

Patterns from Static: Philosophy and the Question Concerning Statistics

Table of contents

Preface	3
1 Philosophy, statistics, and the philosophy of statistics	4
1.1 What is philosophy?	5
1.1.1 A historical approach	7
1.1.2 Core subdisciplines of philosophy	8
1.2 What is statistics?	13
1.2.1 A very short and general primer on statistical inference	13
1.2.2 Pillars of statistical wisdom	14
1.3 What is the philosophy of statistics?	29
1.3.1 Philosophy in statistics	29
1.3.2 Statistics in philosophy	31
1.4 Discussion Questions	32
2 Contextualizing statistics	34
2.1 Types of inductive inference	35
2.1.1 Inference to the best explanation	35
2.1.2 Induction by enumeration	36
2.1.3 Inference from analogy	37
2.2 The problem of induction	38
2.3 The problem of induction and statistical philosophies	40
2.3.1 The falsification solution	40
2.3.2 The Bayesian solution	43

Preface

This is a Quarto book.

To learn more about Quarto books visit <https://quarto.org/docs/books>.

1 Philosophy, statistics, and the philosophy of statistics

I rush from science to philosophy, and from philosophy to our old friends the poets; and then, over-wearied by too much idealism, I fancy I become practical in returning to science. Have you ever attempted to conceive all there is in the world worth knowing—that not one subject in the universe is unworthy of study? The giants of literature, the mysteries of many-dimensional space, the attempts of Boltzmann and Crookes to penetrate Nature’s very laboratory, the Kantian theory of the universe, and the latest discoveries in embryology, with their wonderful tales of the development of life—what an immensity beyond our grasp!

— Karl Pearson, *The New Werther*¹

Over time, science, technology, engineering, and mathematics (STEM) curricula at many colleges and universities have become more and more specialized. Many Americans see higher education as a pathway to a good job, rather than, say, a pathway to educated citizenship (Skorton & Bear, 2018). There are good reasons to view higher education in this way; rising costs make it difficult for students to justify studying subjects that do not have a clear return on investment. STEM fields in general, and statistics and data science in particular, are seen as a great return on investment (Davenport D.J et al., 2012). So why should training in a STEM field include the study of an (ostensibly) esoteric field like the philosophy of statistics? In my view, there are two reasons. First, as Karl Pearson alludes to above, stepping outside of one’s primary STEM concentration, and diversifying one’s skills and knowledge, can be a real joy. Second, as I hope to demonstrate throughout this book, an awareness of philosophical issues in statistics can make us better statisticians and data scientists.

The preceding paragraph suggests that the audience of this book will consist primarily of individuals from STEM fields. But in addition to statisticians who wish to know something more about philosophical issues in their own discipline, I also hope that thi book will find an audience of philosophers that wish to know more about important conceptual issues in statistics. Consequently, this chapter provides an introduction to each discipline to “bring everyone up to speed”.

¹Note that some philosophers do not classify IBE as a type of induction (or deduction); such philosophers carve up the space of non-deductive arguments differently than we have here, to leave space for IBE as its own type of inference. See Chapter 2 of Okasha (2016) for more details.

1.1 What is philosophy?

“Philosophy is a field that, unfortunately, reminds me of that old... joke, ‘those that can’t do, teach, and those that can’t teach, teach gym.’”

— Lawrence Krauss, interview in The Atlantic²

Recently, popularizers of science have suggested that philosophy is a useless undertaking, a waste of time, and something that distracts us from making progress on real problems. For example, in a 2014 interview on the Nerdist Podcast, Neil de Grasse Tyson expressed his irritation with philosophers “asking deep questions” that lead to a “pointless delay in progress” (Dvorsky, 2014). Similarly, Stephen Hawking has claimed that deep questions in science, such as those concerning the fundamental constituents of the universe will only be answered using data from science, such as data coming from space and particle physics. Hawking writes:

Most of us don’t worry about these [philosophical] questions most of the time. But almost all of us must sometimes wonder: Why are we here? Where do we come from? Traditionally, these are questions for philosophy, but philosophy is dead. Philosophers have not kept up with modern developments in science. Particularly physics (Hawking & Mlodinow, 2010).

Given the claims made about philosophy by such respectable figures, one might reasonably wonder why we should embark on a journey into the *philosophy of statistics*; not only might philosophy and statistics seem unrelated; the former, it is claimed, is useless!

We should reject these attacks against philosophy; but in order to understand *why* we should reject them, and ultimately, to justify our study of the philosophy of statistics, we first have to achieve clarity in our conceptual framework. Most pressingly, especially for those of us who are statisticians and data scientists, we must ask: what is *philosophy*?

If one has never studied philosophy in a formal setting, one is likely have certain misconceptions about what academic philosophy is and what philosophers do. It is commonly thought (wrongly, in my view) that philosophy is entirely subjective, vague, imprecise, and incapable of progress (Park, 2020). These misconceptions are often born out of the way that the word ‘philosophy’ is used in colloquial settings. One use of the word ‘philosophy’ captures an individual’s personal outlook on life. For example, Apple co-founder Steve Jobs, at the 2005 Stanford Commencement Address said the following:

Your time is limited, so don’t waste it living someone else’s life. Don’t be trapped by dogma—which is living with the results of other people’s thinking. Don’t let the noise of others’ opinions drown out your own inner voice. And most important,

²Arguably, using BIC for *explanation* rather than *prediction* would require that we know something about the extent to which each input variable in the statistical model is causally related to the output variable. BIC does not, on its own, select for causal relationships, and such relationships are typically what is desired in an explanation.

have the courage to follow your heart and intuition. They somehow already know what you truly want to become (Jobs, 2005).

Colloquially, we might say that this is Steve Jobs' (personal) *philosophy*. Of course, there is nothing wrong with holding a personal philosophy, but holding one does not imply that one has *done philosophy* in the academic or historical sense.

To distinguish between personal philosophies and academic philosophy, let's look at how professional philosophers and professional philosophical organizations attempt to answer the question 'What is philosophy?' In the magazine *Philosophy Now*, artist and philosopher Colin Brookes writes that "philosophy critically examines anything and everything, including itself and its methods. It typically deals with questions not obviously addressed by other areas of enquiry, or those that remain after their activity seems complete" ("What Is Philosophy and How Do We Do It?" 2009), the American Philosophical Association describes philosophy as a field that

pursues questions in every dimension of human life...its techniques apply to problems in any field of study or endeavor. No brief definition expresses the richness and variety of philosophy. It may be described in many ways. It is a reasoned pursuit of fundamental truths, a quest for understanding, a study of principles of conduct. It seeks to establish standards of evidence, to provide rational methods of resolving conflicts, and to create techniques for evaluating ideas and arguments.(American Philosophical Association, 2024)

Finally, Jon Wainwright claims that "philosophy involves the analysis of arguments and concepts...power of reason...weight of evidence...[and] exposes unsupported assertions, prejudice" ("What Is Philosophy and How Do We Do It?" 2009).

Already, we might notice that academic philosophy differs from one's personal philosophy in many ways:

1. Personal philosophies are not necessarily critical examinations.
2. Personal philosophies might well be (and often are) absent of method. We might ask, how did Jobs *arrive* at this philosophy? It's not entirely clear.
3. Academic philosophy critically examines "anything and everything"—including statistics! Philosophy is a very intellectually diverse discipline; personal philosophies are typically much more limited in scope.
4. As might be clear after hearing your uncle's personal philosophy over Thanksgiving dinner, personal philosophies are not always (attempts at) "reasoned pursuits of fundamental truths", and do not always consider evidence, expose unsupported assertions, etc.

In addition to seeing how academic philosophy differs from personal philosophies, we also get a sense of some of the fundamental features of philosophical investigation. We see that reason, evidence, the analysis of arguments, concepts, and assumptions are all core features of philosophy. Given that science also cares about reasons, evidence, and the like, philosophy sounds a lot like science. So, what's the difference? To answer this question, it will be important to consider some of the historical roots of science and philosophy.

1.1.1 A historical approach

To the extent that science is concerned with causes and principles of the natural world, many of the earliest ancient Greek philosophers may also be considered scientists (Curd, 2016). For example, Thales of Miletus (c. 620 B.C.E.—c. 546 B.C.E.) is often identified as the first person to investigate the basic natural principles and the question of the originating substances of matter; therefore, we may consider him a founder of natural science. The historical connection between philosophy and science does not end with Thales; Plato, Aristotle, Francis Bacon, Galileo Galilei, René Descartes, and Isaac Newton were all considered both philosophers and scientists. Aristotle, most often considered a philosopher, made contributions to geology, physics, zoology, biology, and medicine. Descartes and Newton both made important contributions to metaphysics and epistemology—subdisciplines of philosophy—as well as physics and mathematics. In fact, until around the 19th century, what we now call science was called “natural philosophy” (Cahan, 2003).

It was not until the 18th and 19th centuries that philosophy and science started to split apart as two “separate” disciplines. One explanation for this split is that, at around this time in history, many thinkers developed empirically rooted answers to important questions. Once answers became available and more broadly accepted, these fields split apart from philosophy into their own disciplines. Philosophy then, gets stuck with all of the hard questions for which empirically rooted answers are not (yet) available.

This theory, though it may be incomplete (Papineau, 2018), illuminates two important features of philosophy. First, on this view, the charge that philosophy does not make progress—a charge made by Neil de Grasse Tyson, Lawrence Krauss, Stephen Hawking, among others—is misguided. Philosophy *does* make progress; it's just that once it progresses, we often stop calling it philosophy! Second, on this view, we see that philosophers are not “anti-empirical”; they very much care about and value empirical evidence. It just so happens that many of the (important!) questions that they are concerned with are *underdetermined* by all of the available empirical evidence; that is, the available empirical evidence equally supports several different answers to a given philosophical question, and philosophers must resort to other tools. Thus, the difference between philosophers and scientists is not that, somehow, the latter are more intellectually rigorous. Rather, it's that the latter limits herself to questions that, at present, are empirically driven. Such a difference is not disparaging to philosophers. Many of the most important questions about us and our world have not yet been decided by, and perhaps *cannot* be decided by, empirical evidence alone. Such questions—for example, what

makes a just society? what set of criteria clearly demarcate science from pseudo-science?—may be of critical importance. Philosophers use important and imaginative tools of reasoning, such as thought experiments, to discover answers to these questions.

Historically then, it seems that philosophy was a broad category that included the sciences (e.g., physics, biology) as subdisciplines. But now, if philosophy no longer includes the sciences, what is its content?

1.1.2 Core subdisciplines of philosophy

It is standard to parse the discipline of philosophy into several subdisciplines. For simplicity, we will look at four: logic, metaphysics, epistemology, and ethics. We will consider each of these, noting that there is no clean and uncontroversial way to partition the field of philosophy; there is much overlap, between the subdisciplines presented here. Also, we note that many philosophers work in fields denoted *the philosophy of X*, where X is some other field or concept, such as physics, psychology, biology (or science more broadly), mind, mathematics, or...statistics!

1.1.2.1 Logic

As noted above, reason, evidence, and the analysis of arguments are core features philosophy. The branch of philosophy that has as its focus the analysis of arguments is called *logic*. As an entry point into defining logic—and delimit it from other branches of philosophy, and from science itself—consider the following three arguments:

Argument #1

P1 On any given day, if it is raining, then Newman will not go on his postal route.

P2 Today, it is raining.

C So, today, Newman will not go on his postal route.

Argument #2

P1 If Kramer swims in the East River, he will smell bad.

P2 Kramer smells bad.

C So, Kramer swam in the East River.

Argument #3

P1 The car salesman claimed that George's 1989 Chrysler LeBaron convertible was owned by the actor Jon Voight.

P2 The owner's manual shows that the previous owner's last name was Voight.

C Therefore, the previous owner of George’s car was Jon Voight.

In each case, the author of the argument is using the premises—P1 and P2—as reasons to believe the conclusion, C.³ *But in what sense do the premises provide good reasons for believing the conclusion?* Logic, generally defined as the study of correct reasoning, attempts to answer this question. In **Argument #1**, we should note that the premises provide good reasons for believing the conclusion because it is *impossible* for the premises to be true and the conclusion to be false; such an argument is called *deductively valid*, and the premises are said to *logically entail* the conclusion. Arguments that either are or attempt to be deductively valid are called *deductive arguments*.

We might be enticed to give the same analysis of **Argument #2** that we gave of **Argument #1**; however, **Argument #2** is invalid. To see this fact, consider that Kramer might smell bad for a whole host of reasons; he may, for example, have just finished his Karate lesson.

Argument #3 is a bit different in that the premises do not logically entail the conclusion, but they may give good reasons to believe the conclusion—there are not that many people with the last name ‘Voight’, actors like snazzy convertibles, and the salesman’s testimony provides some basis for believing the conclusion. But of course, the car might be owned by *John* Voight the periodontist, not *Jon* Voight the actor. Arguments like **Argument #3**—ones that might provide good reasons to believe the conclusion but don’t *logically entail* it—are called *inductive arguments*.

We should note that the assessments of these arguments is not entirely *empirical*. We need not check anything about the empirical, physical world—e.g., that it is in fact raining—to assess whether **Argument #1** is valid. Rather, many assessments of arguments are based on philosophical reasoning that need not consult with empirical reality. Scientists sometimes assert that reason and logic fall under the purview of science, but historically, it is a branch of philosophy. Further, to the extent that science is concerned with empirical considerations, logic is not a science (though, we note that logic is essential to the proper functioning of science!). In the chapters to come, we will consider the benefits of thinking of statistics as a branch of logic—a branch that helps us reason properly about incomplete, uncertain data.

