

An Introduction to Statistics

Jonas Mikhaeil*

June 2025

(updated October 28, 2026)

Contents

1	An Introduction to an Introduction	3
2	A Perspective on Statistics: The Craft of Aggregation	4
3	Probability	7
3.1	Interpretations of Probability	7
3.1.1	Frequencies	7
3.1.2	Subjective Probabilities	9
3.2	Mathematical Foundations of Probability	10
3.2.1	The Kolmogorov Axioms and a Dutch Book	11
3.2.2	Properties of Probabilities	13
3.2.3	Random Variables	15
3.3	Probability Distributions	18
3.3.1	Expectations and The Problem of Points	18
3.3.2	The Binomial Distribution	20
3.3.3	From Discrete to Continuous Distributions	21
3.3.4	The Normal Distribution	23
3.3.5	The Importance of Variance	24
3.4	Joint and Conditional Distributions	26
3.5	Sampling and the Distribution of the Sample Mean	28
3.6	Large Samples: Asymptotics and the Central Limit Theorem	30

*These notes are still incomplete, and I hope to soon be able to return to them. In the meantime, I apologize for any mistakes or omissions. Any feedback is more than welcome: j.mikhaeil@columbia.edu. I'm deeply grateful for long and insightful discussions about the history of statistics with Alma Steingart. I'm also indebted to Audrey Yang for her invaluable feedback and corrections on early drafts of these notes. Everything that is true in these notes I have learned from others; all mistakes are my own.

4	Estimation and Uncertainty Quantification	32
4.1	Bayesian Inference	33
4.1.1	Bayes and the Pool Table	33
4.1.2	Bayesian Uncertainty Quantification: Posterior credible interval . .	36
4.1.3	Laplace and the Motion of Planets	37
4.1.4	Bayesian Inference for the Normal Distribution	38
4.1.5	Non-informative Priors (Objective Bayes)	39
4.2	Maximum Likelihood	40
4.2.1	Confidence Intervals	46
4.3	A modern perspective on Bayesian Inference	46
5	Science and Hypothesis Testing	46
5.1	Induction, the Backbone of Science: From Confirmation To Methodolog- ical Falsificationism	47
5.2	A Lady Tasting Tea: Milk or Tea First?	49
5.3	Are Sex Ratios Proof Of Divine Providence?	51
5.4	Neyman-Pearson Hypothesis Testing	51
5.5	P-Values	56
5.6	Crisis in Science and Fallacies of Rejection	56
6	Decomposing Variation: Linear Regression	58
6.1	Galton, Decomposing Variation and Regression to the Mean	58
6.2	Pearson and Correlation	60
6.3	Linear Regression: Disentangling Variation	62
7	Causal Inference	62

1 An Introduction to an Introduction

These lecture notes grew out of a calculus-based introduction to statistics class I taught at Columbia in summer 2025. My goal with these notes is simple: Take you, the reader, on a journey that shows *why statistics is exciting and worth studying*. Writing these notes, I often thought of taking the reader along on a hike. Parts of any good hike are strenuous; so too are, at points, these notes. They are challenging, and not everything will immediately make sense for someone who is studying statistics for the first time. I believe that's okay – we only learn when we challenge ourselves. On our hike, we pass through beautiful forests and lush pastures, but we don't have time to stop. Many topics are covered only cursorily, and we have to content ourselves with pointing to more advanced material. Hopefully, you'll find yourself motivated to come back and study statistics in more detail, perhaps even taking more advanced classes or reading books on some of the topics we could only scratch the surface of here.

Throughout these notes, we turn to history to better understand modern-day statistics. This may seem almost absurd in a time of rapid advances in machine learning, data science, and artificial intelligence. Even if statistics is still important, isn't now the time to look ahead instead of gazing into the past? I would argue “no”: it's actually the perfect time to look at the past – to better understand the present. Turning to the struggles of early statistics to establish itself, along with its promises and failures, help us recognize what makes statistics important. Looking at this century-long struggle to realize statistics' promise of transforming the world, we can more clearly see what statistics offers and what its challenges are: creating useful summaries to understand the world. We've become all too used to summaries. It has become so natural for us to think and argue in terms of averages. Think, for example, of vaccination rates, mortality rates, and unemployment rates. We hardly ever stop to think how these quantities were constructed and why (or better, how) they are useful. These aggregates live in an ambiguous realm between facts and fiction: they are “crafted,” and when they are crafted well, they help us better navigate the world. More on this in the next section. As society becomes increasingly structured around data, it is our job as statisticians to understand when and how data is useful.

A brief note on teaching: These notes quite intentionally cover some advanced material. As such, it is unrealistic to expect students studying statistics for the first time to follow them completely. In my teaching of a calculus-based introduction to statistics, I additionally relied on a standard textbook [Devore, 2016]. Each class, I assigned readings from the textbook which set the expectation for the midterm and exam. I then offered office hours and recitations to give students ample opportunity to work through their questions about the material. This opened up classroom time to develop a broader and more conceptual understanding of statistical ideas along the lines of these notes. This particular teaching style worked well for the small and very motivated class I had the fortune to teach.

2 A Perspective on Statistics: The Craft of Aggregation

It is maybe naive to try and answer what an entire academic discipline is all about. And yet, I will try – or at least, I hope to give a brief, inevitably limited, perspective on statistics. The goal is not so much to give you an exhaustive overview of all the questions and problems raised and solved by current-day statistics. My aim is to explain *why statistics is hard* and at the same time *why statistics is useful*.

Gazing back into history, we find one of the roots of statistics in 18th century astronomy. Newton had wrestled the stars from the order of the divine into the human universe. Now, astronomers were busy measuring and mapping celestial objects with a reinvigorated fervor. In their quest, a tricky question arose: how to combine multiple observations of the same object. This question would lead Bernoulli, Laplace and Gauss to develop some of the most fundamental results of statistics [Stigler, 1990]. This origin suggests that statistics is the *logic of measurement*.

This is not the only origin of modern day statistics, however. In the century after astronomers raised their heads to look at the stars, a new kind of state – the centralized nation state – emerged out of the upheavals of the French Revolution. Power previously held by local elites was consolidated into a central government. To wield power, the state needed to collect and organize vast amounts of information about its land and populace – statistics was the tool that allowed the state to make society “legible” [Scott, 1999]. It is actually this tradition of governance, specifically the German school of the 18th century, that gave rise to the name “statistics”.¹

It would take until the 20th century for both these traditions – probability theory birthed out of astronomy and gambling, and statistics as organizing large scale administrative data – to fully merge and give rise to our modern-day understanding of statistics. In both traditions that nurtured statistics, statistics facilitated the movement from disparate individual cases to a comprehensible whole. I suggest that we should think of statistics as the *craft of aggregation*. You may be on board with the part about aggregates, as it seems clear that statistics is (often) concerned with aggregation. But what do I mean by it being a “craft”? My claim is that we should think of statistical aggregates as crafted: they are not fictitious but they are also not above reproach. What makes a statistical aggregate useful is its stability in the face of other people’s scrutiny [Desrosières, 2002]. There is no way of doing this by rote – every application of statistics is different and requires intuition, communication and adaptation of one’s models. Statistics is a craft, not a logic.

To illustrate this, let us look at how statistics established itself in the social sciences. The hope for a quantitative approach to social sciences and especially the anticipation of the usefulness of probabilities to explain society can already be found in Bernoulli’s writings in the 17th century. While there were sporadic attempts to apply probability theory to answer questions about society – for example, Laplace’s attempt to estimate the French population based on a random sample and Poisson’s attempt to understand

¹It’s related to the German word for state - “Staat”. The inventor of the word is thought to be Achenwall (1719-1772) who was part of the University of Göttingen “school of statistics”. [Desrosières, 2002, pg.19]

the fallibility of the judicial system – it would take almost two centuries until significant headway was made. The first real advance of statistics into the social realm was made by Adolphe Quetelet in the 19th century [Stigler, 1990]. Quetelet noted that many time-series of aggregate characteristics of a populace showed surprising regularity. From year to year, birth-, death-, marriage-, and even suicide-rates seemed to be consistent. Think about what kind of shock this must have been: We are creatures of free will, as individuals completely unpredictable – but on the level of the aggregate, the average of our decisions remains constant from year to year. Ever since, sociology has aimed to explain these “law-like” tendencies of society.

The question that we take aim at is how to understand the meaning of these average quantities. Yes, the consistency of average suicide rates is surprising, but what does this average refer to? It is this conceptual challenge, the question of the “realism of the aggregate” that made advances of statistics into the social sciences so arduous [Desrosières, 2002].

Quetelet set out to convince his peers that population averages were “real”. This conviction was grounded by another empirical observation he made. He found that the population distribution of many characteristics, such as height, followed a normal distribution. Guided by what, from a modern vantage point, is understood to be a fallacy, he used the central limit theorem to argue that this normal distribution implies a “constant cause”. Let us try to understand his reasoning. Quetelet’s reasoning parallels sampling from an urn. Suppose you draw many balls from an urn containing black and white balls. After drawing each ball, we put them back so as not to change the constitution of the urn. Quetelet knew that the ratio of black to white balls observed when drawing many balls converges to a normal distribution centered around the (true) ratio of black and white balls in the urn. He then interpreted the (fixed) constitution of the urn as a “constant cause” whereas he understood the fluctuations of the observed ratio from the true ratio as “accidental causes”. The central limit theorem, for Quetelet, guaranteed that “accidental causes” (individual decisions, measurement error, ...) causing deviations from a “constant cause” (the fixed constitution of the urn) would lead to a normal distribution centered around this “constant cause”. He then turned this reasoning on its head and argued that any observed normal distribution implied such an underlying “constant cause”. This “constant cause” then gives reality to the population average. The ratio of individual balls seen is not a mere average: it reveals the true constitution of the urn. Similar, Quetelet argued that the population average is not just an aggregation of many individuals, it reveals a societal reality. But what is this societal “constant cause”, and what is the “constant cause” behind the normal distribution of heights? Quetelet proposed the “average man”: We humans are imperfect copies of an ideal, average human. This ideal, the “average man”, is the “constant cause”; our personal deviations are “accidental”.

Things happen as if the creative cause of man, having shaped the model of the human species, had then like a jealous artist broken his model, leaving inferior artists the task of reproducing it. – Quetelet, taken from [Desrosières, 2002, pg. 76]

This may, in hindsight, sound incredibly naive, but Quetelet’s “average man” was one of the first (partially) successful attempts at creating a meaningful statistical aggregate. It’s unclear what we learn from a mere average but learning about the “average man” reveals the creator’s intent.

During his lifetime, Quetelet was known as the “famous Quetelet”, but despite his influence, he has been largely forgotten – only rescued from absolute obscurity by scholars of the history of statistics [Stigler, 1990, Desrosières, 2002]. Ironically, his legacy is overshadowed by all those that he inspired. His notion of the “average man”, while influential, would soon come under much criticism. While it influenced Durkheim, he would heavily criticize it in his famous work *Suicide*, that is often referenced as being one of the origins of modern quantitative sociology. Durkheim contrasts the “average type” with the “collective type”, which he argues is the true interest of sociology. While most individuals, and hence the average, may have few qualms about “enrich[ing] themselves fraudulently” [Durkheim, 1897, pg. 317], the collective stigmatizes theft. In a sense, Durkheim not only criticizes Quetelet’s argument for the “reality” of aggregates but the use of statistics in sociology more generally. If the “collective”, the explanandum of sociology, and the “average” are disjoint, then it is unclear how sociology can benefit from statistics.

This tension runs through all of applied statistics. In applied statistics, the highly formalized world of mathematical statistics comes in contact with the world of our experience – one that avoids complete formalization and continuously spills out of the conceptual boxes we try to fit it in. As applied statisticians, we need to come up with aggregate descriptions of the world to make use of the tools of mathematical statistics, but we then need to explain and interpret these aggregate concepts. We need to see if they hold up to scrutiny. Once they crumble –and they will– they need to be abandoned or adapted; in either case, a new attempt at crafting a meaningful statistical aggregate must be made.

This suggests a shift from a rigid realism of aggregates – for example, the existence of an “average man” to give reality to the distribution of heights – to a modeling perspective. For our aggregates to be useful, they need not be beyond question; they just need to be stable enough for whatever we plan to do with them. Different endeavors may require more or less stability; think of personal decision making in comparison with scientific inquiry. Regardless of the domain, however, these aggregates will always be questioned and we may need to go back to them and revise them. Compare this with Box’s famous aphorism: “All models are wrong, but some are useful”. This suggests we should understand statistical aggregates as statistical models of the world. Models aren’t true or false; they are crafted to serve a specific purpose.

From this vantage point, there are three central tasks for us applied statisticians: First, we need to come up with models of the world. This is the hardest, and possibly most interesting task, and the one these notes will teach you the least about. Coming up with models requires expertise in your field, experience, and intuition.

Holding a model to be true, mathematical statistics helps us learn about the model based on empirical observations. More specifically, statistics formalizes how to infer a model’s parameters based on data. With this statistical inference also comes a quantifi-

cation of our remaining uncertainty about these parameters. We will cover estimation and uncertainty quantification in Section 4.

Lastly, statistics offers tools to put our models to the test. It is an essential part of science to test our hypotheses and to try our best to falsify them (to find out if they are wrong). We do this via hypothesis testing, which we cover in Section 5.

Before we dive into these aspects of statistics, we need to cover the basics of probability theory. Probability theory is the basic vocabulary of statistics and any rigorous understanding of statistics vitally requires a good understanding of the fundamentals of probability.

3 Probability

Probability is the foundation on which statistics is built. So we begin by asking what seems to be a simple question: “What do we mean by probability?”

3.1 Interpretations of Probability

If the question “What do we mean by probability?” seems simple to you – simple enough that you wonder why we even bother with the question – imagine a friend of yours, previously unconcerned with statistics, asking you this very question. How do you respond? How do you explain to them what probabilities are? Maybe you have spent some time thinking about it and come to an answer that you’re satisfied with; maybe you have found that it is actually difficult to give a satisfying explanation.

The latter may very well be the case. For the better part of the last four centuries, thinkers ranging from brilliant mathematicians – including Laplace, one of the greatest minds we will encounter in this course – to influential philosophers of science (such as Popper and Lewis) have grappled with this question.

As one might guess from the length and persistence of this (still ongoing) debate, there are no simple or univocally accepted answers. The interested reader is referred to the Stanford Encyclopedia of Philosophy’s entry on interpretations of probability.

