The Philosophy of Statistics
An Introduction

Brian Zaharatos

May 9, 2022

1

Philosophy, statistics, and the philosophy of statistics

Jacobs & Wallach (2019) 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!

is entirely subjective, vague, imprecise, and incapable of progress.[4] These

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 convertable 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.[9] *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 testamony 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*.

[9] Of course, most arguments used in philosophy and science are much more complicated complex than the structure given above—two premises and a conclusion. We focus on these simple arguments to make a conceptual point.

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 property about incomplete, uncertain data.

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 di□cult. Perhaps the most famous exposition of the di□culties 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. 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 e□ect—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 e□ect (e.g., one can imagine that Susan licked the envelops 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 casual relationship would require going beyond mere correlations. Although correlations may suggest a causal relationship between two variables, correlations are not su□cient 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

ologies (e.g., hypothesis testing) for generating knowledge. In Chapters 4 and 5, we will learn about, and consider objections raised against, popular statistical methods.

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

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 Statistican 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 su□cient 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 [illegible] 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.

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 0.5(,0.5(-33y)0.2(p))-0339f[(y)0.2(i)0.2(0.5(-33y))-261.8(m)-0.3((i)(-33y)(e)--0.5(ser)-0.Ty

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,

x	$P(X = x)$
0	$1/70 \approx$

parameter.[23] 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 ($i = 1, \ldots, n$). But importantly, that isn't the end of the story. $\bar{X}$ for our sample won't be exactly equal to μ

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:

1a. If we increased speed by one mile per hour, we would need to increase our stopping distance β_1 units, on average.

2a. 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 squares[27]—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

Design

> No aphorism is more frequently repeated in connection with field trials, then 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) and the newly approved Esketamine[28]; 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.[29] To simplify our example, consider just two medical treatments, the SSRI citalopram, and Esketamine; and one talk therapy treatment, CBT.

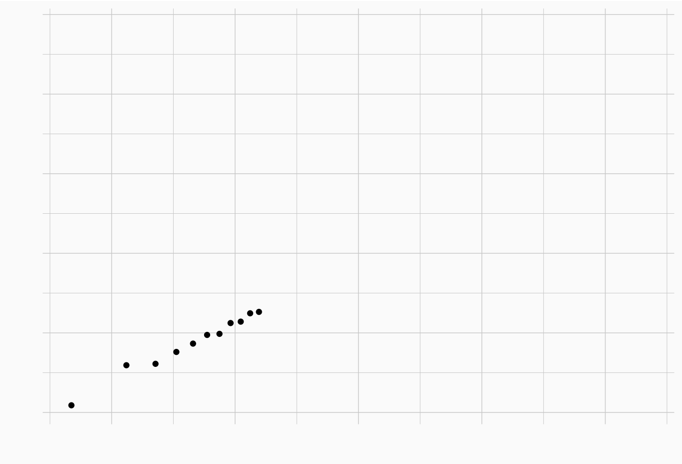
We can think of each treatment as a categorical variable, called a *factor*, with two *levels*: either the treatment has been given to a patient at the specified dosage and schedule, or it hasn't. We might imagine that patients receiving citalopram will receive 40 mg, once per day; patients receiving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

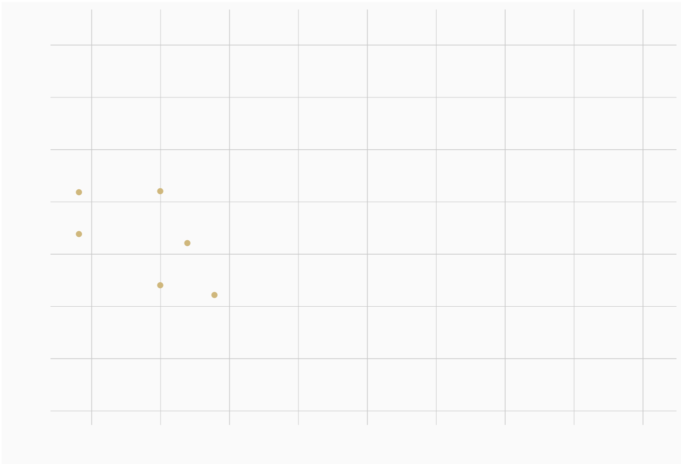
One procedure for testing the e□ectiveness of treatments for depression might be to consider only one factor; that is, administer a treatment, and only that treatment, and measure its e□ect on depression. For example, we might administer 40 mg of citalopram once per day, for 6 weeks, to a group of n_1 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 e□ective 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*. B2sinp

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.