1.1.2.2 Metaphysics

What does it mean to say that *X causes Y*? On the surface, this may seem like an easy question. The gas pedal *caused* the car to move forward. The toxic envelope glue *caused* Susan’s death. But deciding on what causal relations exist in the world can be, in fact, quite difficult. Perhaps the most famous exposition of the difficulties of causality are given by the 18th century philosopher David Hume. As an empiricist philosopher, Hume believed that knowledge of a causal relationship between any two objects must be based strictly on experience.

³A more modest version of the conclusion of enumerative induction is (C) Therefore, *the next* unobserved instance of *A* will have property *p*.

But, according to Hume, experience can only reveal temporal relationships—that *Y* occurred *after* *X* occurred—and contiguity—that *X* and *Y* have been in contact. Experience cannot establish a *necessary* connection between cause and effect—that *Y* happened as the result of *X*—because one can imagine, without logical contradiction, a case in which the cause does not produce its usual effect (e.g., one can imagine that Susan licked the envelopes but did not die). According to Hume, we mistakenly believe that there are causes in the world because past experiences have created a habit in us to think in this way. Really, we have no *direct knowledge* of anything more than spatial and temporal contiguity; anything else that we infer about causality in the world lies beyond direct experience (Morris & Brown, 2019).

Hume’s discussion of causality should be concerning to those of us interested in statistics and science. Many would agree that modern science relies heavily on statistical methods to attempt to provide information about causal relationships; but it seems reasonable to ask whether statistical methods are well-equipped to account for anything more than correlations among variables. But establishing a causal relationship would require going beyond mere correlations. Although correlations may suggest a causal relationship between two variables, correlations are not sufficient for establishing a causal relationship.

The question about the nature of causality can be thought of as a *metaphysical* question. Metaphysics is the study of the fundamental nature of reality. Why is there something rather than nothing? Are space and time discrete or continuous? What is time, and what does it mean for entities to persist through time? Since metaphysics is not constrained by the need for empirical verification, some might think of metaphysics as asking *why?* in a larger domain than science typically does. However, we should note that (good) metaphysics ought to be consistent with known empirical results of science and ought not be internally contradictory.

The scientifically-oriented reader—perhaps in agreement with de Grasse Tyson, Hawking, and Krauss—might posit that metaphysical questions like the ones given in the previous paragraph are ultimately a waste of time. However, developments in philosophy in the twentieth century suggest that it is not so easy to dismiss metaphysics. Culminating in the mid-twentieth century, a movement called *logical positivism* (also known as *logical empiricism*), composed of scientists and empirically minded philosophers, sought to do away with metaphysics. Logical positivists adhered to what is sometimes called the *verifiability criterion of meaning*. This criterion states that only claims that can (at least in theory) be verified empirically, or claims that are logical tautologies, count as genuine, meaningful knowledge (Dphil, 2009). All other claims—e.g., metaphysical claims about causality, god, the nature of being, etc.—are meaningless. For example, following Hume, the logical positivists believed that causal relations were not directly observed, and could not be directly measured; thus, claims about causal relations were meaningless.

It is generally accepted that, with respect to the verifiability criterion of meaning, the logical positivist program is untenable, for at least two reasons (Fetzer, 2010). First, the criterion itself is thought to be self-refuting. After all, the proposition “only claims that can (at least in theory) be verified empirically, or claims that are logical tautologies, count as genuine, meaningful

knowledge” is neither about the physical world, nor is it a logical tautology.⁴ The second criticism of the verifiability criterion—which may be particularly interesting to statisticians—is closely related to data collection. That claim C can be verified empirically assumes that one can go out into the world and collect data relevant to C . But we might wonder: what principles guide decisions about which data are relevant to C , and which are not? Surely, data collection is guided, at least in part, by theory;⁵ to see this, consider measurements taken by a bulb thermometer. Such thermometers rely on, among other things, a theory about the way in which liquid takes up space at different temperatures. Importantly, we might challenge the use of an anomalous temperature reading by challenging whether the particular thermometer used was calibrated properly, and calibration relies on the underlying liquid-temperature theory. If theory guides our data collection processes, then “empirical verification” is no longer entirely empirical; it is tainted by theory. As such, the verifiability criterion seems suspect, and we might entertain the meaning of metaphysical claims; long live metaphysics!

1.1.2.3 Epistemology

Above, we saw that the nature of causality was a metaphysical question. But, suppose, in some future utopia, metaphysicians have uncovered the nature of causality; that is, the question *what is a causal relation?* has been answered. This fact in itself would not lay to rest all philosophical questions related to causality. Even if we have defined a causal relation, we might still wonder how to *gain knowledge* about causal relations. For example, an account of what it means for cigarette smoking to cause cancer does not necessarily provide an answer the question *how do we know that cigarette smoking causes cancer?*

What does it mean when we say that an agent A *knows* a claim C , for example, that “the Moors invaded Spain in the 8th century”? Clearly, in order to know C , A must actually *believe* it. If A doesn’t believe C , it would be odd to say that A actually knows C . Similarly, it would be odd to give A ’s belief the status of knowledge if C weren’t, in fact, *true*. Even if, for some reason, A believed that “ $2 + 2 = 5$ ”, this belief would not constitute knowledge. Finally, according to the canonical view of knowledge, first espoused by Plato, a *true belief* is not sufficient for claiming knowledge; knowledge also requires *justification*. Suppose that A had no idea whether C were true, and decided to believe it based on a coin flip. Such a belief, even though true, would hardly count as knowledge because A had no justification for the belief in C .⁶

In addition to asking for a definition of knowledge, epistemologists are also interested in, among other things, questions about sources of knowledge—e.g., given that our perception is fallible, under what conditions is it reliable for producing knowledge?—the limits of knowledge—e.g., are there some questions for which the answer is unknowable?—and the meaning of justification. Because science is thought to play such an important role in knowledge generation,

⁴See Bryda (2013) for evidence of the claim that there are such similarities.

⁵My explanation of Hume’s argument relies on Henderson (2018).

⁶Hypothesis testing was developed separately by Fisher, on the one hand, and Neyman and Pearson on the other. The version often taught is a blend of these two methods.

epistemologists are especially interested in scientific discoveries, and the methodologies that lead to such discoveries.

Many epistemologists are familiar with, and make use of, statistics in their work. Some make use of statistical methodologies as frameworks for reliable knowledge generation—as a way to update beliefs based on new information. Others interrogate the reliability of certain statistical methodologies (e.g., hypothesis testing) for generating knowledge. In **?@sec-frequentist** and **?@sec-Bayesian**, we will learn about, and consider objections raised against, popular statistical methods.

1.1.2.4 Ethics

In 2017, neuropathologist Dr. Ann McKee published a paper examining the brains of 202 deceased football players. Of the 111 NFL players examined, 110 of those were found to have chronic traumatic encephalopathy (CTE) (Ward et al., 2017). CTE is a degenerative disease believed to be caused by repeated blows to the head and can only be diagnosed after death; so, there is no way to know how many living NFL players have the disease. Although McKee’s sample of brains of NFL players was far from random—many of the brains in the sample were from players whose families suspected that CTE was present—there is still some scientific basis for concluding that NFL player’s run a serious risk of developing CTE. About 1,300 former players have died since the McKee’s group began studying CTE; so, even if every one of the other 1,200 players had tested negative—an implausible scenario—the minimum CTE prevalence would be close to 9 percent. This rate is vastly higher than in the population of non-football players (Ward et al., 2017).

Typically, we think about sports in terms of *personal preference*. As with many other preferences—whether we prefer the mountains or the beach; bananas, apples or oranges; Apple or Android; vanilla or chocolate—sports preferences seem personal; you might enjoy football, and I might enjoy hockey, and there is no compelling reason why either of us should change our preference. The study of CTE, however, challenges this view about sports, at least with respect to football. It appears that playing football comes with serious risk. We might ask whether one *ought to* play football given those risks. Further, we might ask whether we, as a society, *ought to* idolize and support a game that encourages millions of young people to risk serious injury for a very small chance of success.

Whatever you think about these questions—and reasonable people might disagree about the answers—it seems clear that there is a *moral* or *ethical* component to them. Almost always, when we ask questions about what we *ought to* do, either as individuals, small groups, or as a society, we are asking ethical questions. Ethicists ask a wide range of questions, including: What does it mean to live a *good* life? Is it possible to derive what we *ought to* do from what is the case?⁷ Do we have special obligations to the global poor? Ought we eat animals? Is abortion

⁷If this formulation of Bayes’ theorem does not look familiar to you, do not worry. We will discuss Bayes’ theorem in detail in **?@sec-Bayesian**.

permissible? What obligations do we have to the environment? Ought we make consequential decisions about mortgage loans based on uninterpretable machine learning algorithms? We will consider some ethical questions throughout this book.

1.2 What is statistics?

To contextualize the discipline of statistics, it might be helpful to recall a distinction made in Section 1.1.2.1—the distinction between deductive and inductive logic. Recall that an argument is deductive just in case the premises *logically entail* the conclusion. That is, it is impossible for the premises to be true and the conclusion to be false. By contrast, an argument whose premises do not logically entail the conclusion is inductive. Of course, inductive arguments can be very strong; the fact that objects, in the past, have an acceleration due to gravity of (approximately) 9.81 m/s^2 provides good reasons to believe that future objects will have this same acceleration due to gravity. But, this conclusion doesn't necessarily follow; we can *conceive* of a world in which physical laws might change. What methods reliably produce strong inductive arguments? In empirical domains that allow for the collection of data, inferential statistics can be thought of as a set of methods for drawing conclusions about the world from limited information. The conclusions go beyond the data at hand, and thus, the arguments that statistics presents are inductive.

This analysis gives a very high level contextualization of statistics. Where do we go from here? What are some of the actual methods or principles that statistics utilizes to reliably draw conclusions? First, it will be instructive to introduce some terminology to help understand inference problems. Then, we will consider seven foundational principles of statistical theory and practice.

1.2.1 A very short and general primer on statistical inference

As mentioned above, inferential statistics can be thought of as a set of methods used for drawing conclusions about the world from limited information. The limited information is given in a *dataset* or *sample*, and will consist of *variables of interest* measured for each of n *units* in the sample (the entities about which we want to learn). The set of all of the units about which we want to learn—including all units in the sample, and almost always, units not in the sample—is called the *target population*.

For example, suppose that we are interested in learning about the spending practices of customers of artist *A*. To do so, we might ask a randomly selected group of $n = 25$ people at an artist *A* concert some questions: their age, gender, income, cash on hand, proportion of times they've purchased merchandise at a concert before, etc. In this case, the units are individual concertgoers of artist *A*; the sample consists of the $n = 25$ randomly chosen concertgoers of whom we asked questions; the population consists of all potential concertgoers of artist *A*; the

variables of interest are age, gender, income, cash on hand, proportion of times merchandise has been purchased, etc.

We might be interested *describing* or *summarizing* individuals in the sample. Some examples might be: how much cash does the typical person in the sample have on hand? Or, what proportion of people in the sample have never purchased merchandise at a concert before? But such summaries are limiting in that they only tell us about this sample, and not about the larger population.

Alternatively, we might be interested in *inferring* a particular feature of the entire population—such features are called *parameters*—based on the sample. For example, we might be interested in inferring the average income of potential concertgoers of artist *A*. Or, we might like to predict how likely is it that a particular person will purchase an item given that they are 28 years old, female, earn \$45,000 per year, have \$35 in hand, and have purchased merchandise at 10% of the concerts that they’ve attended before. To make such inferences, we need to do more than simply summarize samples. Importantly, to conduct statistical inference, we need to construct a statistical model that represents the data well. We will discuss some particulars about statistical models and inference methods in later chapters. For now, with this setup in hand, we will turn to some features—or pillars of statistical inference—that different inference methods have in common.

1.2.2 Pillars of statistical wisdom

In *The Seven Pillars of Statistician Wisdom*, Stephen M. Stigler attempts to answer an important question posed above: what are some of the actual methods or principles that statistics utilizes to reliably draw conclusions? In doing so, Stigler formulates a possible answer to the question *what is statistics?*, by presenting seven principles that form a conceptual foundation for statistics as a discipline. He writes:

In calling these seven principles the Seven Pillars of Statistical Wisdom, I hasten to emphasize that these are seven *support* pillars—the disciplinary foundation, not the whole edifice, of Statistics. All seven have ancient origins, and the modern discipline has constructed its many-faceted science upon this structure with great ingenuity and with a constant supply of exciting new ideas of splendid promise. But without taking away from that modern work, I hope to articulate a unity at the core of Statistics both across time and between areas of application Stigler (2016).

It should be emphasized that these principles—aggregation, information, likelihood, inter-comparison, regression, design, and residual—are not necessary and sufficient conditions for what constitutes statistics; for example, the aggregation of information is not necessarily an example of a statistical analysis, and the omission of experimental design does not disqualify an analysis from being statistical. Instead, we might think of analyses counting as “statistical” as having a *family resemblance* to one another (Wittgenstein, 2001 (1953)), and Stigler’s pillars

are common to many (but not all). We discuss each of these pillars in turn, and highlight places where each pillar borrows from or makes use of philosophy, emphasizing again that statistics can be understood as a branch of philosophy. Note that Stigler (2016) takes a historical approach to the pillars; the approach here is less historical and more conceptual.