Here we only want to mention two interpretations of probability that highlight two very important facets of what we may aim to capture with the notion of probability. We will begin with a frequentist perspective on probability. This perspective is based on the wish to objectively reason about observed variation. We will then continue to subjective probabilities, which relate probabilities to our personal degrees of belief. In a sense, these views are complementary. Some situations or fields of research require us to try to quantify variation in data objectively. Other times, there is a need to quantify (possibly subjective) uncertainty and to reason about how to make decisions in the face of this uncertainty.

3.1.1 Frequencies

For a paradigmatic example of probability, consider tossing a fair coin. Say, we toss the coin ten times and record whether it lands “heads” (H) or “tails” (T). We may now be

tempted to identify the probability of the coin landing “heads” with the observed (or realized) frequency of “heads” in our sequence. Doing so, we quickly run into trouble. Toss a coin yourself; it’s most likely not going to land “heads” exactly five times. Do we conclude that your coin is biased? Of course we don’t; it’s an integral part of our notion of probability that we should expect variation. Of course, a coin is not always going to land *exactly* five times “heads” in a sequence of ten tosses. Ten is just not enough tosses, you might insist.

Can we solve the problem by tossing the coin a thousand times, a hundred thousand times? Feel free to try, but again, we should not expect to find exactly half of the tosses to land “heads”. What we can expect, however, is that the observed bias – the deviation from exactly half the tosses landing “heads” – shrinks.² Intuitively, you may not be surprised to find 7 out of 10 tosses to give “heads” but you should be surprised if 7000 out of 10,000 tosses come up “heads”.

Following this logic, we can try to identify probabilities with the limiting relative frequencies we would observe if we had tossed the coin infinitely many times. One of the most influential statisticians of all time, Ronald Fisher, was a proponent of this perspective:

[P]robability is the most elementary of statistical concepts. It is a parameter which specifies a simple dichotomy in an infinite hypothetical population, and it represents neither more nor less than the frequency ratio which we imagine such a population to exhibit. For example, when we say that the probability of throwing a five with a die is one-sixth, we must not be taken to mean that of any six throws with that die one and one only will necessarily be a five ; or that of any six million throws, exactly one million will be fives ; but that of a hypothetical population of an infinite number of throws, with the die in its original condition, exactly one-sixth will be fives. – [Fisher, 1922, pg. 312]

At first glance this might sound compelling – and we will see later on that limits of relative frequencies are related to probabilities. On the level of definitions, the story is unfortunately not as simple. Even when considering infinite sequences, finding your coin to always land “heads” is not (logically or physically) impossible. It is just highly improbable. The *law of large numbers*³ tells us that for coin tosses, the relative frequency of “heads” approaches the true probability with probability 1 (we say, *almost surely*). Our definition of probability is now at risk of becoming circular. Trying to define probability, we cannot take recourse to the claim that infinite sequences whose limiting relative frequency is not the true probability (such as “heads” for all eternity) are highly improbable – because to explain what we mean by “highly improbable”, we already need to understand what we mean by “probability”.

While not a convincing definition (and it is unclear if one can be given), it is clear that we often want to use probabilities to quantify observed variation objectively. The lack

²The quantitative aspect of this will be covered later

³see Section 3.6

of a convincing definition should not stop us. We will see that, circularity aside, and embracing some degree of pragmatism, relative frequencies are linked to probabilities, and we will also see that, even for finite sequences, observed frequencies can be expected to be close to the true probability. Many of the methods we will cover will, in spirit, be very close to this frequentist perspective on probability.

3.1.2 Subjective Probabilities

So far, we have tried to explain probabilities in terms of the variation of the outcomes of a random event, in particular a coin toss. Another influential perspective on probability is to understand probabilities as beliefs. In 1847, De Morgan wrote: “By degree of probability, we really mean, or ought to mean, degree of belief” [De Morgan, 1847]. In this view, probabilities are a tool to quantify our uncertainty – we commonly refer to these as *subjective probabilities* because they quantify our personal beliefs and can differ from person to person.

What do we mean when we say that we *believe* a coin to be fair, that we *believe* that “heads” and “tails” are equally likely? There is a long history of thinking about subjective probabilities in terms of betting behavior. De Finetti, one of the pioneers of subjective probabilities, writes:

Let us suppose that an individual is obliged to evaluate the rate p at which he would be ready to exchange the possession of an arbitrary sum S (positive or negative) dependent on the occurrence of a given event E , for the possession of the sum pS ; we will say by definition that this number p is the measure of the degree of probability attributed by the individual considered to the event E , or, more simply, that p is the probability of E (according to the individual considered; this specification can be implicit if there is no ambiguity). – [de Finetti, 1990]

In more modern phrasing, we could say that your subjective probability p is your degree of belief in the occurrence of event E . This degree of belief in E is defined as your willingness to buy a bet for $\$p$ which pays out $\$1$ if E occurs (or is true). In the context of our coin example this means the following: If you believe a coin to be fair (i.e., that the probability of it coming up “heads” is 50%), you’d be willing to buy a bet on it coming out “heads” for 50 cents. If you are afraid that the coin is biased, we can probe your degree of belief about how biased you think the coin is (essentially, your subjective probability of “heads”) by asking you how much you would be willing to spend on the bet. If, for example, you have strong reasons to believe the coin is biased towards “heads”, you should be willing to spend more on the bet.

The metaphor of betting is a conceptual tool to use decision-making to quantify degrees of belief in terms of the willingness to buy a bet at a certain price.

It turns out that if you want to avoid guaranteed losses, these subjective probabilities have to fulfill the same axioms as frequentist probabilities – that is, they need to fulfill the laws of classical probability theory. If your subjective probabilities do not obey

(classical) probability calculus, an antagonist can exploit your betting behavior to win money from you regardless of the outcomes of the bets. These series of bets, bets that collectively guarantee loss, are called Dutch books. The Dutch book argument shows equivalence between the classical (Kolmogorov) probability calculus and (rational) subjective probabilities. We'll discuss this in more detail in the next section (Section 3.2.1).

There are multiple concerns with this subjective perspective. First and foremost, probabilities are frequently used in a scientific context. Scientists, unlike gamblers, are not concerned with winning a bet (or decision-making more broadly) but with explaining the world. We may wonder how accurately the gambling metaphor describes the structure of scientific beliefs. More broadly, even when embracing the decision-making perspective of the Dutch book argument, it is unclear if our human beliefs are (or possibly can be) structured as probability distributions. Various psychologists have shown people's beliefs to drastically violate probability calculus, see for example [Kahneman et al., 1982]. If you wonder if this is an educational issue, I want to point out that ever since Laplace – that is, for more than 200 years – there has been debate among some of the most renowned statisticians about which probability distribution should be used to describe absolute lack of knowledge – the state of ignorance. Imagine for example you find a coin in front of yourself that I tell you may or may not be biased. I don't give you any hints about its bias and you find yourself in an absolute state of ignorance. What distribution describes your belief about this coin's probability p to land “heads”? This may not be as trivial as it seems and we will continue this discussion in Section 4.1.5.

3.2 Mathematical Foundations of Probability

Probability is an integral part of the vocabulary of statistics. Regardless of how we choose to interpret probabilities, they rest on classical probability theory as their mathematical foundation. In this section, we introduce Kolmogorov's axioms that define probabilities in mathematical terms. We then go on to discuss some of the most important concepts in probability theory⁴ (such as conditional probability and random variables). Before we do so, we need to introduce some basic terminology.

With statistics, we aim to shed light on some unknown phenomenon. We do so by imagining (or conducting) an experiment that probes the relevant part of the world and returns a (random) result. As a running example, think about tossing a coin. You may be wondering if the coin is biased (and if so, by how much). Through the course of these notes, you'll learn how to answer this question. The first step in this endeavor is to identify all possible outcomes of your experiment. In statistical jargon, we speak of the sample space:

Definition 3.1 (Sample Space). The sample space of an experiment, denoted by S , is the set of all possible outcomes of that experiment.

⁴Unless otherwise stated, all results in this section and their proofs can be found in [Casella and Berger, 2024]. This is also a great resource for readers that want to learn more about the mathematical details.

In the case of flipping a coin, the sample space is

$$S = \{H, T\},$$

where H stands for “heads” and T for “tails”. Once we have defined the sample space, we can define the events that can occur in our experiment.

Definition 3.2 (Event). An event is any collection of possible outcomes of an experiment. That is, any subset of S is an event.

Example 3.1. Can you list all events for the experiment of flipping a coin?

3.2.1 The Kolmogorov Axioms and a Dutch Book

Now that we have defined the sample space of all possible outcomes of a random experiment, we want to ask how probable any possible event of this experiment is. In Section 3.1, we have discussed two interpretations of probability: Probability as describing long term frequencies of repeated random experiments or subjective probability as a quantification of degrees of belief. Following this discussion, mathematically formalizing probability may seem like a very daring endeavor. It turns out, however, that the axioms on which all of probability theory rests are extremely simple. Despite their apparent simplicity, it took mathematics a long time to mature enough to allow for such a condensed formalization of probability theory. Our modern axiomatic understanding of probability is quite young. While probabilities have been used since at least the 17th century (we will see an early use of probabilities to solve problems in gambling by Pascal in the year 1654 in a later section), the mathematical foundations we’ll introduce in this section go back to Andrey Kolmogorov’s work from 1933.

Definition 3.3 (Kolmogorov Axioms). Given a sample space S , a probability function⁵ is a function P that satisfies:

1. $P(A) \geq 0$ for all events A .
2. $P(S) = 1$.
3. If $A_1, A_2, \dots$ are pairwise disjoint events, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Remark. *The definition as stated above is not completely accurate; it is wrong in a very subtle way that almost broke all of mathematics. As long as the sample space S is finite (or at least countable), the above is completely accurate. In the case of an uncountable sample space S (think of the case where the outcome of your experiment could be any real number $\mathbb{R}$), things become very weird, however. The problem is that the set of all subsets of an uncountable set includes sets that break probability theory. Let me very briefly try to explain why: It turns out that probabilities can be thought of as a (very*

⁵Technically, we need to specify the domain of the function to be a sigma algebra. In the conditions below, the events are then restricted to be elements of the sigma algebra. This detail does not matter for this draft, the interested reader is referred to, for example, [Casella and Berger, 2024]

abstract) notion of the “volume” of an event – we will see that to calculate the probability of an event A , we will need to integrate: $\int_{x \in A} p(x) dx$. Notice that this looks very similar to how we would calculate the volume of a shape A in Euclidean space $\int_{x \in A} dx$.

In 1924, Stefan Banach and Alfred Tarski discovered a paradox at the heart of mathematics: Take a solid ball in 3 dimensions⁶, Banach and Tarski discovered that if you chop up the ball in the right way and move and rotate these pieces, without any stretching, bending, or adding of new points, you can re-assemble the pieces to give two balls, each of the same volume as the original one. This is the famous Banach-Tarski paradox. Its result is clearly counter-intuitive and came as quite a shock for the mathematical community at the time.

It turns out that for the notion of “volume” to make sense, we need to restrict ourselves to only considering “measurable” sets, those that do not lead to the above paradox.

In terms of probabilities, this illustrates why we may need to restrict which subsets of an uncountable sample space count as events. We therefore specify a suitable σ -algebra on which our probability function is defined. This is where our discussion unfortunately needs to end. If this is interesting to you, any advanced probability textbook or lecture will cover this!

After this slight detour, let us now return to Kolmogorov’s axioms. In words, Axiom #1 states that there are no negative probabilities and Axiom #2 ensures that every experiment has an outcome (something must happen). Axiom 3) is slightly more involved but still intuitive. Disjoint events are events that cannot occur simultaneously. When throwing a die, the event of rolling a one, $A_1 = \{1\}$, and the event of rolling a four, $A_2 = \{4\}$, are disjoint. Axiom #3 now requires that the probability of either A_1 or A_2 happening (that is, rolling either a one or a four) is just the sum of the probability of rolling a one and the probability of rolling a four. Make sure this makes sense to you! Notice that for events that are not disjoint, for example the event of rolling a one or a four, $A_3 = \{1, 4\}$, and the event of rolling a four or a six, $A_4 = \{4, 6\}$, probabilities do not simply add up like this. What is amazing about these axioms is that despite their simplicity, all properties of probabilities (for example how to calculate the probability for these overlapping events) are already completely implied by them. We see some of these implications in the next section.

While these axioms are intuitive, there has been some effort to derive them from other principles. As discussed in Section 3.1.2, subjective Bayesians have managed to derive these axioms from avoiding exploitable losses in betting. In doing so, they have related decision making under uncertainty to (classical) probability theory.

It’s illuminating to walk through their argument.⁷ The goal is to show that probabilities (as degrees of belief) that quantify how much you are willing to bet need to follow Kolmogorov’s axioms. If they don’t, there will exist a set of bets guaranteeing a net loss for the side violating the axioms – a so-called Dutch Book.

⁶This does not work with a physical ball made out of atoms. For this to work, you need to get your hands on a mathematical ball, that is, a ball in 3-dimensional Euclidean space.

⁷This argument is taken from the Stanford Encyclopedia of Philosophy entry on “Dutch Book Arguments”.

Suppose you are betting against someone and you're betting on the claim H (for example, H : the coin comes out "heads"). If you win the bet, you win the stake S . Your degree of belief in H determines at which price qS you are willing to buy the bet. This table summarizes the situation. Suppose your beliefs violate Kolmogorov's

H	Payoff
True	$S - qS$
False	$-qS$

axiom #3.⁸ Let $Q(H)$ be your betting quote q for claim H . Now suppose we are betting on two mutually exclusive claims H_1 and H_2 . Think, again, of throwing a die: we could for example bet on the die landing on a one $H_1 = \{1\}$ and landing on a four $H_2 = \{4\}$. Now suppose your degree of belief in $Q(H_1 \cup H_2)$ violates Kolmogorov's third axiom, that is, $Q(H_1 \cup H_2) \neq Q(H_1) + Q(H_2)$. Without loss of generality, assume $Q(H_1 \cup H_2) < Q(H_1) + Q(H_2)$. I will now buy three bets from you. First, I will offer you the bet that pays \$1 if H_1 is true and \$0 otherwise, which you will be willing to buy for $Q(H_1)$. Similarly, I will offer you a bet that pays \$1 if H_2 is true and \$0 otherwise, which you will, according to your degree of belief, buy for $Q(H_2)$.

