true for a subpopulation. In addition as shown in the last subsection, failures of an experiment do not tell too much. We have the feeling that the problems with statistics are keeping scientists so busy that they may have lost focus on the perhaps most fundamental aspects of science: specification and understanding of the mechanisms that produce a particular outcome. Without such understanding it is impossible to predict future outcomes or to be confident that an empirical finding will be replicated in a new setting.

To have high confidence in any empirical result requires a corresponding theory that specifies the necessary and sufficient conditions. Even experimental findings that are generally reproducible cannot have a high level of confidence without a corresponding theoretical explanation because one cannot be sure that a new experimental condition will show the same result.

Consider the potential differences between two Magnetic Resonance Imaging (MRI) machines shown in Fig. 12.1. The MRI machine at the top right is located in Lausanne, Switzerland (top left), while the MRI machine at the bottom right is located in West Lafayette, Indiana (bottom left). How do engineers know that these two machines work similarly? There are many differences between Lausanne and West Lafayette that could, potentially, alter the behavior of the MRI machines. Lausanne has nearby mountains, a lake, stone buildings, and typical residents eat fondue and speak French. West Lafayette has nearby soybean fields, a river, brick buildings, and typical residents eat hamburgers and speak English. How should we know that these differences do not make the MRI machines behave differently? It is not enough to know that other MRI machines seem to function similar to each other; after all, every new machine is in a new environment and it is not feasible to test every possible environment.

Engineers have confidence in the behavior of the MRI machines because they understand how the machines work. For example, modern MRI machines use the properties of

Fig. 12.1 Two MRI machines in Lausanne, Switzerland (top) and West Lafayette, Indiana (bottom)

superconductivity, which was experimentally discovered in 1911. Even though superconductivity could be reproduced in many settings, it was not until the 1930s that a quantitative theory explained superconductivity. Further work in the 1950s explained superconductivity in terms of a superfluid of Cooper pairs, thereby connecting superconductivity to condensed matter physics and quantum mechanics. This kind of understanding allows scientists to identify the necessary and sufficient conditions to produce superconductivity and to predict its properties in MRI machines and elsewhere. Engineers have confidence that MRI machines will work not because previous studies have shown that they do work but because the field's theoretical understanding of superconductivity (and many other aspects of MRI machines) predicts that they will work despite some environmental differences.

As another example, consider the plague, which killed nearly one-third of people in Europe centuries ago. French scientist Paul-Louis Simond established in 1898 that fleas from rats transmitted the plague. Careful experiments justified this mechanism (he also identified the bacteria *Yersinia pestis*, which was infecting the fleas), as he showed that when fleas jumped from an infected rat to a healthy rat, the plague was transmitted. The rat-plague connection served as a (incomplete) mechanism: rats bring the plague. The implication of such a mechanism is clear, to reduce the occurrence of the plague,

reduce the number of rats: keep cats and dogs in the home to serve as rat predators, keep food stuff in sealed packages, set rat traps, avoid contact with live or dead rats (sadly, in the Great Plague of London in 1665, one suspected mechanism was that dogs and cats were spreading the plague, so they were exterminated in great numbers; which probably increased the rat population). When a case of the plague appeared in San Francisco in 1900, the scientific advice was to kill rats (but political conflict prevented this good advice from being generally applied). Note that the rat-plague mechanism does not have to make a quantitative prediction about exactly how many lives will be saved by various actions; it says that almost any action to reduce contact with rats is going to help control the plague. It also indicates what kinds of actions are unlikely to be helpful (e.g., isolation of an infected household). Of course, theories with more accurate mechanisms are even better. Nowadays the plague is kept in check by antibiotics, which directly attack the underlying bacterial cause.

As a final example, consider the effect of anesthesia. The effect of isofluran, one type of anesthetic, has been known for more than a century, and it enables complicated surgeries that would otherwise be impossible. Although the effects are very reliable, the mechanism by which isofluran induces anesthesia is unknown. Moreover, there are occasional failures where patients wake up in the middle of surgery or subsequently have memories about the surgery. If we understood the mechanisms by which isofluran induces anestheisia, we might be able to anticipate and compensate for these failures. Without a theory we do not know whether a failure is noise or hints at some hidden mechanism. In the meantime, doctors use anesthesia with the knowledge that they need to be ready should there be a failure.

Identifying mechanisms and justifying their role in a scientific phenomenon is very difficult to do, but it should be the long-term goal of every scientist. Some scientists may never actually achieve that goal, as they (valuably) spend their time gathering data and testing out ideas; activities that may help future scientists identify and justify mechanisms. Until justified mechanisms exist, scientists can never be confident that a particular effect will show up in a new setting.

In some sense, the long-term goal of science is to (as much as possible) remove the role of statistics by finding the "right" factors and thus reducing variability (see Chap. 6). An understanding of mechanisms promotes new hypotheses that can be rigorously investigated with good experimental designs. For example, an understanding of how vitamins affect bodily organs would explain why vitamins improve the health of some people but hurt the health of other people. Hence, with deeper insights into mechanisms many of the problems and concerns raised throughout this book largely disappear. Thus, to rejuvenate scientific practice the goal should not be to reform statistics but to not need it.

Take Home Messages

1. Many suggestions, such as preregistration, to improve statistical practice do not address the fundamental problems.
2. Good science involves more than just statistics.