1.2.2.1 Aggregation

Aggregation is the combining of observations for the purposes of information gain. At first, aggregation might seem odd. Suppose that we have n individuals, and for each individual, we measure a single variable—e.g., an individual’s yearly income. What does one *gain* by reducing n measurements to a single number, for example, the arithmetic (or *sample*) mean, median, or mode? We typically think of these numbers as *measures of center*; thus, they are meant to tell us about the *average* or *typical* unit under study. But, of course, it might be the case that no unit takes on the mean or median, and in fact, sometimes it is *impossible* for an individual unit to take on these measures of center! So, in what sense are they measuring something typical?

First uses of the sample mean as a measure of center in the social sciences saw criticisms along these lines. For example, as reported in Stigler (2016), the Belgian statistician Adolphe Quetelet used the mean as a way of comparing human populations with respect to a particular variable—e.g., height. Stigler (2016) writes:

Already in the 1840s a critic was attacking the idea. Antoine Augustin Cournot thought the Average Man would be a physical monstrosity: the likelihood that there would be any real person with the average height, weight, and age of a population was extremely low. Cournot noted that if one averaged the respective sides of a collection of right triangles, the resulting figure would not be a right triangle (unless the triangles were all proportionate to one another).

Nevertheless, Quetelet thought that the mean was meaningful, and could stand in as a “typical” individual, or “a group representative for comparative analysis” Stigler (2016). Of course, the practice of using the sample mean to summarize the center of measurements with respect to a given variable is common practice; the sample mean does well at describing what is “typical” in certain contexts, but not in others. The sample mean is not particularly robust to outliers, which means that the addition of outliers can have a large effect on the value. The sample median—the value at which half of the measurements are above and half are below—is more robust to outliers, and thus, in some cases, more appropriate.

Measures of center are not the only forms of aggregation, and in fact, if reported alone, a misleading picture of the data often emerges. For example, it might be important for one to live in a city where the average daily high temperature in the summer months is 70 degrees Fahrenheit. But that information is not enough, because (it is at least conceivable that) a city with such an average might have many summer days with a high temperature of around 30 degrees, and many others with a high temperature of around 100 degrees, such that the

average is around 70 degrees. These sorts of temperature swings are likely not in accordance with the desire to live in a city with an average daily high temperature in the summer months of 70 degrees! Missing in this example is some measure of variability; measures of variability, such as the range and variance, also combine observations for the purposes of information gain and summary, and thus, are aggregations.

It is important to note that aggregation does not just occur as simple summary statistics. For example, consider the statistical model of the form $Y_i = f(x_i; \theta) + \varepsilon_i$, where θ are a vector of parameters, $\theta = (\theta_1, \dots, \theta_p)$; $f(x_i; \theta)$ represents the mean of Y_i at a given x_i and θ ; and ε_i represents random error (with zero mean).⁸ Estimates of θ , found for example, by least squares or maximum likelihood estimation, can be thought of as “weighted aggregates of data that submerge the identity of individuals” Stigler (2016).

Finally, we note that discussions above about measures of center have philosophical and empirical content; the choice of the median over the arithmetic mean as a summary statistic relies on the meaning and understanding of the concepts “typical” or “average”, and empirical considerations alone cannot tell us what is the right meaning of the term “typical” in a given context. Aggregation—a pillar of statistical wisdom—is informed by philosophical considerations!

1.2.2.2 Information

In studying aggregation, we learned that we can gain information by combining observations. Let’s expand upon this idea a bit. Suppose we have a jar full of c candy beans,⁹ where c is unknown. We’d like to estimate c . Our estimation process is as follows: we ask a diverse group of n people to each give an independent estimate of c . Call each estimate X_i , $i = 1, \dots, n$.¹⁰ We then average the n values together, using the sample mean:

$$\bar{X} = \sum_{i=1}^n X_i.$$

Here, we’ve combined observations in a way that increases information about c . That is, \bar{X} will be more precise as an estimator of c than any individual guess, X_i . But how much more precise? What is the relationship between n and precision? How much information do we gain by, say, doubling the number of (independent) guesses? It turns out that, if the standard

⁸We will consider issues related to interpretations of probability—e.g., whether probability *just is* a ratio of frequencies or not—in [?@sec-probability](#).

⁹<https://bit.ly/31Gphnd>

¹⁰We suppose that each person’s guess would be correct, up to some random error or perturbation. Another way of saying this is that, if we could somehow ask each person to give an estimate, record it, erase their memory, and repeat this process many times, on average, they would be correct. Further, we suppose that the random error (i.e., the standard deviation of each guess) is the same across all people.

deviation of each guess is the same—call it σ —then some simple probability theory can give us an answer:

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n} \sum_i^n X_i\right) \stackrel{i}{=} \frac{1}{n^2} \sum_i^n \text{Var}(X_i) \\ &= \frac{1}{n^2} \sum_i^n \sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

Thus, since the standard deviation is the square root of the variance, $sd(\bar{X}) = \sigma/\sqrt{n}$. If we think of information gain as an increase in the precision of our estimator \bar{X} , and we measure precision using the (multiplicative) inverse of the standard deviation, then we see that, to increase the precision of our estimator by a factor of k , we need to multiply the number of guessers by k^2 : $k/sd(\bar{X}) = \frac{k}{\sigma/\sqrt{n}} = \frac{k\sqrt{n}}{\sigma} = \frac{\sqrt{k^2 n}}{\sigma}$. As (Stigler, 2016) writes:

The implications of the root- n rule were striking: if you wished to double the accuracy of an investigation, it was insufficient to double the effort; you must increase the effort fourfold. Learning more was much more expensive than generally believed.

Note that we made some important assumptions when describing information gain and precision in terms of the root- n rule. One important assumption was that the guesses were independent. By independent, we mean that no individual guesser was influenced, either directly or indirectly, by any other guesser. It turns out that, without independence, the derivation above is not correct; $sd(\bar{X})$ will be larger.¹¹ What can we say about information gain in such cases? Intuitively, if guesser X_i influences X_j , we would expect our sample to contain *less* information than if no influence occurred. To quantify how much less, we could calculate an *effective sample size*, n_e , which would be less than n whenever measurements are positively correlated.

1.2.2.3 Likelihood

Consider a thought experiment given by Sir Ronald Fisher in his 1935 work *Design of Experiments* (R. A. Fisher (1935)). A woman at a tea party—let’s call her Elaine—claims that, without looking, she is able to distinguish between two scenarios about a given cup of tea:

1. the cup has been prepared by pouring milk first and then tea;
2. the cup has been prepared by pouring tea first, and then milk.

How might we decide whether Elaine actually has this ability? One option, which Fisher described in R. A. Fisher (1935), is to collect some data—testing Elaine’s ability to distinguish between (1) and (2)—and see how likely those data are under the assumption that Elaine does *not* have this ability. Fisher called an assumption of this type—the status quo, that no effect is present—the *null hypothesis*, denoted H_0 . Fisher describes the data collection as follows:

¹¹Can you derive what $sd(\bar{X})$ should be, assuming that the covariance between the i^{th} and j^{th} guess is $Cov(X_i, X_j) = \sigma_{i,j}$?

We will consider the problem of designing an experiment ... [to be] mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of what the test will consist, namely, that she will be asked to taste eight cups, that these shall be four of each kind. R. A. Fisher (1935)

The goal for Elaine is to correctly identify the four cups of each kind. If Elaine doesn't have this ability—that is, if H_0 is true—then we would expect her to correctly identify all four cups of each kind approximately 1.4% of the time. The full probability distribution¹² for $X = \#$ of cups correctly identified is given in Table 1.1.

Table 1.1: The probability distribution of $X = \#$ of cups correctly identified by Elaine. The possible values are 0, 1, ..., 4.

x	$P(X = x)$
0	$1/70 \approx 0.014$
1	$16/70 \approx 0.229$
2	$16/70 \approx 0.514$
3	$16/70 \approx 0.229$
4	$1/70 \approx 0.014$

We can use this probability distribution to decide whether a given dataset provides *evidence against* the null hypothesis as follows: if Elaine *does* have the ability to distinguish between (1) and (2), then we would expect her to correctly identify all of the cups. This result, $X = 4$, is rare under H_0 . So, if we observe $X = 4$, then we have evidence against H_0 . Conversely, if Elaine correctly identifies zero, one, two, or three of the cups, we don't have enough evidence against H_0 .

Broadly, the use of a probability model to make comparative judgements about data is what we mean by the likelihood pillar. Stigler (2016) writes that

In modern statistics we use a probability measure as at least part of the assessment of differences, often in the form of a statistical test, with roots going back centuries. The structure of a test is an apparently simple, straightforward question: Do the data in hand support or contradict a theory or hypothesis? The notion of likelihood is key to answering this question, and it is thus inextricably involved with the construction of a statistical test.

It is important to note that, in any interesting statistical test, the data in hand will never *strictly* contradict a hypothesis; instead, the data in hand might provide evidence against H_0 in the following way: we might act as if a hypothesis is false if, under that hypothesis, the data

¹²Can you calculate it?

in hand are improbable.¹³ So, in the tea example, we might act as if H_0 : *Elaine does **not** have the ability to distinguish between (1) and (2)* is false, if the data in hand are $X = 4$, because $X = 4$ is improbable under H_0 .

The concept of likelihood is ubiquitous in statistics, stretching far beyond hypothesis testing. As we will see in **?@sec-frequentist** and **?@sec-Bayesian**, likelihoods enter into both frequentist and Bayesian statistical methods, for example, estimating the rate of a disease in a given population. One point of contention between frequentist and Bayesian methods is the role that the likelihood ought to play!

1.2.2.4 Intercomparison

Consider a population where units are pages in this book. Suppose that we want to estimate μ , the average number of words per page in this book.¹⁴ From above, we know that μ is a feature of a population, called a population parameter.¹⁵ It would be tedious to count the number of words on each page to find the true average, μ (let's suppose we don't have software to do this for us!). But, perhaps we can choose a random sample of n pages, and count the number of words on each page in the sample. Then, we can infer something about μ by using information in the sample. Naturally, we could estimate our population μ using the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, where X_i is the number of words on the i^{th} page in the sample ($i = 1, \dots, n$). But importantly, that isn't the end of the story. \bar{X} for our sample won't be exactly equal to μ . And worse, if we had taken a different random sample of size n , the value of \bar{X} would have been different! So, over different samples, \bar{X} is random!

If we'd like to ask how good \bar{X} is at estimating μ ¹⁶—and we should ask this question!—then we should inquire about at least two things:¹⁷

1. Over many samples of size n , on average, what will \bar{X} be?
2. Over many samples of size n , how much variability will \bar{X} have (i.e., what is its variance)?

Some basic probability theory can help us answer these questions. If X_1, \dots, X_n is a random

¹³Does this reasoning sound strong? Some think it is not, as we will see in **?@sec-frequentist**.

¹⁴What is the variable of interest in this example?

¹⁵Other population parameters in this context might be p = the proportion of words per page under four letters in this book, or σ = the standard deviation of the length of words in this book.

¹⁶For a rigorous set of answers to this question, take a course in mathematical statistics!

¹⁷In addition, it might be nice to know things like (1) the shape of the distribution of \bar{X} , and (2) what happens to \bar{X} as $n \rightarrow \infty$.

sample of word counts from pages of this book, then, with respect to (a):

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n} \sum_i^n X_i\right) \\ &= \frac{1}{n} E\left(\sum_i^n X_i\right) = \frac{1}{n} \sum_i^n E(X_i) \\ &= \frac{1}{n} \sum_i^n \mu = \frac{1}{n} n\mu = \mu. \end{aligned}$$

This is important information: it tells us that, *on average* \bar{X} is correct! With respect to (b), we saw above (section 1.2.2.2, in the discussion of information), that the variance of \bar{X} is σ^2/n , where σ^2 is the population variance for each X_i . That is, σ^2 represents how much variability there is in the number of words per page in this book. So, now we know (a) what \bar{X} is on average, and (b) how much \bar{X} varies from sample to sample (if we want that variability in the original units, # of words per page, we can look at σ/\sqrt{n}). These facts provide some ingredients for assessing the *goodness* of \bar{X} as an estimator of μ , and we will return to a more comprehensive analysis of the goodness of estimators, and \bar{X} in particular, in [?@sec-frequentist](#).

But, there's a hidden problem here, which gets at the essence of what Stigler (2016) calls intercomparison: σ^2 is a population parameter, and we don't have a way of understanding the variability in \bar{X} without referring to an *external* quantity, σ^2 ; but in most cases, we won't know σ^2 . Is there a way to use *internal information* to estimate σ^2 , and thus, $Var(\bar{X})$? It turns out that we can estimate σ^2 internally using the *sample variance*:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

How does this substitution impact the accuracy of the analysis of the goodness of \bar{X} as an estimator of μ . The answer to that question depends on the context, and in particular, on the size of n . Stigler (2016) writes:

With large samples, statisticians would with no reluctance replace σ with $\sqrt{\frac{1}{n} \sum (X_i - \bar{X})^2}$ (or by Gauss' preference, $\sqrt{\frac{1}{n-1} \sum (X_i - \bar{X})^2}$) when its value was not otherwise available. Gosset's goal in the article [The Probable Error of a Mean] was to understand what allowance needed to be made for the inadequacy for this approximation when the sample was not large and these estimates of accuracy were themselves of limited accuracy.