Now comes the trick. I will buy a third bet from you that will guarantee me a win. According to your beliefs, you are willing to sell me a bet that pays me \$1 if $H_1 \cup H_2$ comes true and \$0 otherwise for $Q(H_1 \cup H_2)$. Table 1 shows the possible payoffs for you after these three bets. Because of your inconsistent belief that $Q(H_1 \cup H_2) < Q(H_1) + Q(H_2)$, you are guaranteed a loss in all three scenarios.⁹ Can you show that there exist bets

H_1	H_2	Payoff for you
T	F	$(1 - Q(H_1)) - Q(H_2) + (Q(H_1 \cup H_2) - 1)$
F	T	$-Q(H_1) + (1 - Q(H_2)) + (Q(H_1 \cup H_2) - 1)$
F	F	$-Q(H_1) - Q(H_2) + Q(H_1 \cup H_2)$

Table 1: Beliefs that violate Kolmogorov's third axiom lead to a Dutch Book.

that guarantee you a loss if your beliefs violate the other two axioms?

3.2.2 Properties of Probabilities

In the previous section, we have discussed the Kolmogorov axioms of probability. What is possibly most surprising about them is that despite their simplicity, they suffice as a foundation for all of probability theory. We will now discuss some of the basic implications these axioms have for the properties of probability functions.

Theorem 3.1. *If P is a probability function and A is an event, then*

⁸Here we will only consider a violation of finite additivity. By looking at infinite sequences of bets, this argument can be extended to show how violations of countable additivity lead to a loss, see [Adams, 1962]

⁹For the case $Q(H_1 \cup H_2) > Q(H_1) + Q(H_2)$, just play through the same argument but flip the signs.

1. $P(\emptyset) = 0$, where $\emptyset$ is the empty set.
2. $P(A) \leq 1$.
3. $P(A^c) = 1 - P(A)$.

While these properties, especially $P(A) \leq 1$, have the flavor of axioms too, we need not assume them – they are already implied by Kolmogorov’s axioms. The Kolmogorov axioms specify how to calculate the probabilities for the union of mutually exclusive events. The following theorem shows how to calculate the probability of the union of (possibly overlapping) events:

Theorem 3.2. *If P is a probability function and A and B are events, then*

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
2. If $A \subset B$, then $P(A) \leq P(B)$.

The following theorem is of great use, especially in the Bayesian context:

Theorem 3.3 (Law of Total Probability). *If P is a probability function, then*

$$P(A) = \sum_{i=1}^{\infty} P(A \cap C_i) \text{ for any partition}^{10} C_1, C_2, \dots \quad (1)$$

One notion from probability that is essential for statistics is conditional probability. Suppose you throw a die. Before you guess which face it landed on, I take a peek. I go on to tell you that the die landed on an even number (event $B = \{2, 4, 6\}$ occurred). Conditional probability tells you how you should update your beliefs based on this information:

Definition 3.4 (Conditional Probability). Let A and B be events and $P(B) > 0$, then the conditional probability of A given B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (2)$$

The intuition behind conditional probability is that given the new information, the sample space is updated from S to B . To see this, note that for the original sample space S , we had $P(S) = 1$. Now we have $P(B|B) = 1$, that is, by conditioning on the event B , the set B acts as the new updated sample space.

Another crucial notion is statistical independence:

Definition 3.5 (Independence). Two events A and B are independent if

$$P(A \cap B) = P(A)P(B) \quad (3)$$

It is easier to interpret independence by understanding what it implies for conditional probabilities. In the case that two events A and B are independent, the conditional probability of A given B becomes

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A). \quad (4)$$

Intuitively speaking, if A and B are independent, learning B does not tell us anything about A .

Bayes' rule is central to how we update our beliefs given new information. Here, we will give it for the probability of events and discuss its implications for estimation in more detail in Section 4.1.

Theorem 3.4 (Bayes' Rule). *Let $A_1, A_2, \dots$ be a partition of the sample space, and let B be any set. Then for each $i = 1, 2, \dots$,*

$$P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B \cap A_j)}. \quad (5)$$

3.2.3 Random Variables

Oftentimes, we are not interested in the exact probability structure of an experiment but rather in some meaningful summary variable. For example, think of conducting a survey in which you ask 500 participants about whether they support a specific policy or not. For simplicity, we will think of this as a binary response. In this case, the sample space has 2^{500} elements – an ordered string with each participant's response. In this form, just by the sheer bulk of detailed information, it is almost impossible to learn anything from the data. An essential part of statistics is reducing the data in a meaningful way to allow comprehension. In the case of the survey, we may not be interested in each participant's individual response but rather in the overall proportion – what is the support for the given policy at large? The technical tool that allows us this shift from the potentially unwieldy sample space to a meaningful outcome measure is a random variable:

Definition 3.6 (Random Variable). A random variable X is a function from the sample space S to the real numbers, formally $X : S \rightarrow \mathbb{R}$.

Example 3.2. Continuing the example of the survey may help elucidate the concept of a random variable. Let us encode support of a policy as a 1 and opposition as a 0. For simplicity, we consider a tiny survey with three participants. In this case, the sample space has $2^3 = 8$ elements:

$$S = \{000, 001, 010, 100, 011, 110, 101, 111\}.$$

Using a random variable, we want to find a more meaningful description of the possible outcomes of the survey. We may, for example, choose

$$X = \text{Proportion of participants that support the policy.}$$

The random variable X is a function mapping every element in the sample space $s \in S$ to a number in $[0, 1]$:

Element of the sample space s	000	001	010	100	011	110	101	111
$X(s)$	0	1/3	1/3	1/3	2/3	2/3	2/3	1

Remark. *Random variables may be one of the most unfortunately named constructs in mathematics. Notice that they are neither random nor variables. They are (fixed) functions from the sample space to the real line $\mathbb{R}$. All randomness stems from which element of the sample space is realized by the world – the world is random, **not** our description of it.*

Any probability distribution function P over the sample space S implies a probability distribution for the outcomes of the random variable. This probability distribution of the random variable is fully described by its cumulative distribution function.

Definition 3.7 (Cumulative distribution function). The cumulative distribution function (cdf) of a random variable X is defined by

$$F_X(x) = P(X \leq x) \quad \text{for all } x \quad (6)$$

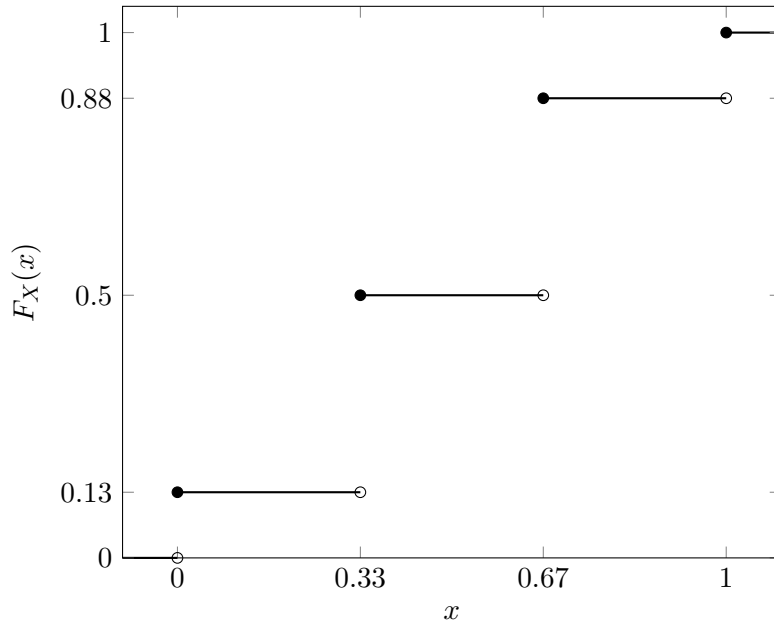


Figure 1: Plot of the cdf $F_X(x)$ in Example 3.3. The black dots indicate right-continuity.

Example 3.3. Let us continue with the survey example from above. For simplicity, we make the widely unrealistic assumption that every survey outcome is equally likely, that is, for all $s \in S$, we have $P(s) = 1/8$. Later in this draft (in Section 3.3.2), we will

discuss a more satisfying model for surveys. In our simplified case, we have

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/8 & \text{if } x \in [0, 1/3) \\ 1/2 & \text{if } x \in [1/3, 2/3) \\ 7/8 & \text{if } x \in [2/3, 1) \\ 1 & \text{if } x \geq 1 \end{cases}.$$

Cumulative distribution functions (or cdfs for short) have three defining characteristics.

Theorem 3.5 (Properties of CDFs). *The function $F(x)$ is a cdf if and only if the following conditions hold*

- $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$.
- $F(x)$ is a nondecreasing function of x .
- $F(x)$ is right-continuous, that is, for every number x_0 , $\lim_{x \downarrow x_0} F(x) = F(x_0)$.

In the above example, the possible outcomes for the proportion of survey participants in support of the policy X had 4 discrete levels. X can only take values from the set $\{0, 1/3, 2/3, 1\}$. This leads the corresponding cdf to be a step function of x – it jumps every time x reaches one of these levels. In this case, we speak of a discrete random variable. When the cdf is a continuous function in x , we speak of X being a continuous random variable.

Definition 3.8. A random variable X is called continuous if $F_X(x)$ is a continuous function of x . A random variable X is called discrete if $F_X(x)$ is a step function of x .

Oftentimes, random variables we work with are not equal but share the same distribution. In this case, we say that they are identically distributed.

Definition 3.9 (Identically distributed random variables). Two random variables X and Y are identically distributed if for every set A , we have $P(X \in A) = P(Y \in A)$.

Example 3.4. For an example of random variables that are not equal but identically distributed, think of tossing a fair coin multiple times. The random variables

$$X = \text{Number of heads observed} \quad Y = \text{Number of tails observed} \quad (7)$$

are identically distributed (convince yourself that this is true). For some points in the sample space, they are unequal, that is, $X(s) \neq Y(s)$ for some $s \in S$.

The notion of identically distributed random variables is central to statistics. We often think of observations as being independent and identical draws from the same probability distribution. This notion of a *sample* (we will discuss it in more depth in Section 3.5) is essential to how we learn about our models of the world from data.

The following theorem states that a random variable's cdf $F_X(x)$ completely determines its probability distribution.

Theorem 3.6. *Two random variables X and Y are identically distributed if and only if $F_X(x) = F_Y(x)$ for every x .*

3.3 Probability Distributions

As statisticians, we should think of probability distributions as models for the world. Some of these models are handcrafted to fit a specific situation perfectly. The *binomial distribution*, for example, puts drawing lottery tickets from an urn into a precise mathematical language. It can also be used to model situations where it doesn't fit quite as perfectly – for example, to model survey outcomes where participants pick between two choices. Another distribution we will encounter, the *normal distribution*, may not as readily reveal what it is a model for. It is, however, as the name implies, ubiquitous in statistics. As we will see, sample means (and other estimators) often tend to have an approximately normal distribution in large samples.

In other more complicated situations, we as statisticians need to build probabilistic models to fit the data that we are interested in. It is our job then too, to see how well these models we came up with actually fit the data. Do they do justice to the parts of the data we care about? Does our model describe the data well enough that we can trust our conclusions?

There are many probability distributions that we won't discuss here but that may nonetheless be very helpful in this endeavor of modeling your data. The important takeaway is that you should think of these distributions as models of the world. Every distribution that you learn about can help you model different kinds of data.

A central property of a probability distribution is its expectation. To better understand the rationale behind expectations, we will travel back in time and watch how Pascal used the notion of an expectation to solve a thorny gambling problem that — I can't help but imagine — caused many fights and cost many friendships. To fully solve the problem, we will then learn about the Binomial distribution.

3.3.1 Expectations and The Problem of Points

The earliest results in probability theory were often sparked by examples from gambling. Maybe this should not come as a surprise, as gambling is quintessential decision-making under uncertainty, and its relation to probability is almost tautological. The successful use of probability in other domains, such as the natural and social sciences, was slower and required many subtle conceptual advances (such as understanding measurement errors to be random).

In 1654, Chevalier de Méré, a French liberal thinker, was convinced that mathematics had no bearing on real-life problems. To score the point, he approached one famous Blaise Pascal with the following problem that he thought had no (mathematical) solution: Suppose you and your friend are playing a friendly game in which, beforehand, you equally contribute to a prize pot (of stake S). You decide on a fixed number of rounds, each of which either of you is equally likely to win. Whoever has won more rounds after the game is over wins the pot. So far so good. While you play, the two of you are rudely

interrupted and stopped from finishing the game. How should you divide the prize pot? Remember, you're friends so the division should be fair. It is immediately clear that a "fair" division is more complicated than just cutting the prize in half. One of you may have been close to winning! Let us give the players names: say Alice is playing against Bob. Alice has won s rounds so far, and Bob has won r . Let us call S_A the amount of the prize pot that we award to Alice.

Let us follow Pascal's reasoning to a solution. Let T be the number of wins required to win the game (if you play 11 rounds, for example, you would have to win 6 of them to win the game). Suppose both Alice and Bob are one win away from winning the game, that is, $s = r = T - 1$, when the game is stopped. How should the pot be split? Well, in this case clearly neither of them could complain if the pot is split evenly as they would have equal chances of winning the game if it was continued. That is, $S_A = 1/2S$.

Now, say Alice was two wins away from winning but Bob still just needs one, that is, $s = T - 2$ and $r = T - 1$. How should the pot be divided now? Pascal's ingenious idea was to imagine the game to be continued for one more hypothetical round. If Alice were to win this hypothetical round, we'd be back to the situation discussed before – where we know how to divide the pot fairly. If Bob was to win, he would win the entire pot. In one of these hypothetical worlds, Alice would lose everything, in the other the game would be interrupted and the pot split evenly. Both scenarios are equally likely. Pascal argues, we should weight these outcomes by their likelihood and award Alice the following split $S_A = \frac{1}{2}0.5S + \frac{1}{2}0 = 0.25S$ – a quarter of the pot. For arbitrary s and r , the split can be calculated by induction.

It is worthwhile to take a step back now and consider why this is considered a big leap in the history of statistics. What Pascal advocates for here is the first use of *expectations*.