3. *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





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:

1. *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" (Hoffman, 2015; Ferrer, 1998). 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; J. P. Ioannidis, 2010). Assessing the strength that a meta-analysis brings to a statistical argument is logical, and thus, philosophical, in nature.

2. *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-0.4-5(o)-2.8(-0.5(aci)0.-0.4(s)-0..5-27.0.5(i)0.2.

computed using Bayes' theorem:

$$P(T \mid x) = \frac{P(x \mid T)P(T)}{P(x)}. \tag{1.4}$$

This view of confirmation theory raises many questions. Ostensibly, theories are either true of 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:

$$P(T \mid x) > P(\neg T \mid x). \tag{1.5}$$

Equations (1.3) and (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?

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 swans have been white, and false that (C) All swans, including those yet to be observed, are white. As such, an inference about a hypothesis, H, based on an inductive argument is *risky*, in the sense that we may have taken in good information from the world, and properly encoded that information into a set of premises and assumptions, but drawn incorrect conclusions with respect to H.

Why does this problem arise? Why do we need to draw inferences to hypotheses or theories that go beyond the observations at hand? One reason is that scientific laws are su□ciently general, in the sense that they refer not to particular entities, 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 unobserved, 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).

Inferences to broad generalizations or unobservable entities aren't particular to the physical sciences. For example, 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 measure observable variables, and have a

to believe that demons exist, and even if they did exist, we have no reason to believe that they have the goal of unplugging our phones.

2.1.2 Induction by enumeration

What justifies our knowledge that all electrons have a mass of 9.1×10^{-31}g? Or that a hot stove will burn my hand? Or that there will be a full moon on January 18, 2030?[3] The argument for such knowledge is often of the form (Norton, 2002):

(P1) All *observed* instances of A have had property p.

(C) Therefore, *all* (including unobserved) instances of A

(P3)

$f($

2.4.1 Popper, Fisher, and induction

> *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 (1902 - 1994) 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

To be sure, this view has some problems. For one, we might notice that there is an asymmetry between our ability to reject T as false, i.e., when evidence e contradicts T; and accepting T as true, i.e., when e does not contradict T. In the latter case, strictly speaking, e being broadly consistent with T does not confirm

Under this model, the research hypotheses can be reformulated into statistical hypotheses. Let μ_1 be the mean energy consumption in the unmodified group, and μ_2 be the mean energy consumption in the modified group.

Bayesian inference makes use of this statistical model within *Bayes' theorem*. In particular, it is possible to use Bayes' theorem to produce a probability distribution over hypotheses about the parameter, θ, given data x. Bayesian inference thus allows us to quantify our degree of belief in different hypotheses, i.e., different values of θ. For example, the result of a Bayesian inference might be that, given the modeling assumptions (to be made more explicit below), we are justified in believing $H_0 : \theta \leq 0$ is five times more likely than $H_0 : \theta > 0$.

Consider again the research question about refrigerator energy consumption from the previous section. Let μ_1 be the mean energy consumption in the unmodified refrigerator group, and μ_2 be the mean energy consumption in the modified refrigerator group. For the sake of simplicity, let's assume that we have enough experience with the unmodified group to know that $\mu_1 = 1.5$. Using this assumption, Marilynne's research hypotheses from above are:

R_0 : The motor modification will not impact energy consumption
R_1 : The motor modification will impact energy consumption

Those research hypotheses were translated into statistical hypotheses:

$$S_0 : \mu_1 = \mu_2 \qquad \mu_2 = 1.5$$
$$S_1 : \mu_1 = \mu_2 \qquad \mu_2 = 1.5.$$