When n is small, the students- t distribution allows statisticians to perform rigorous analyses of how good \bar{X} is as an estimator of μ , while using the substitution of S^2 in for σ^2 .

This result is an example of intercomparison, which Stigler (2016) defines as the ability to make statistical comparisons “strictly in terms of the interior variation of the data, without reference to or reliance upon exterior criteria [e.g., σ^2].” If estimating a population mean using

a sample mean was the only context in which intercomparison arose, then intercomparison it would not rise to the status of a “pillar” of statistical wisdom. In fact, the use of interior variation to estimate exterior variation arises in many areas of statistics, including regression, analysis of variance, and more advanced statistical models.

1.2.2.5 Regression

Regression is, at its core, about relationships between variables. Can we predict the sales of a product from the amount of money spent on advertising it? Do changes in meteorological conditions—e.g., temperature, windspeed, humidity—lead to systematic changes in atmospheric ozone concentration? What can we say about the relationship between the heights of parents and the heights of their children? Questions like these clearly require a framework that can model several (well, at least two) variables, at least some of which are measured with some uncertainty (“statistical noise”).

To get a sense of the fundamentals of linear regression, consider the `cars` dataset, which comes with the R statistical programming software.¹⁸ The data give some measurements of the speed of cars and the distances taken for those cars to stop. A priori, you might guess that the distance that it takes for a car to stop will increase as a function of the speed that the car was traveling. The plot in Figure 1.1 confirms this suspicion. But what is the relationship? More specifically,

Suppose that we increased speed by one mile per hour; how much, on average, would we need to increase our stopping distance by?

How could we predict stopping distance for a new speed? We can answer these questions with regression.

Given the plot in Figure 1.1, it might be reasonable to assume that there is an approximately linear relationship between speed (x) and distance (Y); that is

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

where β_0 is the intercept and β_1 the slope of the line relating speed and distance, and ε captures what we mean by “approximately linear”. More precisely, ε is a random variable centered around zero (i.e., mean zero), and models nonsystematic variability in the measurement process. That is, for each value of x_i , the value of Y_i is perturbed off of the true line $f(x; \beta_0, \beta_1) = \beta_0 + \beta_1 x$ (up or down) by a random draw from the random variable ε_i . Notice that f is given as a function of x , and the fixed, unknown parameters are specified after the semicolon.

If we knew the values of β_0 and β_1 , we could answer questions 1. and 2. above:

If we increased speed by one mile per hour, we would need to increase our stopping distance β_1 units, on average.

¹⁸<https://www.r-project.org>

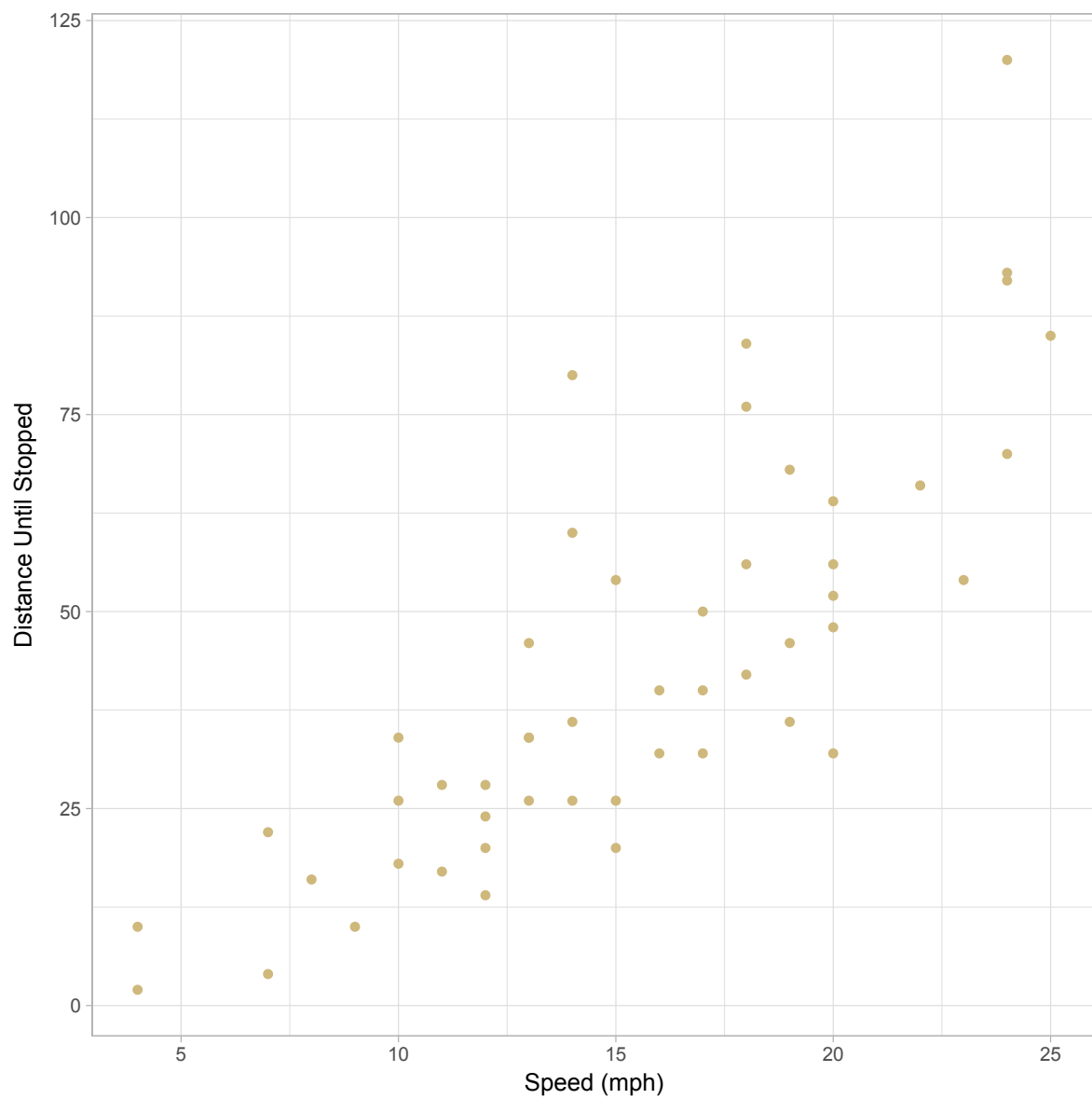


Figure 1.1: A plot of the speed of cars and the distances taken to stop.

To predict stopping distance for a new speed, x_0 , we could compute $f(x_0; \beta_0, \beta_1) = \beta_0 + \beta_1 x_0$. Unfortunately, these answers involve unknown quantities (parameters) β_0 and β_1 . An important component of regression is to *estimate* β_0 and β_1 based on the data. The *estimators* of β_0 and β_1 , call them $\hat{\beta}_0$ and $\hat{\beta}_1$, could then replace β_0 and β_1 in 1a. and 2a. above. Note that estimation can be done in the frequentist framework—through, for example, maximum likelihood estimation or ordinary least squares¹⁹—or in the Bayesian framework—through, for example, the maximum a posteriori estimate.

A careful reading of the questions posed in this section reveals a few important distinctions related to the goals of regression. For example, the first question in the first paragraph is about prediction—if we know the amount of money spent on advertising in a particular region, can we predict, to some degree of accuracy, sales? In constructing a regression model used for making a prediction, we are not necessarily concerned with whether that model is an accurate depiction of the world. Rather, we are concerned with whether it can tell us something useful about the *response variable*—sales in dollars—based on known measurements of the *predictor variable*—dollars spent on advertising.

By contrast, the second question in the first paragraph refers not to prediction, but to “systematic changes” in the response—atmospheric ozone concentration—based on changes in the predictors—temperature, windspeed, and humidity. Here, prediction might be an auxiliary goal, but language about systematic changes seems to suggest something more; in particular, we might want to *explain* the rise in atmospheric ozone concentration in terms of changes in meteorological conditions. The need for an explanation seems to point toward an accurate depiction of the world, meaning that our model should, in some sense, model the world (e.g., through a law of nature). Models that provide explanations often raise the issue of causation. Do the predictor variables *cause* the response? In what sense? What does it mean for X to cause Y , anyway? These questions that arise in the regression framework have a long and fascinating history in philosophy and the sciences, and we will explore some of them in **?@sec-causation**.

1.2.2.6 Design

No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or, ideally, one question, at a time. The writer [Fisher] is convinced that this view is wholly mistaken. Nature...will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed.—R.A. Fisher in (Stigler, 2016)

Depression is a tricky condition to treat, and there are several treatment options to choose from. Among them are medications, such as selective serotonin reuptake inhibitors (SSRIs)

¹⁹Which are equivalent when $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$.

and the newly approved Esketamine²⁰; and talk therapies, such as cognitive behavioral therapy (CBT) and emotionally focused therapy (EFT). Suppose that we are interested in learning which treatment works best for depression, as measured using the Beck’s Depression Instrument (Jackson-Koku, 2016). To simplify our example, consider just two medical treatments, the SSRI citalopram, and Esketamine; and one talk therapy treatment, CBT.

We can think of each treatment as a categorical variable, called a *factor*, with two *levels*: either the treatment has been given to a patient at the specified dosage and schedule, or it hasn’t. We might imagine that patients receiving citalopram will receive 40 mg, once per day; patients receiving Esketamine will receive 28 mg in the form of a nasal spray, twice per week.

One procedure for testing the effectiveness of treatments for depression might be to consider only one factor; that is, administer a treatment, and only that treatment, and measure its effect on depression. For example, we might administer 40 mg of citalopram once per day, for 6 weeks, to a group of n_1 people, and administer a placebo to a separate group of n_2 people; neither group receives Esketamine or CBT. Then, we could compare groups with respect to their average levels of depression. Such a procedure is called a *one factor at a time*, or OFAT, design, because it only varies one factor, while keeping all others constant.

An OFAT design is an intuitively plausible design for learning about an effective treatment, and has a long history. As reported in Stigler (2016), the Arabic medical scientist Avicenna, 1000 CE, comments on the importance of experimenting by changing only one factor at a time in his discussion of planned medical trials in his *Cannon of Medicine*. But as Fisher suggests in the quote above, “asking nature one question at a time” has disadvantages. For example, when compared with carefully designed experiments that vary more than one factor at a time, OFAT designs require more resources (such as more time and medication); are unable to estimate interactions between treatments (for example, whether Esketamine is only effective in conjunction with CBT); and, produce less precise estimates of the effects of each treatment (Czitrom, 1999).

Factorial designs are used as alternatives to OFAT designs. In factorial designs, we consider two or more factors, and allow factors to vary at the same time. To continue with our example above, imagine that we wanted to consider both citalopram and Esketamine. The administration of each would be a factor (and thus, we have a 2×2 factorial design). If a patient received 28 mg of Esketamine twice per week, we might assign them a variable $E = 1$; otherwise, we would assign $E = 0$. Similarly, if a patient receives 40 mg of citalopram, once per day, we might assign $C = 1$, and $C = 0$ otherwise. Importantly, in designing our experiment, it is desirable to have individuals with all combinations of E and C , i.e., $E = 1$ and $C = 1$; $E = 1$ and $C = 0$; $E = 0$ and $C = 1$; $E = 0$ and $C = 0$.²¹ Allowing all factors to vary, rather than just one, we are able to estimate interactions, for example, the extent to which taking both Esketamine and citalopram is better than taking either one alone. Of course, factorial designs

²⁰See Meisner (2019) for information about this new treatment for depression.

²¹As long as we have no reason to believe that this would be harmful or unethical.

exist for two-factor experiments with several levels—e.g., different doses of each drug—and for multi-factor experiments.

Factorial designs are an important example of the design pillar in statistics. Many other important principles in experimental design that help us decide whether an experimental treatment is effective are described in Fisher’s *Design of Experiments* (R. A. Fisher, 1935). Here are some examples:

Randomization. In a randomized experiment, units (e.g., individuals) are assigned to treatment groups (e.g., citalopram vs placebo) according to some random process (e.g., a coin flip). The use of randomization helps block the negative effect of confounding variables. For example, suppose that, in our depression study, subjects were *not* chosen by random, but instead by convenience: we assigned CBT to all University of Colorado Boulder students because they had easy access to talk therapy and CBT; all other individuals in the experiment were not given CBT. In such a case, the effectiveness of CBT is confounded (at least) by education level—it may be that University of Colorado Boulder students, or individuals with some college education respond better to CBT than the general population.

Blocking. Blocking is a technique for including a factor (or factors) in an experiment that lead to undesirable variation in the outcome. In a sense, we are able to control for that variation. In a *randomized block design*, units are first divided into blocks, and then, within each block, units are randomly assigned levels of the treatment. For example, in our depression study, we might group subjects by their education level—no HS diploma, HS diploma only, bachelor’s degree, master’s degree, terminal graduate degree (e.g., PhD)—and then, within each level, randomly assign CBT.