Definition 3.10 (Expectation of (discrete) Random Variables). Suppose X is a random variable with finitely many possible outcomes $x_1, \dots, x_K$. Then X has expectation

$$\mathbb{E}[X] = \sum_{k=1}^K P(X = x_k)x_k. \quad (8)$$

The amount of the pot that we should award Alice is her *expected* payoff at the time the game was interrupted! At the moment the game is interrupted (with $s = T - 2$ and $r = T - 1$), Alice has a 25% chance of winning the game and the entire pot. A fair division of the pot due to interruption is to split it according to the expectations of the players at the moment of interruption. Let us make this a bit more formal. Let X be the amount Alice wins if the game would have been played to the end. Because the game was interrupted, X is a random variable. X can take two values: If Alice wins, $X = S$ (she wins the entire pot S). If she loses, $X = 0$ (she doesn't win anything). Her expected winnings depend on her probability of winning the game at the time it's interrupted. Concretely, $\mathbb{E}[X] = P(\text{Alice wins})S + P(\text{Alice loses})0$.¹¹ Pascal's idea was to split the pot according to this expectation, $S_A = \mathbb{E}[X]$.

¹¹The symbol $\mathbb{E}$ is often used to denote expectations. Using just E is also common.

In the scenario discussed above in which Alice is two wins away from winning the entire game and Bob only one, we have $\mathbb{E}[X] = P(\text{Alice wins})S + P(\text{Alice loses})0 = 0.25S$. Exactly the amount we reasoned it should be intuitively.

Expectations are, to this day, used to figure out how to value uncertain future outcomes. On a more conceptual level, the justification for the use of expectations depends on our interpretation of probabilities. If we follow a frequentist interpretation of probabilities, then the expectation above is the amount Alice would win (on average), if we would restart the game after it has been interrupted and would play from there (infinitely) many times. This seems a reasonable basis for deciding on how to split the pot if Alice and Bob (or whoever sets the rule) play many, many of these games – a casino, maybe. What if Alice and Bob just happen to play once? In this case, subjective probabilities may be a more helpful perspective. From this point of view, the question would be: how much should Alice be willing to pay to restart the game? Her expectation of winning the game is the answer to this question.

The important takeaway is that expectations are often a useful tool for decision-making when outcomes are uncertain.

3.3.2 The Binomial Distribution

We still owe de Méré a full solution (don't worry, Pascal provided one). In the previous section, we focused only on the situation where $s = T - 2$ and $r = T - 1$, but how do we deal with arbitrary s and r ? The answer will come from learning about the Binomial distribution.

First, we need to take a step back. In every round of the game described above, both Alice and Bob are equally likely to win. We can think of them tossing a coin, for example. The *Bernoulli* distribution is the probabilistic model for coin tosses. Suppose X is a random variable with outcomes 1 (for “heads”) and 0 (for “tails”). If we denote the coin's probability to come out “heads” with p , the random variable X is said to be Bernoulli distributed with probability p (in the above example, the example of a fair coin, we have $p = 1/2$). In mathematical notation, we write $P(X = 1) = p$ and $P(X = 0) = 1 - p$.

Definition 3.11 (Bernoulli distribution). A random variable X taking values 0 or 1 with

$$P(X = x) = p^x(1 - p)^{1-x}, \quad (9)$$

where $x \in \{0, 1\}$, is said to be Bernoulli(p) distributed.

As a shorthand to say that a random variable X is Bernoulli distributed (with probability p), we write $X \sim \text{Bernoulli}(p)$.

Now suppose we toss a coin multiple times, say N times. Let X now be the number of “heads” you see within these N tosses. The model for this situation is the Binomial distribution. What is the probability that the coin landed “heads” (exactly) k times? Any sequence with k (out of N) heads has probability $p^k(1 - p)^{N-k}$. The number of sequences with k “heads” is given by the binomial coefficient $\binom{N}{k}$. Putting

things together, we find that the probability of k “heads” out of N coin tosses is $P(X = k) = \binom{N}{k} p^k (1 - p)^{N-k}$.

Definition 3.12 (Binomial Distribution). A random variable X with distribution

$$P(X = k) = \binom{N}{k} p^k (1 - p)^{N-k} \quad (10)$$

for $k \in \{0, \dots, N\}$ is said to be Binomial(N, p) distributed and we write $X \sim \text{Binomial}(N, p)$.

We are now in a position to fully respond to de Méré’s challenge: Suppose Alice has won s and Bob has won r games when the game is interrupted. Both s and r are now arbitrary. To win, Alice would need to win another $n = T - s$ (hypothetical) rounds, and Bob would need to win $m = T - r$ rounds. The probability that Alice would win if the game was continued is equal to her winning at least n out of $n + m - 1$ rounds. The probability of winning (exactly) k out of $n + m - 1$ rounds is $\binom{n+m-1}{k} 0.5^k (1 - 0.5)^{n+m-1-k}$. Her chance of winning the entire game if it was continued is thus

$$P(\text{Alice wins}) = \sum_{k=n}^{n+m-1} \binom{n+m-1}{k} 0.5^k (1 - 0.5)^{n+m-1-k}$$

and we should award her $S_A = P(\text{Alice wins})S = S \sum_{k=n}^{n+m-1} \binom{n+m-1}{k} 0.5^k (1 - 0.5)^{n+m-1-k}$.

Example 3.5 (Modeling a survey). Think about how the Binomial distribution may help us when modeling a survey. Why is the Binomial distribution a better model for surveys than the naive uniform distribution over all survey outcomes we’ve assumed in a previous section?

3.3.3 From Discrete to Continuous Distributions

A binomial random variable is an example of a discrete random variable – a random variable that only takes a finite number of values. Say, for example, $X \sim \text{Binom}(N, p)$: then X can only take on numbers in the set $\{0, \dots, N\}$. The probability $P(X = x)$ as a function of x is called the probability mass function.

Definition 3.13 (Probability Mass Function). The probability mass function (pmf) for a discrete random variable X is given by

$$f_X(x) = P(X = x). \quad (11)$$

The probability mass function describes how much “mass” of the overall probability distribution is carried by each possible value of x . Now consider we keep increasing N : more and more values for X become possible and the “mass” carried by each individual value x becomes smaller and smaller. If we send N all the way to infinity and allow

(uncountably)¹² infinitely many values, we arrive at a continuous random variable – a random variable that can take on (uncountably) infinitely many values.

Let us stick to the metaphor of “mass”. Imagine a block of iron with total mass of 1kg (or 1 pound, if you are so inclined). Now suppose you subdivide the block into four (not necessarily) equal sub-blocks. Each of these sub-blocks has a specific mass and all of them together add up back to 1kg. Now imagine you keep dividing the blocks into ever finer and finer sub-blocks. With every division, the mass of every sub-block becomes smaller but all of them keep adding up to 1kg. In the limit of smaller and smaller blocks, the concept of “weight” of each sub-block becomes useless – the weight of each sub-block will tend towards 0 with its size decreasing. Instead, it is much more enlightening to think about density. To see why, imagine that you switch out one of the four original sub-blocks with feathers (making sure to keep the overall weight of all four blocks fixed at 1kg). This sub-block of feathers carries less mass than the other blocks. In the limit of ever finer subdivisions, this distinction between feathers and iron gets lost, however, if we talk about mass. Focusing on density, we can distinguish between the parts of the block that are iron (high density) and the part that is feathers (low density) even in the limit of infinitesimal divisions: Let $\rho(x)$ be the density of the block at position x . We can then integrate over any volume A to obtain the mass of this volume $M(A) = \int_A \rho(x)dx$. Volumes of the same size in the parts of the block that are feathers will carry less weight than those that are filled with iron.

For continuous random variables, the notion of a probability density is similarly central.

Definition 3.14 (Probability Density Function). The probability density function (pdf) $f_X(x)$ for a continuous random variable X is given by the function that satisfies

$$F_X(x) = \int_{-\infty}^x f_X(t)dt \quad \text{for all } x. \quad (12)$$

The probability (mass) of any specific value x of a continuous random variable X is zero, that is, $P(X = x) = 0$. Its probability density is more useful: it tells us how likely X is to fall into any interval $A = [a, b]$. Given its pdf $f_X(x)$, we have $P(X \in A) = \int_a^b f_X(x)dx$.

The following corollary (an application of the fundamental theorem of calculus) gives a simple relationship between continuous pdfs and cdfs.

Corollary 3.1. *In the case of a continuous pdf, we have the further relationship*

$$\frac{d}{dx}F_X(x) = f_X(x). \quad (13)$$

¹²Why do we specify “uncountably” infinitely many here? It turns out that step-functions can have countably many jump-discontinuities on $\mathbb{R}$. The case of countably infinitely many values is thus still covered by our definition of a discrete random variable. For uncountably infinitely many values, think of the random variable being able to take on all values in an interval $[a, b]$, we need the notion of a continuous random variable. Notice that the definition of a continuous random variable excludes cases in which specific values take on positive probability, that is, we require $P(X = x) = 0$ for all x .

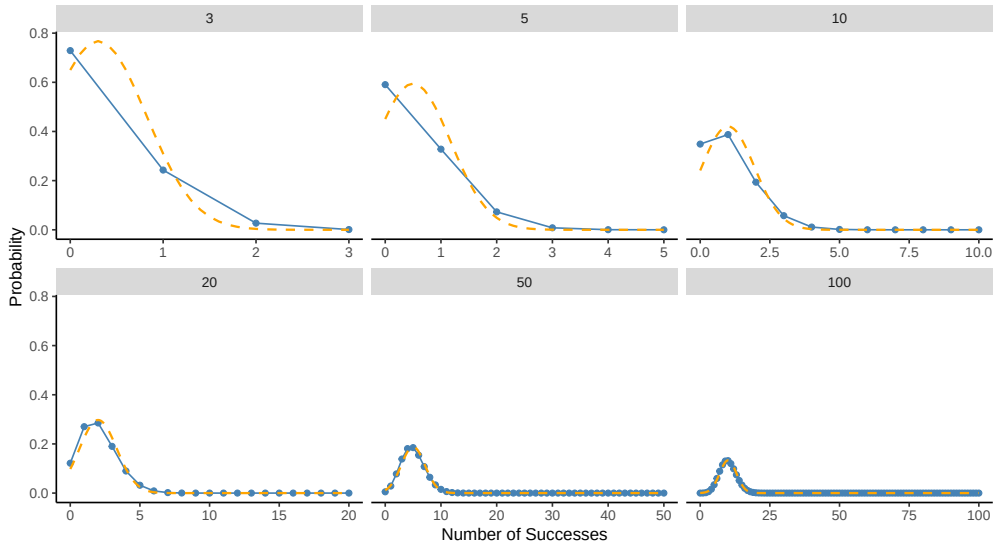


Figure 2: The Normal distribution as an approximation of the Binomial distribution when the number of draws N grows large.

The following theorem lists two essential properties of pdfs and pmfs:

Theorem 3.7 (Properties of pdfs/pmfs). *A function $f_X(x)$ is a pdf (or pmf) of a random variable X if and only if*

1. $f_X(x) \geq 0$ for all x .
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$ for pdfs (or $\sum_x f_X(x) = 1$ for pmfs).

3.3.4 The Normal Distribution

The Normal distribution is ubiquitous in statistics and we will see some of its uses throughout these notes. Its first use is often attributed to de Moivre, who in 1738 was seeking to approximate the Binomial distribution when the number of draws N tends to infinity [Stigler, 1990]. Figure 2 illustrates that as the number of draws N grows larger, the Normal distribution (orange) becomes a better and better approximation. In 1823, Gauss was the first to posit the normal distribution as a probability distribution in its own right. He managed to show that, assuming measurement error followed this normal distribution, the average of multiple observations is the best guess at the true value of what was measured (see Section 4.1.3). In his honor, the Normal distribution is also often called the *Gaussian* distribution. Laplace would cement the central role of the normal distribution by proving that (under very weak assumptions) the deviations of an average from its mean asymptotically tend to be normally distributed – this is the famous central limit theorem, which we will discuss in detail in Section 3.6. In 1860, Maxwell established an early example of the occurrence of the normal distribution in

natural phenomena – due to the inherent symmetry of the motion of molecules in a gas, their velocities can be found to be normally distributed Maxwell [1860].

A random variable is said to be normally distributed if it has the following pdf:

Definition 3.15 (Normal Distribution). A random variable X with pdf

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (14)$$

is said to be normally distributed with mean μ and standard deviation σ . We write $X \sim \mathcal{N}(\mu, \sigma)$.

We have already seen the use of expectations for discrete random variables. We will now define the expectation of a continuous random variable.

Definition 3.16 (Expectation of (continuous) Random Variables). Suppose X is a continuous random variable with probability density $f(x)$, then its expectation is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx. \quad (15)$$

Example 3.6 (Mean of a Normally Distributed RV). Let X be normally distributed with mean μ and standard deviation σ . That is, $X \sim \mathcal{N}(\mu, \sigma)$. The expectation of X is given by its mean:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu.$$

The next theorem gives an essential property of expectations:

Theorem 3.8 (Linearity of Expectation). *Let X and Y be random variables and let a, b and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist, we have*

1. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
2. $\mathbb{E}[ag_1(X) + bg_2(X) + c] = a\mathbb{E}[g_1(X)] + b\mathbb{E}[g_2(X)] + c$
3. *If $g_1(x) \geq g_2(x)$ for all x , then $\mathbb{E}[g_1(X)] \geq \mathbb{E}[g_2(X)]$.*

3.3.5 The Importance of Variance

Another important property of a random variable beyond its expectation is its *variance*.

Definition 3.17 (Variance). The variance of a random variable X is defined as the expectation of its squared distance from its mean $\mu = \mathbb{E}[X]$, that is

$$\text{Var}(X) = \mathbb{E}[|X - \mu|^2]. \quad (16)$$

The standard deviation of X is defined as $\text{sd}(X) = \sqrt{\text{Var}(X)}$. We often denote the variance of a random variable with $\sigma^2 \equiv \text{Var}(X)$ and its standard deviation with $\sigma \equiv \text{sd}(X)$.

Example 3.7 (Variance of a normally distributed RV). Let $X \sim \mathcal{N}(\mu, \sigma)$. The variance of X is given by the square of its standard deviation (that is, σ^2)

$$\text{Var}(X) = \mathbb{E}[|X - \mu|^2] = \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2.$$

The next theorem gives two important properties of the variance of random variables.

Theorem 3.9. *Let X, Y be two independent random variables with finite variance and let $\mu = \mathbb{E}[X]$. Then*

1. $\text{Var}(X) = \mathbb{E}[|X - \mu|^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
2. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.