In Bayesian inference, we must start with a prior set of beliefs (or a "prior") about the parameter of interest, in this case, $\theta = \mu_2$. A prior will specify a probability distribution over the relevant values of θ, *before observing the data*. It quantifies our degree of belief in θ before collecting observations. In this case, a reasonable choice might be a normal distribution of θ, centered at 1.5, with variance σ_0^2:

$$\theta \quad N\left(1.5, \sigma_0^2\right).$$

Informally, by selecting this prior distribution, we are stating that we believe it is very likely that the true value of θ is relatively close to 1.5 (i.e., the normal distribution has its peak at 1.5, the value under H_0), and less likely that θ is far from 1.5 in either direction. This prior quantifies our belief that, before observing the data, there is a high probability that the modified group is no different than the unmodified group. The goal of a Bayesian analysis is to update our prior based on the data. This update results in a *posterior distribution*, $\pi(\theta \mid x)$, our degree of belief in θ given the data x. The posterior distribution comes from Bayes' theorem:

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\, \pi(\theta)}{\int f(x \mid \theta)\, \pi(\theta) d\theta}.$$

$\mu_2 \mid \mathrm{x} \quad N$

References

Animal research at the icr. (2019). Retrieved from https://www.icr.ac.uk/our-research/about-our-research/animal-research/animal-research-at-the-icr

Artificial sweeteners and cancer. (2016, Aug). Retrieved from 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. doi: 10.1007/s12045-009-0022-8

Band, P. R., Le, N. D., Fang, R., & Deschamps, M. (2002). Carcinogenic and endocrine disrupting e□ects of cigarette smoke and risk of breast cancer. *The Lancet*, *360*(9339), 1044?1049. doi: 10.1016/s0140-6736(02)11140-8

Baumer, B. S., Kaplan, D. T., & Horton, N. J. (2017). *Modern data science with r*. CRC Press,Taylor Francis Group.

Brown, D. L. (2019, Aug). *'you've got bad blood': The horror of the tuskegee syphilis experiment.* WP Company. Retrieved from https://www.washingtonpost.com/news/retropolis/wp/2017/05/16/youve-got-bad-blood-the-horror-of-the-tuskegee-syphilis-experiment/

Bryda, E. C. (2013). *The mighty mouse: the impact of rodents on advances in biomedical research.* Journal of the Missouri State Medical Association. Retrieved from 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.

Cohon, R. (2018, Aug). *Hume's moral philosophy.* Stanford University. Retrieved from https://plato.stanford.edu/entries/hume-moral/#io

Curd, P. (2016, Apr). *Presocratic philosophy.* Stanford University. Retrieved from https://plato.stanford.edu/entries/presocratics/

Czitrom, V. (1999). One-factor-at-a-time versus designed experiments. *The American Statistician*, *53*(2), 126?131. doi: 10.1080/00031305.1999.10474445

Davenport D.J, T. H., Patil, McAfee, A., & Brynjolfsson, E. (2012, Oct). *Data scientist: The sexiest job of the 21st century.* Retrieved from 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. doi: 10.1007/978-3-540-29678-2_3346

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. doi: 10.18491/bijop.59053

Faraway, J. J. (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.

Fisher, R. (1935). *The design of experiments. 1935*. Edinburgh: Oliver and Boyd.

Gelman, A. (2009). *Statistical modeling, causal inference, and social science.* Retrieved from https://statmodeling.stat.columbia.edu/2009/07/

from https://www.journalism.org/2018/06/18/distinguishing-between-factual-and-opinion-statements-in-the-news/

Morris, W. E., & Brown, C. R. (2019, Apr). *David hume.* Stanford University. Retrieved from https://plato.stanford.edu/entries/hume/

Norton, J. D. (2002, Apr). *A survey of inductive generalization.* University of Pittsburg. Retrieved from 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, Sep). *Is philosophy simply harder than science?* The Times Literary Supplement. Retrieved from https://www.the-tls.co.uk/articles/public/philosophy-simply-harder-science/

Paul, C., & Brookes, B. (2015, Oct). *The rationalization of unethi-*

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, Dec). Data dredging, bias, or confounding. *Bmj*, *325*(7378), 1437?1438. doi: 10.1136/bmj.325.7378.1437

Stigler, S. M. (2016).