Replication. Replication is the repetition of an experiment on many different units. In the blocking example above, we might only recruit two subjects at each education level, and within each education level, randomly assign CBT or no CBT. Here, there would be no replication within blocks. However, to derive more reliable estimates of effects, we might recruit several subjects at each education level and randomly assign CBT or no CBT. If a treatment is actually effective, e.g., CBT does reduce depression, then aggregating over replications should reflect that fact; if a treatment is not effective, e.g., CBT does *not* reduce depression, then replication will guard against coincidences, such as a subject receiving CBT and a reduction in their depression by chance, or for some other reason.

1.2.2.7 Residual

We can learn by trying explanations and then seeing what remains to be explained.—Stephen Stigler (Stigler, 2016)

Consider again the `cars` dataset, discussed in the section on regression above. Recall that this dataset gives some measurements of the speed of cars and the distances taken for those cars to stop. We decided that there is an approximately linear relationship between speed (x) and

distance (Y): $Y = \beta_0 + \beta_1 x + \varepsilon$. After fitting the model—i.e., using measured (x, Y) pairs to estimate β_0 and β_1 —we might use the model to explain something about stopping distance, or predict stopping distance for a new speed not measured in the original dataset. But how do we know that the model fits well? Is the *assumed* linear relationship the *true* relationship between these variables?

Statisticians answer this question by analyzing the residuals of the model. To define the model residuals, and to understand why they are helpful in assessing fit, let's decompose the model into two components: a fixed, structural component, given by $f(x; \beta_0, \beta_1) = \beta_0 + \beta_1 x$, and a random component, given by ε . We assume that the measurement process is noisy, resulting in random normal errors: $\varepsilon \stackrel{iid}{\sim} N(0, \sigma^2)$. Suppose that we took our response variable Y , and subtracted from it the structural part of the model; we'd be left with the error term:

$$Y - f(x; \beta_0, \beta_1) = \varepsilon \tag{1.1}$$

So, if we could perform this operation, $Y - f(x; \beta_0, \beta_1)$, and if we could check that the result were normal, then we would have a sense of whether the model fit well or not; if the structure of the model has been specified correctly, then the distribution of $Y - f(x; \beta_0, \beta_1)$ should be normal, as assumed. But, recall that we do not know β_0 and β_1 , and estimate them from the data; the estimates are denoted $\hat{\beta}_0$ and $\hat{\beta}_1$. This estimation changes things. Instead of Equation 1.1, we now have

$$Y - f(x; \hat{\beta}_0, \hat{\beta}_1) = \hat{\varepsilon}, \tag{1.2}$$

which is the definition of the residual for this model. How does this help us with assessing fit? Well, we could think of $\hat{\varepsilon}$ as an estimate of the error term, ε , and thus, check the normality of $\hat{\varepsilon}$. *If* the model is specified correctly, then we should expect that $\hat{\varepsilon}$ will be approximately normally distributed. In Figure 1.2, we see a [qqplot](#) of the (standardized) residuals, which is one way of assessing normality. Notice that some points deviate from the line $y = x$, which suggests that the residuals deviate from normality. This suspicion is further corroborated by Figure 1.3, where a plot of the (standardized) residuals against fitted values, $\hat{Y} = f(x; \hat{\beta}_0, \hat{\beta}_1)$, shows some structure—a slight downward linear trend—rather than random scatter around $y = 0$.

Analyses of the residuals of a statistical model can be a powerful tool in assessing its fit. It can alert practitioners to issues with their given theory—as specified by a statistical model—and can suggest that a simpler or more complicated theory might better explain the phenomena in question.

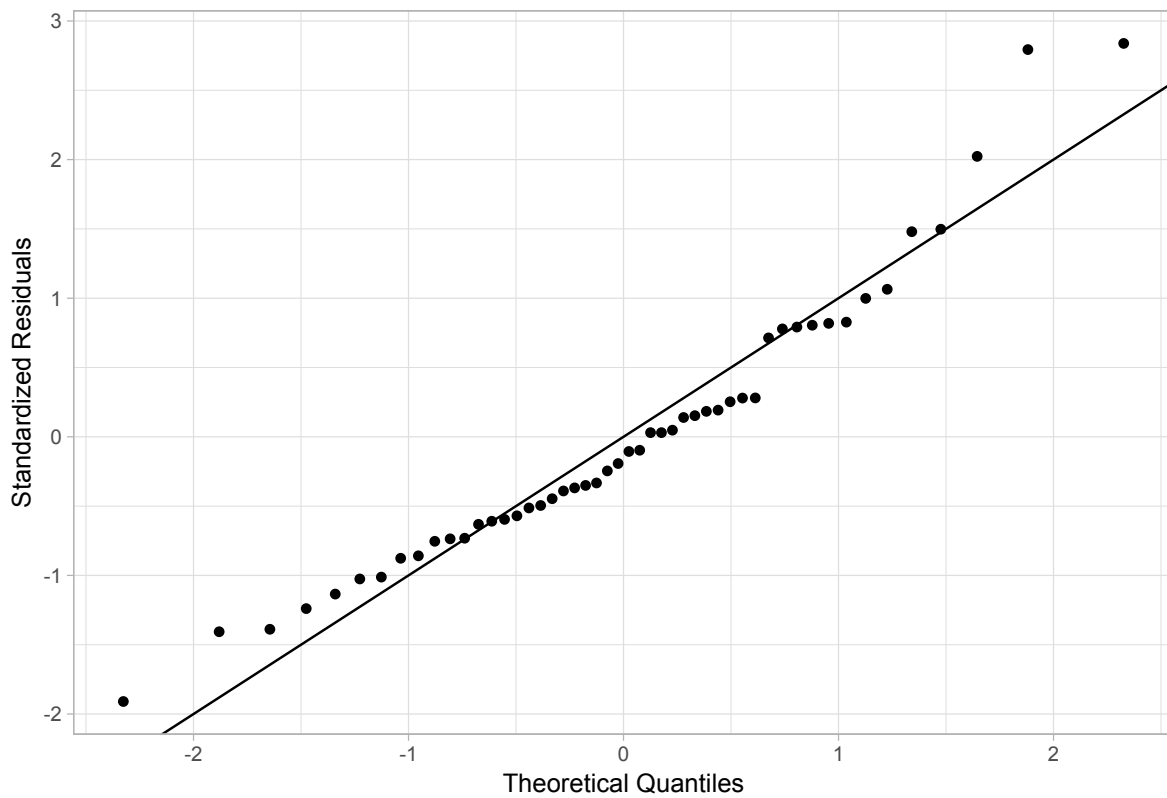


Figure 1.2: A qqplot of the (standardized) residuals from the linear model fit to the cars dataset. If the residuals are normal, we would expect to see them gather along the solid black line. In this qqplot, we see some deviations for small and large quantiles, suggesting some deviation from normality.

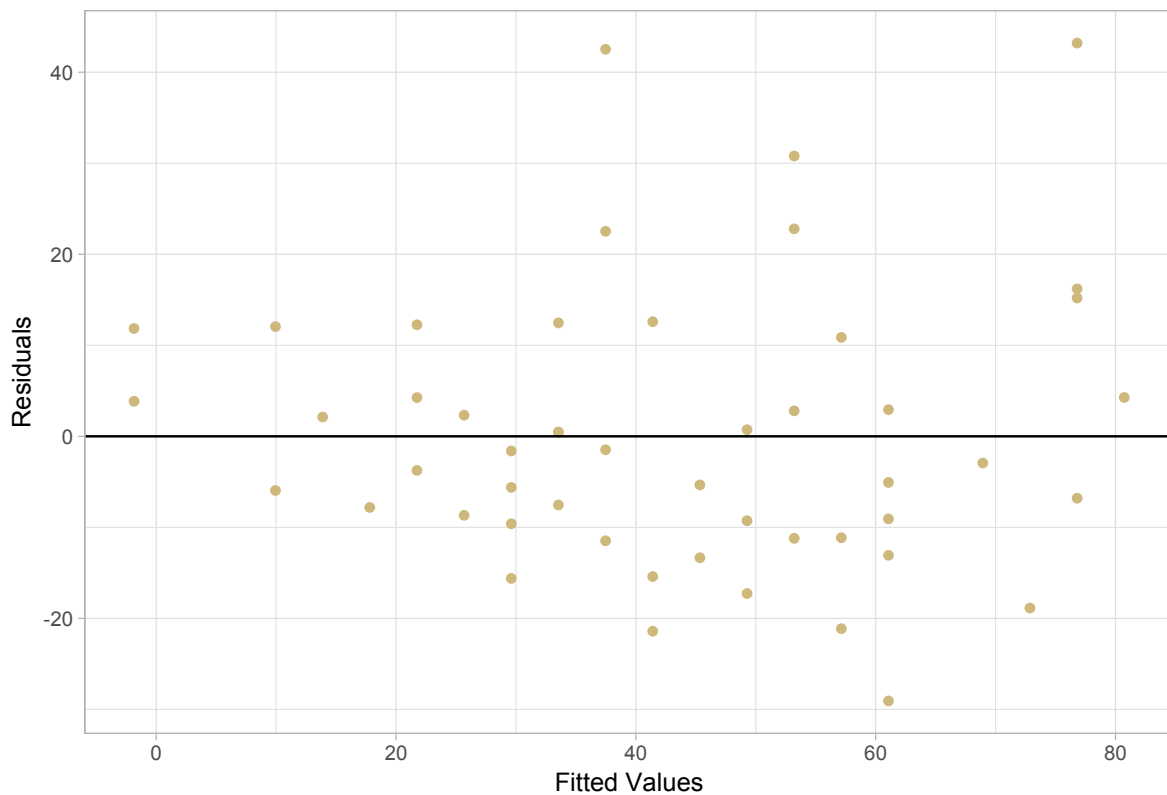


Figure 1.3: A plot of the residuals against fitted values, extracted from the linear model fit to the cars dataset. In this plot, if the model fits correctly, we would expect to see points scattered around the line $y = 0$, with many points close to $y = 0$, and few points far above or below. In this case, we see a slight downward trend in the points, suggesting that the model is not specified correctly. In addition, we also see higher variation in the residuals at larger fitted values.

1.3 What is the philosophy of statistics?

Now that we have a sense of some important features of philosophy and statistics as distinct disciplines, we are in the position to think about how they might be related. Broadly, there are two ways:

Philosophical issues in statistics. The use of statistics to solve real scientific problems requires, either implicitly or explicitly, certain philosophical commitments. Philosophers of statistics, and philosophically oriented statisticians, are interested in critically evaluating those commitments to decide whether they are justified. Many philosophical commitments receive attention in the practice of statistics and data science. For example, the inability to replicate many scientific results is often blamed on the inherent defectiveness of frequentist statistical methods, such as hypothesis testing (Ioannidis, 2005). To launch an effective critique of frequentist methods, one must often address the underlying philosophical and logical principles in play. Much of this book will deal with these sorts of issues, that is, philosophical issues that arise in statistics.

Statistical methodologies in philosophy. Many philosophers use statistical tools to attempt to solve important philosophical problems, such as the problem of induction (Chapter 2), scientific theory confirmation, and various problems in the philosophy of mind. Of course, attempts to utilize, for example, Bayesian tools to solve problems in scientific confirmation theory, may run into broad objections about the Bayesian tools themselves; so, the sorts of issues that arise in (1) are relevant here. We end this chapter by briefly considering an example from both (1) and (2).

1.3.1 Philosophy in statistics

The relationship between breast cancer and behaviors such as smoking and alcohol consumption has been studied extensively. In 2002, a report published in *Lancet* claimed that moderate drinking was not associated with a higher risk of breast cancer. With respect to smoking, the report found that premenopausal women who smoke had an increased risk of breast cancer, but that postmenopausal women had a significantly reduced risk of breast cancer (Band et al., 2002). Months later, in a report published in *The British Journal of Cancer*, a different group of researchers concluded that, *in women who reported drinking no alcohol*, smoking was not associated with breast cancer, and go on to conclude that “smoking has little or no independent effect on the risk of developing breast cancer” (Hamajima et al., 2002).

Both reports used *observational*, rather than *experimental*, data. In an observational study, researchers do not manipulate any variables or impose any treatments.²² In particular, both reports mentioned above made use of a type of observational study called a *case-control* study. Studies of this sort identify the *case*, i.e., a group known to have an outcome. In these studies above, groups of women with breast cancer constituted the case. Then, *controls* are identified,

²²Reasons for not controlling for variables or imposing treatments may be logistical—i.e., it would be costly, or impossible—or ethical.

i.e., a group known to be free of the outcome. Many variables are measured within each group. The goal of a case-control study is to look back in time to determine associations between the outcome and other variables (e.g., breast cancer and smoking) (Lewallen & Courtright, 1998).

In the Band et al. (2002) study, a questionnaire was sent to 1431 women under 75 years old with breast cancer; these women were listed on the population-based British Columbia cancer registry between June 1, 1988, and June 30, 1989. Questionnaires were also sent to 1502 age-matched controls, randomly selected from the 1989 British Columbia voters list. A subset of 318 and 340, respectively, replied. Researchers assessed the effects of alcohol consumption and smoking (separately for premenopausal and postmenopausal women), and adjusted for confounding variables (Band et al., 2002). The Hamajima et al. (2002) study is a *meta-analysis*, which combined data from many studies of the type conducted in Band et al. (2002).