Variance is a measure of dispersion: It captures how spread out possible values of a random variable are around its expected value $\mu = \mathbb{E}[X]$. In particular, we often want to quantify how likely values of a random variable X are to fall outside intervals around its expected value. Probabilities of the form $P(|X - \mu| \geq k)$ (where $k > 0$) are a natural quantification of our uncertainty about X .

Example 3.8 (68–95–99.7 rule). For normally distributed random variables, the 68–95–99.7 rule helps to memorize how likely the random variable is to stray away from its mean. Let $X \sim \mathcal{N}(\mu, \sigma)$, then

$$\begin{aligned} P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) &\approx 68\% \\ P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &\approx 95\% \\ P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &\approx 99.7\% \end{aligned}$$

We see that we should expect 32% of draws to fall outside the 1σ interval around the mean (that is, $[\mu - \sigma, \mu + \sigma]$), but only 0.3% of draws to stray more than 3 standard deviations away from the mean.

For general distributions, there is no such neat rule exactly quantifying our uncertainty based on the distribution's variance. Chebycheff's inequality, however, establishes an upper bound on these probabilities of a random variable to stray away from its mean in terms of the distribution's variance:

Theorem 3.10 (Chebycheff's Inequality). *Let X be a random variable with finite variance $\sigma^2 < \infty$ and mean $\mu = \mathbb{E}[X]$. Then for any real number $k > 0$,*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \tag{17}$$

Notice that Chebycheff's inequality does not give us exact probabilities the same way the 68–95–99.7 rule does for the normal distribution. It shows us, however, that for all distributions with finite variance, we can bound our uncertainty about a random variable based on its variance.

3.4 Joint and Conditional Distributions

Oftentimes, we want to talk about multiple random variables and their relationship. In the next section, for example, we want to discuss the properties of a sample of multiple draws from a probability distribution. To do so, we need to introduce the notion of the joint distribution of multiple random variables. First, we extend the notion of a random variable, which maps the sample space to $\mathbb{R}$, to a random vector, which maps the sample space to vectors:

Definition 3.18. An n -dimensional random vector is a function from a sample space S into n -dimensional Euclidean space $\mathbb{R}^n$.

Similarly to a random variable, a random vector can also be described by a pmf or a pdf depending on whether its cdf is discrete or continuous. When dealing with a random vector, we not only have to specify the distribution of its components separately but also their relation. We begin with the discrete case:

Definition 3.19 (Joint pmf). Let (X, Y) be a discrete random vector. We then call the function $f(x, y)$ from $\mathbb{R}^2$ into $\mathbb{R}$ defined by $f(x, y) = P(X = x, Y = y)$ the joint pmf of X and Y .

Definition 3.20 (Expectation of discrete random vector). Let (X, Y) be a random vector with pmf $f(x, y)$. Then for any function $g(x, y)$ (defined for all possible values of (X, Y)), the expectation of $g(X, Y)$ is given by

$$\mathbb{E}[g(X, Y)] = \sum_{(x, y) \in \mathbb{R}^2} g(x, y) f(x, y) \quad (18)$$

While it is not enough to specify the distributions of the components of a random vector, the joint distribution fully determines these *marginal* distributions of the components of the vector:

Theorem 3.11. Let (X, Y) be a discrete bivariate random vector with joint pmf $f(x, y)$. Then the marginal pmfs of X and Y , $f_X(x) = P(X = x)$ and $f_Y(y) = P(Y = y)$, are given by

$$f_X(x) = \sum_{y \in \mathbb{R}} f(x, y) \quad \text{and} \quad f_Y(y) = \sum_{x \in \mathbb{R}} f(x, y). \quad (19)$$

Similarly to the discrete case just discussed, we can also extend the notion of a pdf to a (bivariate) random vector.

Definition 3.21 (Joint pdf). A function $f(x, y)$ from $\mathbb{R}^2$ into $\mathbb{R}$ is called the joint pdf of the continuous random vector (X, Y) if for every $A \subset \mathbb{R}^2$,

$$P((X, Y) \in A) = \int_A \int f(x, y) dx dy. \quad (20)$$

Definition 3.22 (Expectation of continuous random vector). Let (X, Y) be a random vector with pdf $f(x, y)$, then for any real-valued function $g(x, y)$, the expectation value of $g(X, Y)$ is given by

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy \quad (21)$$

Theorem 3.12. Let (X, Y) be a continuous bivariate random vector with joint pdf $f(x, y)$. Then the marginal pdfs of X and Y , $f_X(x)$ and $f_Y(y)$, are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad -\infty < x < \infty \quad (22)$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad -\infty < y < \infty \quad (23)$$

The joint distribution of a (bivariate) random vector implies the conditional distribution of one of its components given its other. The notion of a conditional pdf (or pmf) for a random variable is similar to the notion of conditional probabilities for events discussed in Section 3.2.2.

Definition 3.23. Let (X, Y) be a bivariate random vector with joint pdf (or pmf) $f(x, y)$ and marginals $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$, the conditional pdf (or pmf) of Y given that $X = x$ is

$$f(y|x) = \frac{f(x, y)}{f_X(x)}. \quad (24)$$

Two random variables are independent when their pdf (or pmf) factorizes.

Definition 3.24 (Independence of Random Variables). Let (X, Y) be a bivariate random vector with joint pdf (or pmf) $f(x, y)$ and marginals $f_X(x)$ and $f_Y(y)$. Then X and Y are called independent if for every $x, y \in \mathbb{R}$, we have

$$f(x, y) = f_X(x) f_Y(y) \quad (25)$$

The following theorem gives important properties of independent random variables.

Theorem 3.13. Let X and Y be independent random variables. Then

1. For any $A, B \subset \mathbb{R}$, we have $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$, that is, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent events.
2. Let $g(x)$ be a function of only x and $h(y)$ of only y . Then

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]. \quad (26)$$

3.5 Sampling and the Distribution of the Sample Mean

[T]he process of the reduction of data is, even in the simplest cases, performed by interpreting the available observations as a sample from a hypothetical infinite population – Fisher [1922, p. 313]

In probability theory, we imagine our pdf $f(x)$ to represent an infinite population. Think about an urn containing infinitely many balls, each with a number inscribed on it. You can then think of the pdf $f(x)$ as the frequency with which a number x appears in the urn. From this infinite population, we now randomly draw a bunch of samples. Think of randomly pulling balls out of the urn; because the urn contains infinitely many balls, we do not have to worry about removing balls changing the urn’s constitution. We can think of these samples as random variables $X_1, \dots, X_n$ all being independent and with the same distribution $f(x)$. One of the main tasks of statistics is learning about properties of $f(x)$, say its mean, from the random sample $X_1, \dots, X_n$.

Remark. *You may wonder how this urn containing infinite balls relates to the real world. This is a valid question. Most discussions of probability theory just posit a random sample $\{X_i\}_{i=1}^n \stackrel{iid.}{\sim} f(x)$. In real world applications, the logic is different. There is an ebb and flow between observations and a model. As applied statisticians, it is beneficial to think of the pdf $f(x)$ as a model for your observations: You collect some observations and then through modeling work, you come up with a probabilistic model $f(x)$ that hopefully adequately describes your data. The data can then be thought of as a random sample from the model $f(x)$ and we infer its parameters. If the model doesn’t fit the data, we need to go back to the drawing board.*¹³

To rigorously define what we mean by a random sample, we need to extend the definitions of the previous section from the bivariate to the general case.

Definition 3.25. If $X_1, \dots, X_n$ are all discrete random variables, then their joint pmf is

$$f(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n). \quad (27)$$

If all the variables are continuous, then their joint pdf $f(x_1, \dots, x_n)$ fulfills

$$P((X_1, \dots, X_n) \in A) = \int_A f(x_1, \dots, x_n) dx_1 \dots dx_n \quad \text{for all } A \subset \mathbb{R}^n. \quad (28)$$

When dealing with n random variables, we define (mutual) independence as follows.

Definition 3.26 (Mutual Independence). Random variables $X_1, \dots, X_n$ are called mutually independent if their joint pdf (or pmf) factorizes

$$f(x_1, \dots, x_n) = f_{X_1}(x_1) \dots f_{X_n}(x_n) \quad (29)$$

¹³This holistic perspective on modeling that incorporates building, fitting, checking and adapting a model can, for example, be found in recent work on Bayesian Workflow [Gelman et al., 2020].

Remark. *It is important to note that pairwise independence between all random variables does **not** imply mutual independence (the reverse is true, however). Mutual independence plays a more important role and the modifier “mutual” is often dropped – independence of random variables usually refers to mutual independence.*

Equipped with the notion of (mutual) independence, we can now give a mathematical definition of a random sample.

Definition 3.27 (Random sample). Random variables $X_1, \dots, X_n$ are called a random sample of size n from the population $f(x)$ if $X_1, \dots, X_n$ are mutually independent and identically distributed random variables with marginals $f(x)$. Alternatively, we say $X_1, \dots, X_n$ are independent and identically (often abbreviated to iid.) distributed with pdf (or pmf) $f(x)$.

Two essential sample statistics are the sample mean and the sample variance:

Definition 3.28 (Sample mean). The sample mean $\bar{X}$ is defined as the arithmetic average of the values of a collection of random variables $X_1, \dots, X_n$, that is,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i. \quad (30)$$

Definition 3.29 (Sample Variance). The sample variance of a collection of random variables $X_1, \dots, X_n$, is defined as

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2. \quad (31)$$

The sample standard deviation is $S = \sqrt{S^2}$.

The following theorem shows that the sample mean and the sample variance are unbiased under very weak assumptions. It also provides the variance of the sample mean.

Theorem 3.14 (Properties of Sample Mean and Variance). *Let $X_1, \dots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then*

1. $\mathbb{E}[\bar{X}] = \mu$,
2. $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$
3. $\mathbb{E}[S^2] = \sigma^2$.

In the important case that our random sample is drawn from a normal distribution, the following theorem gives the exact distribution of the sample mean.

Theorem 3.15. *Let $X_1, \dots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma)$, then*

1. $\bar{X}$ and S^2 are independent,
2. $\bar{X}$ has $\mathcal{N}(\mu, \sigma/\sqrt{n})$ distribution.

The scaling of the standard error with $1/\sqrt{n}$ together with Chebycheff's Inequality (Theorem 3.10) tells us something quite non-intuitive. The “information” contained in, let's say, the first 10 samples is greater than the information contained in the next 10 samples. To double the precision of our knowledge about μ , we must not collect double the amount of data – as one might naively suspect – but quadruple it.

3.6 Large Samples: Asymptotics and the Central Limit Theorem

Theoretical statistics can be very challenging. In the previous section, we have described properties of the sample mean – one of the simplest sample statistics. While we showed that it's unbiased under very weak assumptions, we were only able to provide its exact distribution for the case of a random sample from a normal distribution. You may wonder: what happens if we are concerned with a random sample from the Binomial distribution?

Questions like these can be extremely hard to answer but we have a very powerful tool at our disposal: asymptotics. We often study the behavior of statistical quantities as the sample size grows to infinity, $n \rightarrow \infty$. These results are often significantly easier to obtain than exact calculations and serve as great approximations in large samples. One of the most famous of these “approximations” is the *central limit theorem* (CLT). It's referred to as “central” because of its importance for probability theory¹⁴ The central limit theorem answers the question we began this section with: it gives an asymptotic approximation of the distribution of the (centered) sample mean under very general conditions.

Before we dive into the details of the central limit theorem, we will discuss another famous theorem, the law of large numbers. When we discussed the interpretations of probability, we mentioned that relative frequencies tend towards the “true probability”. This is the central pillar of the frequentist interpretation of probabilities that associates limiting frequencies with probabilities. We discussed that the ratio of “heads” in a sequence of fair coin tosses will tend to $1/2$. We noticed, however, that there are possible sequences (such as only heads) that do not tend towards the “right” probability. These possible sequences are weird in the sense that they are highly unlikely, we would never expect to flip only heads for all eternity when flipping a fair coin.

Exactly these insights are generalized and formalized with the law of large numbers:

Theorem 3.16 ((Strong) Law of Large Numbers). *Let $X_1, X_2, \dots$ be iid. random variables with mean $\mathbb{E}[X_i] = \mu$ and variance $\text{Var}(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, then for every $\varepsilon > 0$,*

$$P\left(\lim_{n \rightarrow \infty} |\bar{X}_n - \mu| < \varepsilon\right) = 1. \quad (32)$$

¹⁴At least George Pólya interpreted “central” in this way. The mathematician Lucien Le Cam interprets it in the sense that “it describes the behaviour of the centre of the distribution as opposed to its tails”. No one disagrees, however, about the central place the CLT occupies in modern statistics.

We say $\bar{X}_n$ converges to μ almost surely.

With asymptotics we aim to develop useful approximations for large samples. We often want to approximate complicated distributions; to do so, we need the concept of convergence in distribution.

Definition 3.30 (Convergence in Distribution). A sequence of random variables $X_1, X_2, \dots$ is said to converge in distribution to a random variable X if

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x) \quad (33)$$

at all points x where $F_X(x)$ is continuous. We then write $X_n \xrightarrow{d} X$.

We are now ready to return to the central limit theorem, one of the most impressive results of probability theory. In colloquial terms, the central limit theorem shows that many random causes, in their individuality completely chaotic, can on aggregate create a new kind of statistical order. It is hard to match Galton's poetic description of the theorem:

I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error" [Galton is referring to the CLT]. The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along. – Galton [1889, p. 66]

Equipped with this definition of convergence in distribution, we can formally state the central limit theorem:

Theorem 3.17 (Central Limit Theorem). Let $X_1, X_2, \dots$ be a sequence of iid. random variables with mean $\mathbb{E}[X_i] = \mu$ and variance $\text{Var}(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then

$$\lim_{n \rightarrow \infty} G_n(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \quad (34)$$

for all $x \in \mathbb{R}$. We often write $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} \mathcal{N}(0, 1)$.

A first version of the central limit theorem, one for Binomial random variables, was developed by de Moivre in 1733. While we now recognize this as a first step towards one of the most central results in probability theory, it took almost a hundred years until Laplace rescued it from obscurity in 1812 and developed it into the result we know today.

In doing so, Laplace not only made great mathematical strides but also helped illuminate the heavens. In his lifetime, astronomers were beginning to understand the intricate dynamics of our solar system. Some of the questions they asked had deep philosophical implications: Believing the universe to be eternal, they began to worry if the heavens—no longer the sphere of the divine but a deterministic mechanism precisely following Newton’s laws—were stable. It was already well known that the three-body problem (the gravitational dynamics of three planets) had chaotic solutions; a pressing question arose: “Will the solar system fall apart”?

In an effort to answer this question, many astronomers began to rigorously measure positions of celestial objects. But how was one to combine these measurements? This was a question that vexed some of the greatest minds, including Laplace and Gauss. Laplace was set to solve this problem, and his central limit theorem played a crucial role. We will discuss his solution in Section 4.1.3.

Remark. *You may wonder at what sample size n a sample is large enough to warrant applying the central limit theorem – or any other asymptotic results. This is not an easy question. As mentioned, asymptotics are used as approximations; they never hold exactly because we never have infinitely many observations. What’s worse is that many results, such as the central limit theorem, are limit results. Applying them for any finite sample is, in a sense, a logical fallacy [van der Vaart, 1998]. No matter how large you choose n , a sequence may exhibit completely arbitrary behavior for its first n elements even if it eventually converges to a limit. For any finite n , limit results make absolutely no guarantees. Sometimes, like in the case of the central limit theorem, there are theoretical results that give finite sample guarantees: The Berry-Esseen theorem bounds the approximation error of the CLT in finite samples. Often, we use asymptotics without such theoretical results that guarantee the adequacy of our approximation. In these cases—and this is a general rule whenever you want to use asymptotic results—we rely on simulations to check whether asymptotic results are good approximations for the specific sample size we are dealing with.*

Example 3.9. We have discussed before that the Binomial distribution can asymptotically be approximated by a normal distribution. Convince yourself that the central limit theorem gives another perspective on why this is the case.

4 Estimation and Uncertainty Quantification

In the previous section, we have encountered the binomial and the normal distribution as important probabilistic models of the world. An important question, and one that has vexed the most brilliant mathematicians working on the eve of probability theory (from Bernoulli to Fisher), is how to infer their parameters from data. Before we dive into these questions, it is helpful to clarify some terminology.

Definition 4.1 (Statistic). Let $X_1, \dots, X_n$ be a random sample from some population and let $T(x_1, \dots, x_n)$ be any function whose domain includes the sample space

of $(X_1, \dots, X_n)$. The random quantity $T = T(X_1, \dots, X_n)$ is called a statistic. The probability distribution of $T = T(X_1, \dots, X_n)$ is called its sampling distribution.

In inference, our goal is to find statistics that are “good” estimates of the population parameter of interest. A statistic’s sampling distribution tells us how much the specific statistic would vary if we had seen other realizations of the data. In a frequentist setting, it is thus a natural basis for uncertainty quantification.

We call the population parameter that we aim to learn about the *estimand*. The statistic we use to estimate the population parameter, we call an *estimator*. The estimator is a random variable; its value evaluated at the observed sample point is called an *estimate*.

There are many estimation strategies (M-estimation, empirical risk minimization, ...). Here, we will only discuss the two most prevalent and, arguably, most important estimation techniques: Bayesian inference and maximum likelihood.

4.1 Bayesian Inference

Thomas Bayes, who lends his name to one of statistics’ most famous theorems (you’ve guessed it: Bayes’ rule), wrote an essay entitled *An Essay Towards Solving a Problem in the Doctrine of Chances* [Bayes, 1763].¹⁵ While it is Bayes’ only paper on statistics and one of the first applications of what we nowadays recognize as Bayesian inference, it remained relatively obscure for a long time. Instead of pondering the implications this may have for the importance of catchy or enlightening titles, we will see what we can learn from Bayes about how to do inference.¹⁶ He sets out to solve the following problem:

Problem

Given the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named. [Bayes, 1763, p. 376]

Clearly, Bayes was considering the same problem we find ourselves before today.

4.1.1 Bayes and the Pool Table

In *An Essay Towards Solving a Problem in the Doctrine of Chances*, Bayes asks us to imagine a table – he wasn’t as specific but feel free to imagine a billiard table – onto which two balls are rolled. Ball one is rolled once and is equally likely to end up anywhere on the table. We are (for simplicity) only interested in its position along the left-right dimension of the table, call this position of ball one θ . For simplicity, let $\theta \in [0, 1]$. If

¹⁵It was read to the Royal Academy of Science in 1763 after Bayes’ death.

¹⁶A brief warning to the historically interested reader is in order: While we will follow Bayes’ example, we will do so with modern statistics in mind. While this may misrepresent Bayes’ original ideas, I hope that with the knowledge of hindsight, this exposition will be more pedagogical.

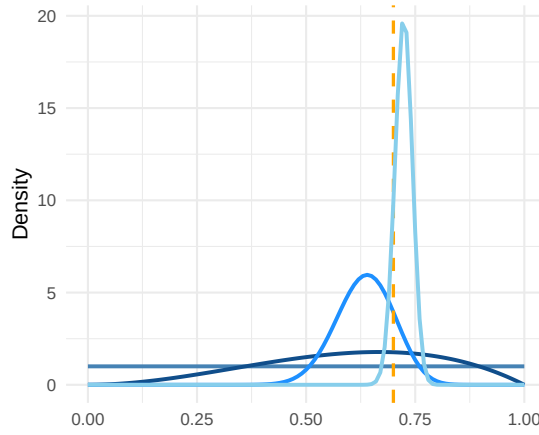


Figure 3: Sequence of Posteriors for Bayes' example. The more data we accrue, the more the posterior concentrates.

$\theta = 0$, ball one landed completely on the left of the table. For $\theta = 1$, on the other hand, the ball landed on the right-most part of the table.

We now repeatedly throw a second ball and record whether it has landed to the left or right of the first ball. We call these observations $Y_1, \dots, Y_n \in \{0, 1\}$ (where we record a 1 if the second ball lands to the left of ball one and 0 if it lands on its right.) What can we say about the position of the first ball (and remember this position is unknown, we can't just peek and see where it is) based on the observations $Y_1, \dots, Y_n$? Following Bayes' problem statement, in modern notation, we are after the probability distribution $p(\theta|Y_1, \dots, Y_n)$. This is a big conceptual leap!

In our discussion of probability distributions, we considered the parameters of our distributions θ (such as p in the binomial distribution and μ in the normal distribution) to be fixed numbers. We have then calculated probabilities of random draws y from this distribution $p_\theta(y)$. This likelihood, $p_\theta(y)$, tells us how probable our observations are for any fixed value of θ . Bayes argues that we can turn this argument on its head. We now consider the unknown θ to be random. Our goal is to calculate the probability density of θ given the observation y we've made, that is, we want to calculate $p(\theta|y)$.

The probability $p(\theta|y)$ is called the *posterior* and it can be calculated with Bayes' rule:

Theorem 4.1 (Bayes' rule). *Let θ and $y_1, \dots, y_n$ be random variables, then*

$$p(\theta|y_1, \dots, y_n) = \frac{p(y_1, \dots, y_n|\theta)p(\theta)}{p(y_1, \dots, y_n)}. \quad (35)$$

Bayes' rule has three ingredients:

- The likelihood $p(y_1, \dots, y_n|\theta)$. This is our (probabilistic) model for our data y and this model contains the parameter θ which we want to learn about.
- $p(y_1, \dots, y_n) = \int p(y_1, \dots, y_n|\theta)p(\theta)d\theta$ is called the evidence and it normalizes the numerator so that the posterior is a proper probability distribution. $p(y_1, \dots, y_n)$

can be hard to compute. In these notes, we will only consider simple examples for which the posterior has closed-form solutions. In general, computational methods (such as MCMC) are used to approximate the posterior.

- $p(\theta)$ is the prior. This is the probability that we assign θ *before* we see the data. Bayes cleverly chose an example in which the prior has a physical interpretation. Because we throw the first ball uniformly, it is easy to justify why $p(\theta)$ should be uniform and to explain the meaning of the prior. This is often more tricky. In most situations, the parameter is not thought of as being literally randomly drawn from an objective distribution (that we then can use as a prior). Instead, priors are often interpreted from the perspective of subjective probabilities: The prior is our (subjective) belief about θ before we see the observations y . This could be a state of absolute ignorance, say, we know nothing whatsoever about θ . One of the strengths of Bayesian inference is, however, that it can also incorporate any kind of knowledge about θ that we may have beforehand. Think, for example, about knowledge we may have from earlier measurements or theoretical constraints.

Let us follow Bayes' footsteps. We have rolled the first ball onto the table in such a way that it was equally likely to come to a halt anywhere. We thus have $p(\theta) = \text{Unif}[0, 1]$ for the prior distribution of ball one's location θ – before observing any of the other throws (remember, no peeking at the first ball's position allowed), ball one might be anywhere on the table with equal probability.¹⁷

We have made n observations, $Y_1, \dots, Y_n$, by repeatedly rolling a second ball that lands either to the left (we then record $Y_i = 1$) or to the right ($Y_i = 0$) of ball one. Where the individual ball throws fall is irrelevant to determine the original ball's position θ , we are only interested in the total number of balls that (out of the n throws) landed on the original ball's left. That is, $Y = Y_1 + \dots + Y_n$. Notice that $Y \sim \text{Binom}(n, \theta)$ is binomially distributed with parameter θ . This determines the likelihood $p(Y = k|\theta)$ of the data.

Lastly, we need to evaluate the denominator, that is, the evidence $p(Y = k)$:¹⁸

$$\begin{aligned} p(Y = k) &= \int_0^1 \binom{n}{k} \theta^k (1 - \theta)^{n-k} d\theta \\ &= \frac{1}{n+1}. \end{aligned} \tag{36}$$

With this, we have all the ingredients to apply Bayes' rule:

$$\begin{aligned} p(\theta|Y = k) &= \frac{p(Y = k|\theta)p(\theta)}{p(Y = k)} \\ &= (n+1) \binom{n}{k} \theta^k (1 - \theta)^{n-k}. \end{aligned} \tag{37}$$

¹⁷We are actually side-stepping some of Bayes' subtle reasoning about the prior here. We will discuss his reasoning and problems with priors on completely unknown quantities in Section 4.1.5.

¹⁸To see why this is true, you can look up the Beta-function.

This is a so-called *Beta-distribution*. The posterior mean is given by $\mathbb{E}[\theta|Y = k] = \frac{k+1}{n+2}$.¹⁹ If all throws of the second ball land to the right of ball one, that is, $k = 0$, our best guess for the original ball's position is $\mathbb{E}[\theta|Y = 0] = \frac{1}{n+2}$: we guess that it's almost all the way on the left of the table. If, on the other hand, all throws land to the first ball's left (that is, $k = n$), then we have $\mathbb{E}[\theta|Y = n] = \frac{n+1}{n+2}$ – the original ball is most likely almost all the way on the right of the table.

4.1.2 Bayesian Uncertainty Quantification: Posterior credible interval

In his problem statement, Bayes did not only set out to find a method to determine the best guess at a parameter θ given observations, he specifically aimed to also quantify our uncertainty about θ . Calculating “the chance that the probability of [the parameter] lies somewhere between any two degrees of probability” can easily be obtained once we have computed the posterior probability. Continuing our example from the previous example, say, we want to know how probable it is to find θ between θ_1 and θ_2 having observed $Y = k$, then we can easily do so by integrating the posterior:

$$\begin{aligned} P(\theta_1 \leq \theta \leq \theta_2 | Y = k) &= \int_{\theta_1}^{\theta_2} p(\theta | Y = k) d\theta \\ &= \int_{\theta_1}^{\theta_2} (n+1) \binom{n}{k} \theta^k (1-\theta)^{n-k} d\theta. \end{aligned}$$

This integral, known as an incomplete beta integral, has no general closed-form solution, which presented some difficulties for Bayes. Nowadays, with advances in numerical methods and the use of computers, evaluating integrals of this type is straightforward.

Being able to evaluate probabilities of this kind, we can now give our uncertainty about θ by reporting posterior *credible intervals*. An interval $[\theta_1, \theta_2]$ is a γ -credible interval if the probability of finding θ (given observations Y) in this interval is γ . It is common to report a 95% credible interval, in this case, we require

$$P(\theta_1 \leq \theta \leq \theta_2 | Y) = 0.95.$$

Credible intervals are not unique. There can be many, in fact infinitely many, γ -credible intervals. Oftentimes, we report the smallest credible interval (SCI; which is sometimes also called the highest density interval). The γ -SCI is defined as

$$\gamma\text{-SCI} = \arg \min_{[\theta_1, \theta_2] \subset \mathbb{R}} \{\text{length}([\theta_1, \theta_2]) = \theta_2 - \theta_1 : P(\theta_1 \leq \theta \leq \theta_2 | Y) = \gamma\}.$$

SCIs always contain the median; for unimodal posteriors, they also contain the mode.

Why are SCIs a useful tool for uncertainty quantification? The more we know about a parameter θ , the narrower our posterior. Similarly, the SCI will get tighter and tighter. By reporting the SCI, we can communicate our uncertainty without having to detail the entire posterior.

¹⁹To convince yourself that this is true, you can calculate the expectation yourself or look up the properties of the Beta-distribution.

4.1.3 Laplace and the Motion of Planets

Let us travel back in time and consider one of the problems that may have kept Laplace up at night. It is the year 1774 and many astronomers take measurements of the positions of celestial objects. How do we infer an object’s position from multiple (flawed) observations? This was a vexing question at the time. Newton had shown that the stars are not guided by divine providence but follow precise mathematical laws like clockwork. These equations ruling the heavens, however, could be shown to allow chaotic solutions; solutions in which the solar system may fall apart with planets violently being flung into deep space.

To assure themselves of the stability of the sky, astronomers began to rigorously map the stars and planets. For a determinate answer, they would need to find out how to combine observations. We may be tempted to immediately jump to an answer: “Take the average”, I hear you say. While this may seem very intuitive to us, Laplace was working hard on a convincing justification for doing so.

Let’s make this a little more formal. Suppose μ is the object’s position and let y_i be an observation of its position. Laplace’s crucial idea was that if we consider the measurement error $\varepsilon_i = y_i - \mu$ to be random (and symmetrical), then we can reason equally from the unknown parameter μ to the observation (i.e., $y_i = \mu + \varepsilon_i$) or, and this is the crucial insight, from the observation to the unknown parameter (i.e., $\mu = y_i - \varepsilon_i$). Laplace reasoned that we can hold the observations *fixed* and consider the conditional probability of the object’s position given these observations. This is a big conceptual leap!