The results from the two reports are, at least on their surface, in tension (if not, outright in contradiction) with one another: one suggests that smoking is a risk factor for breast cancer; another suggests that smoking is not a risk factor if we “control” for alcohol consumption (e.g., there may be an interaction between alcohol consumption and smoking). One practical implication of this tension is that, if one were to attempt to make behavioral changes based on these studies, it’s not clear what behaviors ought to be adopted. The correct adoption of a particular behavior depends on, among other factors, the reliability of the statistical analyses used, and there are a number of conceptual issues that bear on the reliability of these analyses. Many of these conceptual issues, while related to empirical content, are not empirical in and of themselves, and thus, I count them as philosophical. Some important philosophical issues that arise are:

How does using a meta-analysis strengthen the inductive support of the conclusions being drawn? It is often thought that combining several studies together into a meta-analysis can “create a single, more precise estimate of an effect” (Ferrer, 1998; Hoffman, 2015). A correctly performed meta-analysis that creates a more precise estimate of an effect would increase the inductive support of the conclusion being drawn; but in practice, few meta-analyses meet all the criteria for correctness, and thus, the inductive support provided by meta-analyses can be weak (Hoffman, 2015; Ioannidis, 2010). Assessing the strength that a meta-analysis brings to a statistical argument is logical, and thus, philosophical, in nature.

How does each study avoid, or fail to avoid, data dredging? Data dredging is a set of fallacious procedures that result in claimed associations when, in fact, no associations exist. One popular type of data dredging is post hoc multiple comparisons, which arises when many claims are tested simultaneously, after the data have been collected. When a large number of claims are tested without adjustments being made to the testing procedures, the large majority of findings will be inadequately supported, i.e., they will be false positives (Smith, 2002). But there is no universally agreed upon method for adjusting testing procedures for multiple comparisons. In choosing a particular method, one is advancing (either explicitly or implicitly) a set of *values*, e.g., conservatism about avoiding a particular type of error. We will revisit this issue in ?@sec-frequentist.

Does the fact that only a subset of chosen subjects respond to a questionnaire impact the conclusions being drawn? Even if the original group sent the questionnaire was randomly chosen, the subset of actual respondents is likely not a random sample from the desired population. If questionnaire response is correlated with a confounding variable, conclusions drawn will be weakly supported.

Even if the associations discovered are real, what can we conclude about causal relationships? The strength of support lent to causal conclusions based on analyses of observational studies is disputed. Some argue that “case-control studies may prove an association but they do not demonstrate causation” (Lewallen & Courtright, 1998). Others argue that causal conclusions *can* be drawn from case-control studies and, more broadly, observational studies (Persson & Waernbaum, 2013). Further, among those who believe that observational studies can support causal conclusions, there is disagreement as to which methods provide the strongest inductive argument (Gelman, 2009; Pearl, 2009)

1.3.2 Statistics in philosophy

There are several areas of philosophy that make use of statistical methodology in advancing solutions to philosophical problems. One example is in scientific confirmation theory. Generally, given a scientific theory T , scientists use empirical evidence to attempt to confirm or refute T . As a simple example, consider the ‘scientific theory’ T : *All ravens are black*. How might one confirm or refute T ? Immediately, we notice that there is an asymmetry; to refute T , one only needs to observe a single non-black raven. However, to conclusively confirm T , one needs to show that *all ravens, even those yet to be observed* are black. That is a much harder task. But, suppose that many, many ravens have been observed, and all of them have been black. Does this add some confirmatory support to T ? Intuitively, it does, and Bayesian confirmation theorists have made attempts to formalize this intuition by quantifying the degree to which new observations consistent with a theory T actually confirm T .

Let’s consider one simple attempt at a Bayesian confirmation theory. Let x be a new observation; some have proposed that a theory T is confirmed by x just in case the probability of the theory given the new observation is greater than the probability of the theory without the observation (Mayo, 2018):

$$P(T|x) > P(T). \quad (1.3)$$

In Equation 1.3, $P(T)$ is the *prior probability* that the theory is true, and $P(T|x)$ is the posterior probability that the theory is true, given the observed evidence, x . The posterior probability can (at least in theory!) be computed using Bayes’ theorem:

$$P(T|x) = \frac{P(x|T)P(T)}{P(x)}. \quad (1.4)$$

This view of confirmation theory raises many questions. Ostensibly, theories are either true or false, i.e., they are assigned uninteresting probabilities: either zero or one. So, does it make sense to assign non-zero and non-unit probabilities to theories? What could that probability mean? Further, what does it mean to assign a prior probability to a theory, i.e., $P(T)$? If we have no evidence bearing on that theory, then what probability should we assign to it (we need *some* prior to use Bayes' theorem!)? Finally, as Mayo (2018) suggests, Equation 1.3, while intuitively plausible, has its problems and rival proposals. For example, we might say that T is confirmed by x just in case the probability of the theory given the new observation is high in some absolute sense, at least greater than the negation of that theory given the new observation:

$$P(T|x) > P(\neg T|x). \quad (1.5)$$

Equation 1.3 and Equation 1.5 provide different accounts of theory confirmation. How can we decide between the two? Formal epistemologists use statistical (especially Bayesian) tools to work on these issues.

The goal of this chapter has been to provide a shared framework to think through important issues in the philosophy of statistics. We saw that philosophy is rooted in a shared commitment to providing reasons for particular views about the world, and has a close historical connection to the sciences. Philosophers often care about empirical content, but often, the arguments that they advance depend on concepts (e.g., values, metaphysical commitments) that go beyond empirical content. We also saw that (inferential) statistics can be thought of as a set of inductive methods used to draw general conclusions about the world from limited information. In remaining chapters, we will compare, contrast, and explore the inductive strength of particular statistical methodologies.

We continue in the next chapter by expanding upon the inductive nature of statistics. What is induction, and what forms can it take? What are some general principles that make statistical methodologies strong, in the inductive sense? Do any of the competing statistical methodologies provide solution to the longstanding philosophical problem of induction?

1.4 Discussion Questions

What is a reasonable working definition of philosophy? Of statistics?

Describe some ways in which academic philosophy differs from “personal philosophies”.

What are some important issues that arise in the philosophical study of logic? Metaphysics? Epistemology? Ethics? Philosophy of Statistics?

What is the verifiability criterion of meaning? What are some problems with this criterion? What bearing does this have on metaphysics as a discipline?

In the discussion of hypothesis testing in Section [1.2.2.3](#), we reasoned as follows: we might act as if a hypothesis is false if, under that hypothesis, the data in hand are improbable. Is this strong reasoning? Can we think of an example in which it is not?

What is the relationship between philosophy and science?

In what sense do the “pillars of statistical wisdom” provide a definition of statistics?

What is the relationship between philosophy and statistics?

Fisher writes, “Nature...will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed.” What does he mean by “asking nature a single question”, and how might doing so not be optimal?

What is the difference between an observational study and an experiment? For what reasons might we prefer the former?

Describe some interesting issues that arise in Bayesian confirmation theory. For example, Bayesians assign probability values to theories. Is that coherent?

Which “confirmation theory” given in Section [1.3.2](#) do you prefer and why?

2 Contextualizing statistics

The general body of researches in mathematical statistics during the last fifteen years is fundamentally a reconstruction of logical rather than mathematical ideas, although the solution of mathematical problems has contributed essentially to this reconstruction.

— R.A. Fisher, *The Logic of Inductive Inference*

In Chapter 1, we saw that inductive arguments are such that, even if the premises are true, the conclusions may be false. For example, it might be true that

P: up to the current time, t , all observed ravens have been black

and false that

C: All ravens, including those yet to be observed, are black.

Arguably, most scientific and statistical arguments are inductive in this way: the available data, and modeling assumptions (encoded in premises) do not guarantee the veracity of the inferred scientific theory or statistical hypothesis (the argument’s conclusion). Most inferences to theories or hypotheses *go beyond* the observations at hand. Scientific laws are sufficiently general, in the sense that they refer not to particular entities or cases, but broad categories. For example, Hubble’s Law of Cosmic Expansion states that $V = h \times d$, where V is galaxy’s recessional velocity, h is a parameter representing the rate of universe expansion, and d is the galaxy’s distance from a reference galaxy. Hubble’s Law is not only about the relationship between velocity and distance for galaxies that have been observed, but about the relationship between distance and velocity for *all*, including yet-to-be-observed galaxies. Further, the constant, h , is strictly speaking, an *unobservable*; it represents “the constant rate of cosmic expansion caused by the stretching of space-time itself” Bagla (2009). h can only be *inferred* through scientific or statistical methods rather than directly observed.

Inferences to broad generalizations or unobservable entities aren’t particular to the physical sciences. In the social sciences, psychologists are often interested in measuring unobservable psychological traits, called *latent variables*, such as general intelligence, g , self-esteem, or extroversion. To “measure” latent variables, psychologists must first measure observable variables—e.g., responses to a questionnaire—and have a statistical model describing how the latent variables relate to what was measured.

In this chapter, we study statistical inference as a form of inductive inference. What forms can inductive inference take? What problems arise in attempting to justify inductive inference? How do statistical models contribute to the proper foundation for inductive inference and, by extension, scientific knowledge? How strong are the arguments that justify statistical methodologies? By expanding upon induction and these related questions, we gain a broader and contextualized view of the nature of statistical inference. From there, we will be in the position to begin to evaluate different statistical methodologies.

2.1 Types of inductive inference

To better understand inductive inferences, it may be helpful to study different types of inductive inference. Here, we will study three types: inference to the best explanation, induction by enumeration, and inference from analogy. For more information on types of inductive inference, see Vickers (2006).

2.1.1 Inference to the best explanation

Today, Estelle woke up late. She was in a rush to get ready, and quickly grabbed her phone off of the charger on her way out of the house. Soon after, on her way to work, she noticed that her phone battery was only at 20 percent. Oh no! What could be the explanation for why her phone was not charged to (or near) 100 percent? There are many *logically possible* explanations. Here are a few:

H_1 Estelle plugged her phone in properly the night before, but, unbeknownst to her, the power went out for a long period of time, and as a result, her phone did not charge.

H_2 Estelle plugged her phone in properly the night before, but the phone charging cord is faulty and no longer working, and as a result, her phone did not charge.

H_3 Estelle plugged her phone in properly the night before, but a being from another planet visited her room and unplugged it for most of the night. As a result, her phone did not charge.

H_4 Estelle, in fact, didn't plug her phone in properly the night before, and as a result, her phone did not charge.

Our intuition says that some of these explanations are plausible, and others are not. For example, in the absence of additional information, H_1 , H_2 , and H_4 seem plausible. H_3 seems implausible because we have no good reasons to believe that such beings exist or can travel to Earth, and even if they did and could, we have no reason to believe that they have the goal of unplugging our phones.

Now, suppose that Estelle thinks a bit more, and remembers two things: First, she remembers that the digital clock on her stove displayed the correct time on the way out of the house. Second, she remembers that a few other times in the last month, she’s plugged in her phone improperly, and once she secured the connection, her phone charged without issue. This information changes which explanations are plausible. In particular, H_1 now becomes much less plausible, and H_4 becomes much more plausible. In fact, we might infer that H_4 is *the best explanation* for the fact that the phone is only charged to 20 percent, based on the information at hand.

The reasoning employed in this example is a type of inductive inference¹ called *inference to the best explanation* (IBE). Generally, IBE might be characterized as the process of “accepting a hypothesis on the grounds that [it] provides [a] better explanation of the given evidence comparing to the other competing hypotheses” (Erdenk, 2015). Notice that IBE is clearly not deductive, because there is no requirement that, with limited information, the best explanation is logically entailed by the observed phenomena. In the example above, H_2 has not been eliminated on the basis of logical impossibility; rather, it just seems less plausible than H_4 .

In science, we often use statistical models to provide explanations for the phenomena that generated the data. Statistical models can help construct such explanations. In many cases, there will be several candidate models for a particular set of data. For example, we might like to explain atmospheric ozone concentration based on certain known conditions, such as temperature, windspeed, humidity, and concentration of certain pollutants, such as sulfur dioxide. Many plausible models could be constructed with respect to these data—some models might include possible pollutants as explanatorily relevant to the variation in atmospheric ozone concentration, while other models might exclude (some of) these pollutants. Statisticians have come up with processes to select a “best” model among the candidates. Some criteria that measure “best”—for example Bayes’ Information Criterion (BIC)—might be thought of as a formalization of inference to the best explanation. That is, among several explanations (models) of the regularities in the data, BIC selects a “best” explanation by balancing goodness of fit with simplicity Faraway (2015).²

2.1.2 Induction by enumeration

What justifies our knowledge that all electrons have a mass of $9.1 \times 10^{-31}\text{g}$? Or that a hot stove will burn my hand? Or that there will be a full moon on January 18, 2030? The argument for such knowledge is often of the form (Norton, 2002):

¹Note that some philosophers do not classify IBE as a type of induction (or deduction); such philosophers carve up the space of non-deductive arguments differently than we have here, to leave space for IBE as its own type of inference. See Chapter 2 of Okasha (2016) for more details.

²Arguably, using BIC for *explanation* rather than *prediction* would require that we know something about the extent to which each input variable in the statistical model is causally related to the output variable. BIC does not, on its own, select for causal relationships, and such relationships are typically what is desired in an explanation.

P: All *observed* instances of A have had property p .