Let us make this even more concrete. Let us assume that the measurement errors are distributed normally, that is, $\varepsilon_i \sim \text{Normal}(0, \sigma)$.²⁰ For simplicity, we will assume to know the standard deviation σ of our observations. This is equivalent to saying that our observations are normally distributed with mean μ , i.e., $y_i \sim \text{Normal}(\mu, \sigma)$. The position μ of the celestial object is the parameter of the distribution that determines our observations. Our goal is to calculate the probability of values of μ given the observation y we’ve made, that is, we want to calculate the posterior $p(\mu|y)$.

Laplace argued that the prior should be thought of as flat, that is, uniform over all possible values of μ . This is sometimes referred to as the principle of indifference. We will talk more about this in the section on non-informative priors, see Section 4.1.5. For Laplace, however, this was not a statement about how we ought to quantify our beliefs in a state of ignorance – or a metaphysical statement about the distributions of μ in the world. He argued that, for simplicity, the “causes” of y should be divided such that all of them are equally likely. Defining μ such that $\pi(\mu)$ is flat was simply advice to simplify

²⁰Here we are jumping a little bit ahead in the history of science. In 1774, Laplace did not yet know what a reasonable error distribution might look like. For him to use what we now know as the normal distribution, he first had to derive one of his most famous results – the central limit theorem. In 1774, given only very vague constraints on the error distribution, he was only able to derive how to combine *three* observations. In 1809, Gauss would show that only for normally distributed errors is the arithmetic mean the best guess at the true value. Laplace came across this result in 1810 and immediately recognized its importance. Together with his central limit theorem, Gauss’ argument solved the problem.

calculations [Stigler, 1990].

4.1.4 Bayesian Inference for the Normal Distribution

Let us follow through on the actual calculations: Assume μ is the true position of an object which is measured to be at positions $y_1, \dots, y_n$. Due to measurement error, the n observations are flawed. Based on the central limit theorem, we may posit that a normal distribution is a good approximation for the measurement error. We hence assume that $y_i \stackrel{iid.}{\sim} \mathcal{N}(\mu, \sigma)$. To simplify matters, we will assume that the spread σ of the measurement error is known – we often know the accuracy of our measurement devices.²¹

Example 4.1 (Normal Likelihood - Flat Prior). Now suppose we have a flat prior on μ , that is, $p(\mu) = 1$. Notice that this is not a proper probability distribution. We call priors like this *improper priors*. If the data is informative enough, the posterior is a valid posterior (i.e., a probability distribution) even if the prior isn't. This is the case in our situation here. With this (improper) prior and the normal likelihood above, we have Normal - Flat

$$\begin{aligned} p(\mu|y_1, \dots, y_n) &= \frac{p(y_1, \dots, y_n|\mu)p(\mu)}{p(y_1, \dots, y_n)} \\ &= \mathcal{N}(\bar{y}, \frac{\sigma}{\sqrt{n}}), \end{aligned} \tag{38}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is the average of the measurements.

Notice that we have succeeded in showing what Laplace set out to show. The posterior mean, our best guess at the object's location μ , is *the average* of the observations, that is, $\bar{y}$! We can also see that our uncertainty about the object's location μ decreases with the number of observations n , specifically with $\frac{1}{\sqrt{n}}$.

Example 4.2 (Normal Likelihood - Normal Prior). A strength of Bayesian inference is its capability to make use of prior information. Suppose the object we are studying is actually well known and has, for example, already been observed during the course of the previous years. We might be able to make a reasonable guess about its location even before collecting the observations $y_1, \dots, y_n$. Bayesian inference allows us to incorporate this prior knowledge. We may, for example, a priori believe that $\mu \sim \mathcal{N}(\mu_0, \sigma_0)$. With this informative prior, we have the following for the posterior:

$$\begin{aligned} p(\mu|y_1, \dots, y_n) &= \frac{p(y_1, \dots, y_n|\mu)p(\mu)}{p(y_1, \dots, y_n)} \\ &= \mathcal{N}(\mu_n, \sigma_n), \end{aligned} \tag{39}$$

²¹Conceptually it's not a problem to also estimate σ from the data. Any textbook on Bayesian inference will teach you how to do that, see for example [Gelman et al., 2013].

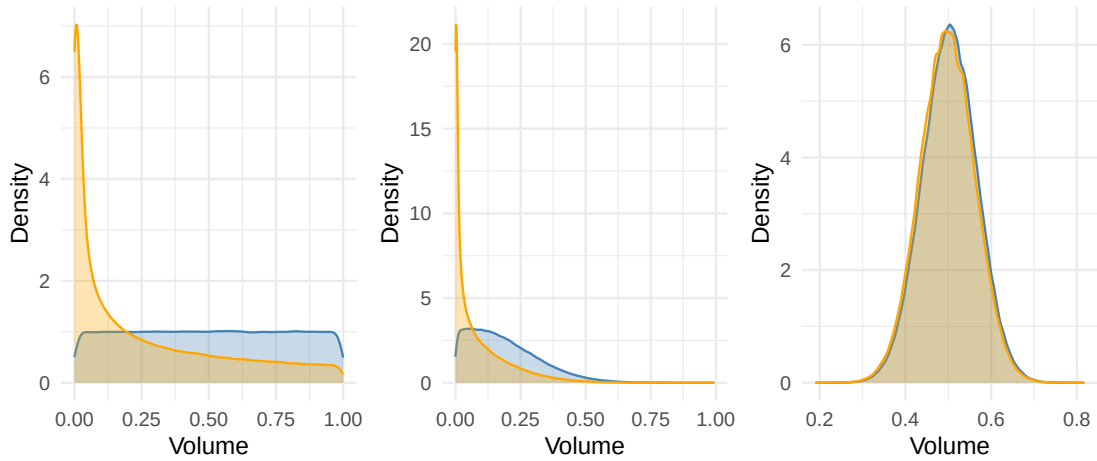


Figure 4: Comparison of inference based on a flat prior on the length of boxes vs. a flat prior on the volume of the boxes.

where

$$\mu_n = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \quad \sigma_n^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1}.$$

We can see that the informative prior $\mu \sim \mathcal{N}(\mu_0, \sigma_0)$ acted like additional data.

4.1.5 Non-informative Priors (Objective Bayes)

We mentioned multiple times already that it is not straightforward to assign a prior to the state of ignorance. Why is this important? While we often do have strong priors about the phenomena we study, there is also a need for objectivity (or more concretely, inter-subjective agreement). Oftentimes, we may be able to agree on the validity (and relevance) of the data, but we may not share the same priors. In this case, we may want to “let the data speak” for themselves. To do so, we would ideally start from a blank slate, from a prior of complete ignorance. If we can agree on how this *non-informative* prior should look, we could perform “objective” Bayesian inference [Berger et al., 2024].

In our examples, we used the flat prior as a natural choice in the absence of prior knowledge. This principle – assigning the flat prior in case of ignorance – is sometimes called the *principle of insufficient reason* (or the principle of indifference). While this principle is often attributed to Laplace, his reasoning was more nuanced [Stigler, 1990].

Unfortunately, this principle suffers from a crucial flaw; its result depends on the (often arbitrary) choice of how to parameterize a problem – this led many famous statisticians, including Fisher and Feller (see Section 4.2) to heavily criticize Bayesian inference as a whole. Let’s pick an easy example to make this point:

Suppose a factory produces cubes with sides of length somewhere between 0 and 1 meter. If we know nothing about the production process, we may follow the principle

of indifference; we hold a uniform prior ($\text{length} \sim \text{unif}([0, 1])$) over all possible lengths of the sides of the boxes. But now imagine, we had begun our discussion by talking about the boxes' volumes. The volume of the boxes is constrained to be between 0 and 1 cubic meters also. The principle of indifference now leads us to hold a flat prior on the volume ($\text{volume} \sim \text{unif}([0, 1])$). Figure 4 clearly shows that these two priors are radically different – the principle of ignorance leads to starkly different choices of prior based on the largely arbitrary choice of how to parameterize the problem.

Is Bayesian inference doomed? No, there are multiple possible responses to this. There has been a lot of work aimed at constructing better “objective” priors than the flat prior [Berger et al., 2024]. While none of them come without problems, there is another argument that this problem of finding an “objective” prior is maybe less important than it may seem at first glance. Most Bayesian models are data-dominated if enough data is at hand: because the prior appears only once in the calculation of the posterior, but the likelihood is raised to the power of the sample size, we often find that the impact of the prior is washed out. Figure 4 illustrates this point. While the posteriors are still drastically different for small samples (middle panel), the choice of prior becomes irrelevant in large samples (right panel).

4.2 Maximum Likelihood

In 1921, Ronald Fisher published *On the Mathematical Foundations of Theoretical Statistics* [Fisher, 1922] – an absolute tour de force. His paper, which truly lives up to the name, introduces the method of maximum likelihood estimation, one of the central pillars of statistics to this day.

Let us take a step back. Bayesian Inference gives a routine procedure of determining a parameter θ from data and immediately comes with uncertainty quantification. Posterior credible intervals give a measure of how uncertain we are about our posterior estimate of θ . As we’ve discussed, Bayesian inference relies on the specification of a prior, and some of the greatest statisticians of all time criticized the practice of Bayesian inference to the utmost.²² While their critique may look overly harsh from a modern perspective – there has been a lot of work successfully addressing many of the challenges – Fisher’s insight was still tremendously enlightening: The goal of statistics is to quantify how data can illuminate an empirical question. Part of this quest is achieving a convincing degree of objectivity. Whatever these estimates may be used for (think of governance or science), their validity needs to be agreed on by all involved agents. The main obstacle to achieving this consensus when using Bayes, at least during Fisher’s time, was the choice of prior. When nothing is known about a parameter (the state of ignorance), the choice of prior is not unique – and thus, to some degree, arbitrary (we’ve discussed this

²²for example: “There would be no need to emphasise the baseless character of the assumptions made under the titles of inverse probability and Bayes’ Theorem in view of the decisive criticism to which they have been exposed at the hands of Boole, Venn, and Chrystal, were it not for the fact that the older writers, such as Laplace and Poisson, who accepted these assumptions, also laid the foundations of the modern theory of statistics, and have introduced into their discussions of this subject ideas of a similar character.” (Fisher, *ibid.*) A similar sentiment can be found in Feller’s famous textbook on probability [Feller, 1991].

in Section 4.1.5). It was unclear how far Bayesian inference could be the foundation for objective empirical inquiry.²³

Fisher set out to create an objective basis for empirical research. Instead of relying on Bayesian inference for estimation, he developed the estimation procedure of *maximum likelihood*. Similar to Bayesian inference, Fisher envisioned this to be an estimation routine applicable to a wide variety of probabilistic problems that automatically provides a corresponding uncertainty quantification.

Before we dive into maximum likelihood estimation, it's worth following Fisher in rigorously setting out the goals of estimation. We begin with the most fundamental criterion: Any estimator of a parameter θ should, as the sample size n increases to infinity, converge²⁴ to the true value of that parameter. This criterion is called *criterion of consistency*. In modern lingo, we would require our estimator to be *consistent*.

The next two criteria are concerned with the quality of estimators. In the class of all consistent estimators, which one should we pick? A high-level answer would be to pick the one that is most informative about the parameter. Fisher writes: “[I]n large samples, when the distributions of the statistics tend to normality, that statistic is to be chosen which has the least probable error.” (Fisher, *ibid.* pg. 316) He calls this the *criterion of efficiency*. In a modern formulation, efficiency is understood as seeking an estimator with the smallest asymptotic variance. A deep characterization of efficiency would take another two decades. Multiple mathematicians (including Harald Cramér and C. R. Rao) managed to prove that there is a (lower) bound on how well we can do with our estimators: the famous Cramér-Rao bound.

Fisher's third criterion states that the statistic on which the estimator is based “should summarise the whole of the relevant information supplied by the sample.” (Fisher, *ibid.*) This *criterion of sufficiency* is actually closely linked with his second criterion of efficiency. At the end of the 1940's, C.R. Rao and David Blackwell managed to show that estimators can typically be improved – without any risk of hurt – by conditioning on sufficient statistics (statistics that fully capture the data under the probability model in use). We should thus expect any efficient estimator to be based solely on sufficient statistics.

We see that Fisher set a tremendously high goal-post for any estimation routine. As we will see, his method of maximum likelihood estimation was up to the task.

We begin by discussing what we mean by a likelihood function. We have seen before that a probability model for some phenomenon (say, a normal distribution) is associated with a probability density $f(x)$ that tells us how likely it is to see some observation x . To highlight that this probability density is a function of an (unknown) parameter θ (say, the mean of a normal distribution), we could index the pdf and write $f_{\theta}(x)$.

²³It is worth emphasizing that a shift from a frequentist to a subjective perspective of probabilities does not dissolve this problem. While subjective probabilities make a prior (and posterior) distribution of fixed parameters cogent, they do not immediately solve the problem of creating consensus. As long as there is disagreement about the (subjective) prior in a relevant community, Bayesian inference runs the risk of not being able to provide the relevant stability (objectivity) required from empirical inquiry.

²⁴The type of convergence referred to here is convergence in probability.

As a function of x , this is a pdf and fulfills the calculus of probability; for example, it integrates to one: $\int_{-\infty}^{\infty} f_{\theta}(x)dx = 1$. We are now interested in this quantity as a function of the unknown parameter holding x fixed. To highlight this change in perspective, we change the notation and write $L(\theta; x) = f_{\theta}(x)$. We call $L(\theta; x)$ the likelihood of θ given x . Notice that the likelihood (a function of θ) does not behave like a probability – most importantly, it does not integrate to 1.

The likelihood function $L(\theta; x)$ tells us how likely our observation x is for different values of θ . Fisher’s central idea was to pick the value that maximizes this likelihood as an estimate for the unknown parameter θ .

Definition 4.2 (Maximum Likelihood). For each sample point x , let $\hat{\theta}(x)$ be a parameter value at which the likelihood $L(\theta; x)$ attains its maximum (as a function of θ , with x held fixed). We call $\hat{\theta}(X)$ based on a sample X the maximum likelihood estimator (MLE) of the parameter θ .

One problem that Fisher highlighted with Bayesian inference is that the principle of indifference (i.e., choosing a flat prior) leads Bayesian estimation to be dependent on the choice of parametrization. Estimation of θ will lead to a different result than estimation of a function of θ . The following theorem shows that this is not a problem for maximum likelihood estimation.

Theorem 4.2 (Invariance of MLEs). *If $\hat{\theta}$ is the MLE of θ , then for any function $\tau(\cdot)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.*

Example 4.3. Let us go into more detail about what parts of Bayesian inference rubbed Fisher the wrong way. Think back to the example that Thomas Bayes himself gave when he used Bayes’ rule: He performed inference for the Binomial distribution but the example he gave was not the classical one that would most easily come to mind – the one of drawing successive draws from an urn.