C: Therefore, *all* (including unobserved) instances of A will have property p .³

This type of argument—often called *induction by enumeration*, or *enumerative induction*—allows us to generalize from observed regularities to unobserved regularities, and as such, is indispensable to science. Often, induction by enumeration is the only justification that we have for a particular scientific fact, as is the case for the mass of electrons (Norton, 2002). In other cases, such as those that predict the phases of the moon, physical theories describe the necessary causes that produce the effect that A has property p , and we don’t necessarily need to rely on induction by enumeration directly. But the justification for the physical theories themselves seems to rely on induction by enumeration: how do we know that the laws of planetary motion will hold on January 18, 2030, so that our predictive model will be accurate? We know this because all observed phenomena in the universe (A) have had the property of obeying the laws of planetary motion (p), and infer that *all* phenomena—including future phenomena—in the universe will obey the laws of planetary motion. That is, we know they will hold because of induction by enumeration!

2.1.3 Inference from analogy

A 1978 study of the artificial sweetener saccharin concluded that “saccharin is carcinogenic for the urinary bladder in rats and mice, and most likely is carcinogenic in human beings” Reuber (1978). How might we reason from the premise that saccharin is carcinogenic in rats to the conclusion that it is (likely) carcinogenic in humans? We might argue something like the following:

P1: Humans, on the one hand, and rats and mice on the other, share many anatomical, physiological, and genetic properties.⁴

P2: Many of these shared properties are relevant to the development of different types of cancer.

P3: Saccharin has been shown to be carcinogenic in rats and mice.

C: Therefore, cancer is (likely) carcinogenic in humans.

This argument might be strengthened by another premise that claims that often in the past, when a result has been demonstrated in rats, it has also been demonstrated in humans (“Animal Research at the ICR,” 2019). We might interpret such an argument form as an *argument from analogy*. The general form of an argument from analogy might look something like this:

P1’: A and B share properties p_1, \dots, p_n .

³A more modest version of the conclusion of enumerative induction is (C) Therefore, *the next* unobserved instance of A will have property p .

⁴See Bryda (2013) for evidence of the claim that there are such similarities.

P2': A has property p ($p \neq p_i, i = 1, \dots, n$).

C': Therefore, B has property p .

Such an argument is (almost) always categorized as inductive, because it is (almost) never logically inconsistent for B to not have property p . And in fact, to the best of our knowledge as of this writing, **C** is believed to be false; there is “no consistent evidence that saccharin is associated with bladder cancer incidence” (“Artificial Sweeteners and Cancer,” 2016).

Arguments by analogy are often used in science and statistics, as suggested by the saccharin case above. For another example, in *Origin of Species*, Darwin draws analogy between domestic selection by breeders and selective process that arises in nature to argue for natural selection as a key mechanism for evolution Norton (2018).

2.2 The problem of induction

Common to all types of inductive inference is the fact that the inferences from premises to conclusions are risky: even if the premises are true, the conclusion does not necessarily follow. Consider the following inductive inference:

P: In a sample of $n = 100$ University of Colorado Boulder students, 85 students claimed to have some amount of student loan debt.

C^\dagger : Therefore, 85% of all University of Colorado Boulder students have some amount of student loan debt.

How can we *justify* this inference from **P** to C^\dagger ? More generally, what makes inductive inference a reliable form of inference? Can we come up with an argument for the conclusion that **C**: *inductive inferences are justified*? Intuitively, we believe that inductive inference *is* a reliable form of inference, for example, when we believe that the key to our home will work today because it worked yesterday. Many of the conclusions that we draw, including scientific conclusions supported by statistical arguments, rely on inductive inference. However, as philosopher David Hume argued, there is no strong argument for the conclusion that **C**: *inductive inferences are justified*. That is, there is no rigorous justification for inductive inference. This fact is called *the problem of induction*. Let's briefly work through Hume's argument that leads to the problem of induction.⁵

To gain some insights into Hume's argument, let's first consider the ways in which the conclusion of an inductive inference, e.g., C^\dagger , might be wrong. With respect to C^\dagger , it might be the case that the chosen sample is biased in some way; if the sample is biased, then it may be the case that students with student debt had a higher chance (or lower chance) of being chosen for the sample. In that case, we might attempt to take a truly random sample, where every student had the same chance of being chosen. In that case, we could modify our argument:

⁵My explanation of Hume's argument relies on Henderson (2018).

P^\dagger In a *random* sample of $n = 100$ University of Colorado Boulder students, 85 students claimed to have some amount of student loan debt.

C^\dagger Therefore, 85% of all University of Colorado Boulder students have some amount of student loan debt.

This modification does not solve the issue; still, C^\dagger can be false, while P^\dagger true. Even with a large random sample, it is possible that we are unlucky in the sense that the sample percentage differs greatly from the population percentage. A second issue with our argument is that, in inferring from a sample of University of Colorado Boulder students to the population of all University of Colorado Boulder students, we are making some assumptions about the uniformity of nature across time and space. For example, in choosing a random sample, we are assuming that:

- the parameter *percent of University of Colorado Boulder students who have some amount of student debt* stays constant, at least across short periods of time; and
- students that we have not observed are similar in the relevant ways (e.g., with respect to finances and student debt) to students that we have observed.

Taken together, philosophers call generalized versions of these assumptions, the “Uniformity Principle” (UP). The UP states that there is a kind of stability to the world; the parameters that we are attempting to estimate and the laws of nature and regularities that are associated with those parameters stay stable, or themselves change in predictable, lawlike ways. The UP plays a critical role in Hume’s claim that there is no strong justification for inductive inference. First, Hume claims that the UP appears to be assumed in any inductive inference. This claim seems quite plausible: any time that we infer a conclusion based on one of the argument types from section 1.1—e.g., that all observed electrons have mass of $9.1 \times 10^{-31}\text{g}$, therefore all electrons (observed and unobserved) will have this mass—we are implicitly assuming the UP. So, inductive inference cannot be truly justified without some justification for the UP. And in fact, the UP seems like the crucial premise in need of justifying.

Once Hume has established the centrality of the UP, he then notes that any justification of the UP must either be deductive or inductive. That is, the UP will either follow necessarily from the premises (deductive); or it will be possible for the premises to be true but for the UP to be false (inductive). As Hume argues, the UP cannot be justified deductively, because its negation does not imply a contradiction; there is nothing incoherent about a universe that isn’t uniform across space or time. So, the deductive route will not work. But further, the UP cannot be justified inductively, because any inductive argument justifying the UP would *assume* the UP itself, and therefore be circular. Thus, according to Hume, our hopes of justifying inductive inference are hopeless: we have failed to justify the UP, which was a necessary condition for justifying inductive inference.

2.3 The problem of induction and statistical philosophies

Hume’s problem of induction is well-known among philosophers, and especially philosophers of science. To be sure, it is a philosophical problem *about* science and statistics; contemporary practicing scientists and statisticians do not often engage directly with this problem. Some might even claim that worries about the justifications for the UP and inductive inference are just philosophical quibbling: the lack of a bedrock justification for induction and the UP, they argue, are not real problems for science, at least in not in practice. But there are good reasons to engage with these issues. First, many scientific and statistical methods were developed as a response to known problems with inductive inference, including Hume’s problem of induction. The developers of these methods often had the explicit goal of making inductive inferences stronger. The frequentist statistician Ronald Fisher (1890 – 1962) contextualized his work as a kind of inductive logic in various places, including in papers titled “Statistical Methods and Scientific Induction” (Fisher, 1955) and “The Logic of Inductive Inference” (Ronald A. Fisher, 1935). The first Bayesian analyses—including the work of Reverend Thomas Bayes (1702 – 1761) and Pierre-Simon Laplace’s (1749 – 1827)—were also developed to solve Hume’s problem of induction (Clayton, 2021; Stigler, 2018). Engaging with the problem of induction, along with proposed solutions, may help us gain a deeper understanding of the origins, justifications, and utility of standard statistical methods. In turn, we may then be in a better position to critique and apply them correctly.

2.3.1 The falsification solution

Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

— R.A. Fisher, *Design of Experiments*

Philosopher of science Karl Popper recognized that Hume’s problem of induction was, in a certain sense, insurmountable. Popper writes:

Hume, I felt, was perfectly right in pointing out that induction cannot be logically justified. He held that there can be no valid logical arguments allowing us to establish ‘that those instances, of which we have had no experience, resemble those, of which we have had experience’. Consequently ‘even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience’ (Popper, 2010 [1963]).

As a result, Popper made no attempt to solve the problem of induction by *justifying* induction. Rather, he denied that induction was necessary for the proper functioning of science. Instead of generalizing from observations to theories (e.g., scientific laws), Popper believed that science properly functions by first posing the theories, and then testing those theories against particular

relevant data. In this way, the proper justificatory structure of science is *deductive* rather than *inductive*: a scientific theory T , so Popper claimed, can be conclusively falsified given certain empirical evidence. As an example, consider the theory T : *All ravens are black*. This theory can be conclusively and deductively falsified with the observation of (at least) one non-black raven. The argument would be:

P1: If T : *all ravens are black*, then any raven observed will be black.

P2: A white raven was observed.

C: Therefore, T is false.

This general argument form,

P1: If T , then e .

P2: Not e .

C: Therefore, not T .

is valid, and therefore, deductive. For Popper, *falsification*—the process of proposing theories and attempting to refute them—rather than induction, is the real mode of scientific progress.

To be sure, this view has some problems. For one, we might notice that there is an asymmetry between our ability to (1) reject T as false, i.e., when evidence e contradicts T ; and (2) accept T as true, i.e., when e does not contradict T . In case (1), practically speaking, most scientific theories and hypotheses are not as easily and clearly falsifiable as T . Consider the health science hypothesis H : *A high carbohydrate diet causes an increase in body weight*. What evidence would *conclusively* falsify H ? Perhaps, in theory, such evidence exists. We can *imagine* a world in which any time someone increases their carbohydrate intake for several weeks, they also increase their body weight. In such a world, we might argue:

P1: If H : *a high carbohydrate diet causes an increase in body weight*, then any individual observed eating a high carbohydrate diet will see an increase in body weight.

P2: Thom eats a high carbohydrate diet but has *not* seen an increase in body weight.

C: Therefore, H is false.

However, our imagined world is not the real world; in the real world, diet is complicated. There are many other factors that also influence body weight. Strict falsification is much more difficult to attain. Statistical methods attempt to control for these other factors—as well as random variation—to isolate the effect of diet on body weight. But even then, do we know we controlled for all of the right factors? How do we know that we did not leave something out, or that random variation, rather than diet, led us to reject H ? Conclusive falsification seems, in practice at least, unattainable. It is not clear exactly what evidence we could practically attain that would allow us to conclusively falsify most real-world scientific hypotheses. As we will see in [?@sec-frequentist](#), [?@sec-Bayesian](#), and [?@sec-causation](#), statistical philosophies,

including causal inference, can help us make inferences and practical decisions in the absence of conclusive falsification.

In case (2), e being broadly consistent with T does not confirm T , because e will be consistent with other (in fact, infinitely many other) theories, T_i , each of which is not equivalent to T . Yet another observed black raven does not confirm T , and there are many other theories consistent with new observation (e.g., T_1 : *ravens are 60% black and 40% white*). Popper’s solution to this problem is to introduce the notion of corroboration. A theory T is *corroborated* by e if e were produced by a “severe test”. By a “severe test”, Popper means “tests that would probably have falsified a claim if false” (Mayo, 2018). Note that corroboration is not strict confirmation, if by ‘confirmation’ one means *conclusively true*. Instead, corroboration is a building up of support for T , through the right kinds of probes.

If one is familiar with the statistical hypothesis testing developed by Ronald Fisher, Jerzy Neyman (1894 – 1981), and Egon Pearson (1895 – 1980), Popper’s logic of conjecture and refutation should not be entirely foreign.⁶ In statistical hypothesis testing, and in Popper’s falsification paradigm, a hypothesis is put forward, and a statistical procedure is conducted to attempt to falsify it. Interestingly, classical frequentist hypothesis testing starts and ends with conjecture and refutation; there was no formal method for corroboration or so-called severe testing. More recently, philosophers Deborah Mayo and Aris Spanos developed a set of statistical tools that formalize the notion of a severe test, which, when used correctly, can help corroborate hypotheses (Mayo, 2018; Mayo & Spanos, 2011). In **?@sec-frequentist**, we will study the statistical procedures that Fisher, Neyman, Pearson, Mayo, and Spanos have developed to deal with messy, real-world scientific theories and hypotheses.

So, do falsification and hypothesis testing succeed in solving the problem of induction? We will not be able to adequately address this question until **?@sec-frequentist**. But, as we will see, under the statistical assumptions posed in a statistical model, hypothesis testing provides a framework for quantifying uncertainty in our conclusions and behaviors by controlling error rates over the long run. This error control represents an important step forward in strengthening inductive inference: if the modeling assumptions are (roughly) met, we know how often we will be in error in the long run. While this paradigm does make explicit and precise statements about probabilities, it still assumes the UP—i.e., by making claims about “the long run”. But, as we saw in section 1.2, the UP cannot be justified without circularity. So, in failing to avoid the UP, strictly speaking, these statistical method has failed to circumvent the problem of induction. Nevertheless, these methods provide some guidance for belief and action under uncertainty.

⁶Hypothesis testing was developed separately by Fisher, on the one hand, and Neyman and Pearson on the other. The version often taught is a blend of these two methods.

2.3.2 The Bayesian solution

The most popular alternative to Popper’s falsificationist framework—and falsificationist statistical methods like hypothesis testing—is called *probabilism*. Probabilism is the view that conclusions, theories, and hypotheses can be assigned a degree of support through the use of probability theory (Mayo, 2018). Probabilism assigns theories a number between zero and one, which represents, roughly, how plausible the theory is. Perhaps we have the following argument:

Over one million ravens have been observed, and all have been black.

Therefore, *all ravens are black*. Intuitively, T has strong support. Probabilism might assign T a number close to one. It is *possible* that $\neg T$: *some ravens are not black*. But given the lack of evidence, $\neg T$ would be assigned a low number, close to zero.

Various attempts have been made to formalize probabilism as a theory of inductive logic (Bayes & Price, 1763; Carnap, 1962; Cox, 1946). The most famous form of probabilism, with the closest connection to statistical practice, is Bayesian probabilism. Bayesian probabilism makes use of Bayes’ theorem to update probabilities assigned to theories based on observed evidence. For example, suppose that we start out by assigning H : *a high carbohydrate diet causes an increase in body weight* a probability of 0.3. Nutrition researchers studying this hypothesis conduct a study and find that e : *on average, participants on a high carbohydrate diet gained 3 pounds more than those on a low carbohydrate diet*. Suppose that the probability of observing e if H were true is $P(e | H) = 0.8$, and the probability of E if H were false is $P(e | \neg H) = 0.4$. Bayes’ theorem states that⁷

$$\begin{aligned} P(H | e) &= \frac{P(e | H)P(H)}{P(e | H)P(H) + P(e | \neg H)P(\neg H)} \\ &= \frac{(0.8)(0.3)}{(0.8)(0.3) + (0.4)(0.7)} \approx 0.46. \end{aligned}$$

Thus, the updated probability of H , based on observing e , is higher. Probabilism aids inductive logic, in the sense that it provides a number that quantifies the strength of the conclusion (i.e., the theory or hypothesis) given the premises (evidence and assumptions).

Probabilism also has its problems. Many prominent philosophers and statisticians—Popper and Fisher among them—are vehemently opposed to the use of probability to confirm hypotheses. Popper argued that the degree of confirmation that e confers on H is not the same as the probability of H given e (Mayo, 2018; Popper, [1959] 2005). Chapman (2015) argues that, contrary to the starting point of probabilism, *probability* is not equipped to extend deductive logic to reasoning about *plausibility* (i.e., uncertain reasoning). Fisher wrote that “probability is a ratio of frequencies, and about the frequencies of such [hypotheses] we can know nothing

⁷If this formulation of Bayes’ theorem does not look familiar to you, do not worry. We will discuss Bayes’ theorem in detail in ?@sec-Bayesian.

whatever” Fisher (1922).⁸ A primary problem for these thinkers is that probabilism relies on an epistemic interpretation of probability. Such an interpretation allows for probabilities to be assigned to fixed, non-repeatable features of the world. It’s not clear how such probability assignments arise. How did we come up with $P(H) = 0.3$? It is not tied to any repeatable process. It seems like we just made it up! For those that reject Bayesian probabilism, all probabilities must arise from empirical phenomena, and ought to be reserved for events that are (at least theoretically) repeatable.

As with falsification and frequentist hypothesis testing, we might ask: does Bayesian probabilism provide a solution to the problem of induction? Again, we will not be able to adequately address this question until we study Bayesian inference in more depth, in **?@sec-Bayesian**. Bayesian inference provides a formal framework for assessing how evidence bears on different hypotheses. Specifically, under the statistical assumptions, Bayesian inference assigns a “plausibility score”, in the form of a probability, to each hypothesis, e.g., $P(H|e) \approx 0.46$.. These probability assignments represent an important step forward in strengthening inductive inference: if the modeling assumptions are (roughly) met then the probabilities of various hypotheses are interpreted as our degrees of belief in those hypotheses. The higher the probability, the more likely the hypothesis is to be true. While it does make explicit and precise statements about degrees of belief in various hypotheses, it still assumes that the future will be roughly like the past, i.e., it assumes the UP. But, as we saw in section 1.2, the UP cannot be justified without circularity. So, in failing to avoid the UP, this statistical method has failed to circumvent the problem of induction. Nevertheless, Bayesian methods provide some guidance for belief and action under uncertainty.

Although, strictly speaking, the statistical inference methods described in this chapter do not *solve* the problem of induction, they go a long way toward placing induction on a stronger foundation. These methods are also quite different from each other. One champions falsification and refutation, and the other assigns probabilities directly to theories and hypotheses. What allows for these differences? Which one provides a stronger foundation for inductive inference? Are there other statistical inference paradigms that do better? The goal of the next few chapters will be to answer these questions.

American Philosophical Association. (2024). *Philosophy resources for undergraduate students*. <https://www.apaonline.org/page/undergraduates>

Animal research at the ICR. (2019). In *The Institute of Cancer Research*. <https://www.icr.ac.uk/our-research/about-our-research/animal-research/animal-research-at-the-icr>

Artificial sweeteners and cancer. (2016). In *National Cancer Institute*. <https://www.cancer.gov/about-cancer/causes-prevention/risk/diet/artificial-sweeteners-fact-sheet>

Bagla, J. S. (2009). Hubble, hubble’s law and the expanding universe. *Resonance*, 14(3), 216?225. <https://doi.org/10.1007/s12045-009-0022-8>

⁸We will consider issues related to interpretations of probability—e.g., whether probability *just is* a ratio of frequencies or not—in **?@sec-probability**.

- Band, P. R., Le, N. D., Fang, R., & Deschamps, M. (2002). Carcinogenic and endocrine disrupting effects of cigarette smoke and risk of breast cancer. *The Lancet*, 360(9339), 1044?1049. [https://doi.org/10.1016/s0140-6736\(02\)11140-8](https://doi.org/10.1016/s0140-6736(02)11140-8)
- Bayes, T., & Price, R. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late rev. Mr. Bayes, f. R. S. Communicated by mr. Price, in a letter to John Canton, a. M. F. R. s. *Philosophical Transactions of the Royal Society of London*, 53, 370?418. <https://doi.org/10.1098/rstl.1763.0053>
- Bryda, E. C. (2013). The mighty mouse: The impact of rodents on advances in biomedical research. In *Missouri medicine*. Journal of the Missouri State Medical Association. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3987984/>
- Cahan, D. (2003). *From natural philosophy to the sciences: Writing the history of nineteenth-century science*. The University of Chicago Press.
- Carnap, R. (1962). *Logical foundations of probability* (Vol. 2). Citeseer.
- Chapman, D. (2015). *Probability theory does not extend logic*. <https://metarationality.com/probability-and-logic>
- Clayton, A. (2021). *Bernoulli's fallacy: Statistical illogic and the crisis of modern science*. Columbia University Press.
- Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1), 1–13.
- Curd, P. (2016). Presocratic philosophy. In *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/presocratics/>
- Czitrom, V. (1999). One-factor-at-a-time versus designed experiments. *The American Statistician*, 53(2), 126?131. <https://doi.org/10.1080/00031305.1999.10474445>
- Davenport D.J, T. H., Patil, McAfee, A., & Brynjolfsson, E. (2012). Data scientist: The sexiest job of the 21st century. In *Harvard Business Review*. <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>
- Dphil, M. S. (2009). Meaning (verification theory). *Encyclopedia of Neuroscience*, 2253?2256. https://doi.org/10.1007/978-3-540-29678-2_3346
- Dvorsky, G. (2014). *Neil deGrasse tyson slammed for dismissing philosophy as useless*. <https://archive.ph/FHurA>
- Erdenk, E. A. (2015). Two tokens of the inference to the best explanation: No-miracle argument and the selectionist explanation. *Beytulhikme An International Journal of Philosophy*, 5(1), 31. <https://doi.org/10.18491/bijop.59053>
- Faraway, J. James. (2015). *Linear models with r*. CRC Press, Taylor & Francis Group.
- Ferrer, R. L. (1998). Graphical methods for detecting bias in meta-analysis. *FAMILY MEDICINE-KANSAS CITY-*, 30, 579–583.
- Fetzer, J. (2010). *Carl hempel*.
- Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 17(1), 69–78.
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, 309?368. https://doi.org/10.1007/978-1-4612-0919-5_2
- Fisher, R. A. (1935). The design of experiments. 1935. In *Oliver & Boyd Edinburgh, Scotland*. Oliver; Boyd.

- Fisher, Ronald A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, 98(1), 39–82.
- Gelman, A. (2009). Resolving disputes between j. Pearl and d. Rubin on causal inference. In *Statistical Modeling, Causal Inference, and Social Science*: https://statmodeling.stat.columbia.edu/2009/07/05/disputes_about/
- Hamajima, N., Hirose, K., Tajima, K., Rohan, T., Calle, E., Heath, C., & Coates, R. (2002). Alcohol, tobacco and breast cancer—collaborative reanalysis of individual data from 53 epidemiological studies, including 58 515 women with breast cancer and 95 067 women without the disease. *British Journal of Cancer*, 87(11), 1234–1245. <https://doi.org/10.1038/sj.bjc.6600596>
- Hawking, S., & Mlodinow, L. (2010). *The grand design*. Bantam Books.
- Henderson, L. (2018). The problem of induction. In *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/induction-problem/#Bib>
- Hoffman, J. I. E. (2015). Chapter 36 - meta-analysis. In J. I. E. Hoffman (Ed.), *Biostatistics for medical and biomedical practitioners* (pp. 645–653). Academic Press. [https://doi.org/https://doi.org/10.1016/B978-0-12-802387-7.00036-6](https://doi.org/10.1016/B978-0-12-802387-7.00036-6)
- Ioannidis, J. P. (2010). Meta-research: The art of getting it wrong. *Research Synthesis Methods*, 1(3-4), 169–184.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8). <https://doi.org/10.1371/journal.pmed.0020124>
- Jackson-Koku, G. (2016). Beck depression inventory. *Occupational Medicine*, 66(2), 174–175.
- Jobs, S. (2005). *Stanford commencement address*. <https://news.stanford.edu/2005/06/14/jobs-061505/>.
- Lewallen, S., & Courtright, P. (1998). Epidemiology in practice: Case-control studies. In *Community eye health*. International Centre for Eye Health. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1706071/>
- Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.
- Mayo, D. G., & Spanos, A. (2011). The error-statistical philosophy. In *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science*. Cambridge University Press.
- Meisner, R. C. (2019). Ketamine for major depression: New tool, new questions. In *Harvard Health Blog*. <https://www.health.harvard.edu/blog/ketamine-for-major-depression-new-tool-new-questions-2019052216673>
- Morris, W. E., & Brown, C. R. (2019). David hume. In *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/hume/>
- Norton, J. D. (2002). A survey of inductive generalization. In *A Survey of Inductive Generalization*. University of Pittsburgh. http://www.pitt.edu/~jdnorton/papers/Survey_ind_gen.pdf
- Norton, J. D. (2018). *The material theory of induction*.
- Okasha, S. (2016). *Philosophy of science: A very short introduction*. Oxford University Press.
- Papineau, D. (2018). Is philosophy simply harder than science? In *The TLS*. The Times Literary Supplement. <https://www.the-tls.co.uk/articles/public/philosophy-simply-harder-science/>
- Park, J. (2020). *On the question of progress in philosophy*. Epoeche Magazine. <https://>

- [//epochemagazine.org/20/on-the-question-of-progress-in-philosophy/](http://epochemagazine.org/20/on-the-question-of-progress-in-philosophy/)
- Pearl, J. (2009). *Myth, confusion, and science in causal analysis*.
- Persson, E., & Waernbaum, I. (2013). Estimating a marginal causal odds ratio in a case-control design: Analyzing the effect of low birth weight on the risk of type 1 diabetes mellitus. *Statistics in Medicine*, 32(14), 2500–2512. <https://doi.org/10.1002/sim.5826>
- Popper, K. ([1959] 2005). *The logic of scientific discovery*. Routledge.
- Popper, K. R. (2010 [1963]). *Conjectures and refutations: The growth of scientific knowledge*. Routledge.
- Reuber, M. D. (1978). Carcinogenicity of saccharin. In *Environmental health perspectives*. U.S. National Library of Medicine. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1637197/>
- Skorton, D. J., & Bear, A. (2018). *The integration of the humanities and arts with sciences, engineering, and medicine in higher education: Branches from the same tree*. The National Academies Press.
- Smith, G. D. (2002). Data dredging, bias, or confounding. *Bmj*, 325(7378), 1437–1438. <https://doi.org/10.1136/bmj.325.7378.1437>
- Stigler, S. M. (2016). *The seven pillars of statistical wisdom*. Harvard University Press.
- Stigler, S. M. (2018). Richard price, the first bayesian. *Statistical Science*, (1).
- Vickers, J. (2006). The problem of induction. In *Stanford Encyclopedia of Philosophy*. <https://stanford.library.sydney.edu.au/archives/sum2016/entries/induction-problem/#CarIndLog>
- Ward, J., Williams, J., & Manchester, S. (2017). 111 n.f.l. Brains. All but one had c.t.e. In *The New York Times*. The New York Times. https://www.nytimes.com/interactive/2017/07/25/sports/football/nfl-cte.html?_r=0
- What is philosophy and how do we do it? (2009). *Philosophy Now*, (79). <https://archive.ph/dPS9D>
- Wittgenstein, L. (2001 (1953)). *Philosophical investigations*. Blackwell Publishing.