Why did Bayes instead consider his Billiard table? He considered the billiard table because the specific setup of rolling two balls (one to determine the parameter θ and one that is thrown multiple times to make the observations) introduces a symmetry that allows us to more naturally think of the parameter of interest θ being distributed with some prior. In the case of urns, it is unclear what the prior distribution $\pi(\theta)$ is supposed to describe. From a frequentist perspective, one that Fisher very much championed, we would have to think of the urn being randomly drawn from some (infinite) population of urns in the universe. Both Fisher, and later Feller, argued that this seems absurd. We may, instead, try to argue that the prior is supposed to describe our subjective beliefs. While this may be less absurd, we still run into a potential problem: there is no clear prior describing our ignorance (see Section 4.1.5).

Equipped with the method of maximum likelihood, we can find an estimator of θ without any use of priors. Let us see how Fisher tackled the problem. Suppose we draw a random sample from an (infinite) urn. In this urn, let θ be the proportion of “successes”. Let the random sample be of size n and denote with X the number of successes in this random sample. We know that the number of successes X is binomially

distributed. That is, $X \sim \text{Binom}(n, \theta)$. We hence have the following likelihood function

$$L(\theta; X = k) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}. \quad (40)$$

To determine the MLE, it is convenient to take the logarithm. Note that the logarithm is a monotone transformation so that the maximum of the log likelihood ($\log L(\theta; X = k)$) is the same as the likelihood:

$$\log L(\theta; X = k) = k \log \theta + (n - k) \log (1 - \theta) + C, \quad (41)$$

where C contains all terms unrelated to θ . We can now obtain the maximum of the log likelihood by differentiation

$$\frac{d}{d\theta} \log L(\theta; X = k) = \frac{k}{\theta} - \frac{n - k}{1 - \theta}. \quad (42)$$

Setting this derivative to 0, we find

$$\hat{\theta} = \frac{k}{n} \quad (43)$$

for the MLE.

Now notice that we obtain the same result if we consider a different parameterization (which we already know from Theorem 4.2).

Fisher asks us to imagine that we had measured probability on the scale $\sin \vartheta = 2\theta - 1$ instead. As he points out, ϑ measures the ratio of “successes” to “failures” in the urn as well as the original parameter θ .

For the new parameterization, the log likelihood becomes

$$\log L(\vartheta; X = k) = k \log 1 + \sin \vartheta + (n - k) \log (1 - \sin \vartheta) + \tilde{C}, \quad (44)$$

where $\tilde{C}$, again, contains all terms unrelated to ϑ . After differentiation, we obtain

$$\widehat{\sin \vartheta} = \frac{2k - n}{n} = 2\frac{k}{n} - 1 = 2\hat{\theta} - 1 \quad (45)$$

as the MLE. This is exactly the transformation of the original MLE.

Fisher’s method of maximum likelihood thus avoids both the need to choose a prior as well as the problem of sensitivity to the choice of parameterization that Bayesian inference with a flat prior faces.

Example 4.4. Now let us continue the problem that Gauss and Laplace faced: determining a planet’s orbit based on flawed observations. If we parameterize the position on the ellipse with ϕ , the coordinate of any point on the ellipse $\mathbf{x}(\phi) = (x(\phi), y(\phi))$ is given by the following equation:

$$(x(\phi), y(\phi)) = (\alpha \cos \phi + a, \beta \sin \phi + b), \quad 0 \leq \phi < 2\pi,$$

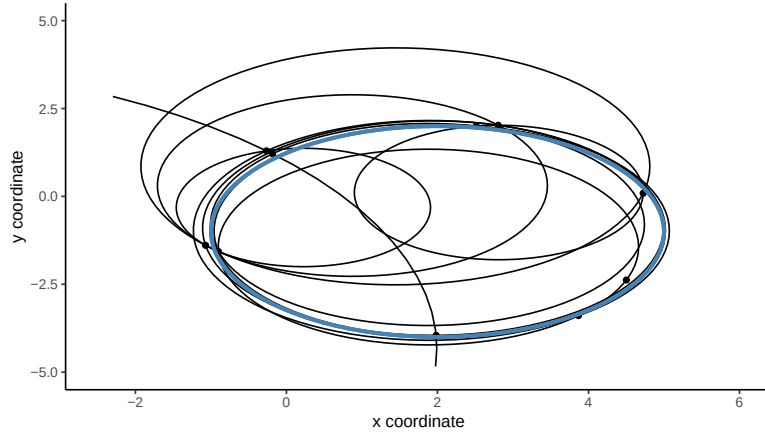


Figure 5: Multiple noisy observations of a planet's trajectory are made. Any combination of 5 points allows us to fit an ellipse. These ellipses (black) are, however, very bad approximations of the true orbit. Assuming normal measurement error, the maximum likelihood estimate for the orbit is equivalent to the least squares solution.

where $\alpha, \beta, a, b \leq 0 \in \mathbb{R}$ are parameters describing the planet's orbit. It is these parameters that we aim to determine from data. Let $\boldsymbol{\theta} = (\alpha, \beta, a, b)$ be the collection of these parameters.

Laplace had argued that, based on the central limit theorem, we should suspect the measurement error to follow a normal distribution. Suppose, for simplicity, we know ϕ perfectly for all of our observations. We can then, following Laplace's reasoning, think of an observation as a normally distributed random vector²⁵

$$\mathbf{X}(\phi) \sim \mathcal{N}_2(\mathbf{x}(\phi), \sigma^2 \mathbf{I}_2).$$

Now, suppose we have many independent observations (all assumed to have the same known measurement error). We then have $\mathbf{X}_i(\phi_i) \sim \mathcal{N}_2(\mathbf{x}(\phi_i), \sigma^2 \mathbf{I}_2)$ (for $1 \leq i \leq n$) independent.

The likelihood of these observations given the parameters $\boldsymbol{\theta}$ is given by

$$\begin{aligned} L(\boldsymbol{\theta}; (\mathbf{X}_i(\phi_i))_{i=1}^n) &= \prod_{i=1}^n \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\|\mathbf{X}_i(\phi_i) - \mathbf{x}(\phi_i)\|^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n \|\mathbf{X}_i(\phi_i) - \mathbf{x}(\phi_i)\|^2\right). \end{aligned} \quad (46)$$

²⁵For simplicity, we will assume the measurement error in the two dimensions x and y to be independent and of the same magnitude σ .

It is convenient to consider the log likelihood

$$\log L(\boldsymbol{\theta}; (\mathbf{X}_i(\phi_i))_{i=1}^n) = -\frac{1}{2\sigma^2} \sum_{i=1}^n \|\mathbf{X}_i(\phi_i) - \mathbf{x}(\phi_i)\|^2 + C,$$

where C contains all terms unrelated to $\boldsymbol{\theta}$.

Taking derivatives, we find what Gauss derived more than 200 years ago:

$$\begin{aligned} \hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta}, \hat{a}, \hat{b}) &= \arg \min_{\alpha, \beta, a, b \geq 0} \sum_{i=1}^n \|\mathbf{X}_i(\phi_i) - \mathbf{x}(\phi_i)\|^2 \\ &= \arg \min_{\alpha, \beta, a, b \geq 0} \sum_{i=1}^n \|\mathbf{X}_i(\phi_i) - (\alpha \cos \phi_i + a, \beta \sin \phi_i + b)\|^2. \end{aligned}$$

The parameters should be determined with the method of least-squares! We will encounter the method of least-squares again when discussing linear regression (see Section 6.3).

Now that we have discussed how to perform maximum likelihood, let us see how it fares in terms of consistency and efficiency. Before we do so, we need to introduce the notion of Fisher information.

Definition 4.3 (Fisher Information). The Fisher Information of a parameter θ given a random variable X with likelihood $L(\theta; X)$ is defined as

$$I(\theta) = \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \log L(\theta; X) \right)^2 \right] \quad (47)$$

It can be shown that this is equivalent to

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \log L(\theta; X) \right].$$

The Fisher information is a way of quantifying how much we can learn about an unknown parameter θ from an observable random variable X . Geometrically, the Fisher information can be understood as the curvature of the likelihood function around the parameter θ . If the curvature is high, the likelihood of an observed sample point x changes rapidly for parameter values close to θ – the likelihood of x lets us learn about θ with high resolution. If, on the other hand, the curvature is small, the likelihood function will be very flat around θ . In this case, it is much harder to distinguish parameter values around θ based on the likelihood of the observed sample point x .

The following theorem shows that the MLE fulfills Fisher's criterion of efficiency; the MLE achieves the smallest possible asymptotic variance:

Theorem 4.3 (Asymptotic Distribution of the MLE). *Let $X_1, \dots, X_n$ be a random sample with likelihood $L(\theta; x)$. Under some regularity conditions²⁶, the MLE $\hat{\theta}$ of θ is asymptotically efficient, that is,*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, I^{-1/2}(\theta)), \quad (48)$$

where $I(\theta)$ is the Fisher-Information evaluated at the true parameter θ .

4.2.1 Confidence Intervals

While we may not always be in a position to say much about the finite-sample properties of an MLE, theorem 4.3 allows us to construct asymptotically valid *confidence intervals* based on the normal distribution. Let X be a random sample from a distribution with parameter θ . Then a γ -confidence interval for θ is any random interval $[l(X), u(X)]$ that has probability γ to cover θ , that is, $P(\theta \in [l(X), u(X)]) = \gamma$. For large enough samples, using the normal approximation (Thm. 4.3), we have

$$P(\theta \in [\hat{\theta} - \frac{2}{\sqrt{n}}I^{-1/2}(\theta), \hat{\theta} + \frac{2}{\sqrt{n}}I^{-1/2}(\theta)]) \approx 95\%. \quad (49)$$

This is just an application of the 68 – 95 – 99.7 rule discussed before.

It's important to distinguish between confidence intervals and credible intervals. When working with credible intervals, we think of θ as a random variable and control the probability of θ being in a fixed interval. The interval is fixed because it's based on the data that we condition on. For confidence intervals, the situation is reversed. We consider θ to be fixed and the interval is random. An analogy may be useful: Think of the game of darts. For credible intervals, we think of throwing the dart (the parameter) at the dart board and we want to make sure that we hit the board 95% of the time. In the frequentist setting, the situation is reversed. The dart is already on the wall (the parameter is fixed). We now throw boards (confidence intervals) at it and want to make sure that 95% of them contain the dart.

4.3 A modern perspective on Bayesian Inference

To be written.

5 Science and Hypothesis Testing

Before we begin our discussion of hypothesis tests, join me on a brief excursion into the realm of the philosophy of science. To understand when (statistical) hypothesis tests are at their best, and why the practice of null-hypothesis significance testing caused a replication crisis, we need to better understand what role empirical evidence plays in science more generally.

²⁶These are discussed, for example, in [Casella and Berger, 2024].

5.1 Induction, the Backbone of Science: From Confirmation To Methodological Falsificationism

Our story begins with Rudolf Carnap, who in the early 20th century tried himself at developing a *logic of induction*. Induction is the process by which we learn from experience: Suppose you try a flavor of ice cream – I personally get much satisfaction from anything dark chocolate – and find that you really like it. Throughout your life, you keep trying ice cream of the same name and color, finding each time that you really enjoy it. By induction, you conclude that you really like ice cream of this type in general.

More formally, we seem to employ an argument of the following form

$$\begin{aligned} &\text{All observed } A_1, A_2, \dots, A_n \text{ have been } B\text{'s} \\ \Rightarrow &\text{All } A\text{'s are } B\text{'s.} \end{aligned}$$

At least since Hume’s *Treatise of Human Nature* (1739), philosophers have been keenly aware of problems with this type of argument. It is not deductively correct:²⁷ The premise “All observed $A_1, A_2, \dots, A_n$ have been B ’s” can be correct without the conclusion being so too. It is not unimaginable for example, that despite how many dark chocolate flavored ice creams I have tried in the past and liked, I may (very unfortunately) one day stumble upon one that I dislike.²⁸

We may hope, and so did Carnap, that Bayesian inference offers a way out of this problem. If we have a prior on the hypothesis “All A ’s are B ’s”, we can use Bayesian inference to calculate how much our belief in B has increased after having observed $A_1, A_2, \dots, A_n$ to be B ’s.²⁹ This would be a *confirmation* account of science in which Bayes’ rule tells us how much our observations confirm our scientific theory. There are problems with this account, however. First, appealing to probabilities does not solve the Humean problem of induction – for the claim that observations of the type “ A_i is B ” warrant that “all A ’s are B ” with some probability is deductively false too. Unfortunately, referring to probabilities makes it not less circular to refer to experience to justify the now probabilistic principle of induction.³⁰

There is a very practical problem with confirmation that I want to mention: When we, as scientists, pick our hypothesis based on the data at hand, confirmation is a very weak property. Oftentimes we collect data to “test” a theory we hold to find the data to not corroborate it. The data seems to tell another story, it seems so very clearly to imply another theory. Naturally, we learn from the data and “test” this new theory on the data that inspired it. To not much surprise, it fits extremely well; the data confirms the new theory. The problem is, and this is a problem as old as science, that we have no

²⁷This type of induction is related to the well-known fallacy of “affirming the consequent”.

²⁸The common example is the following: Suppose you have so far only observed *white* swans and conclude that all swans are white. This is not a deductively valid argument since no matter how many swans you have seen (unless you have seen all of them), there may be some that are black.

²⁹Carnap’s *Logical/Evidential Interpretation* of probability is another interpretation of probability beyond the frequentist and subjective perspectives we have covered in this draft. See the Stanford Encyclopedia of Philosophy entry on Interpretations of Probability.

³⁰There are also specific problems with probabilities as confirmation functions, see Mayo [2018, Section 2.2, pp. 66–74].

reason to expect that this theory will be able to generalize to new data. By picking the theory to explain the data we’ve already seen, we may very easily fool ourselves. As an extreme example, think about a sequence of ten (fair) coin flips. The sequence happens to come out as *HHTHTTTHTT*. Ask yourself, once you’ve seen this sequence, which hypothesis is more likely: I) The coins are unbiased, $P(H) = 0.5$, or II) every throw was done with a fake coin that had two sides either both “heads” or both “tails”. The first two throws were fake “heads” coins, the third a fake “tails” coin, and so on. The hypothesis tailored to the observed sequence explains the data perfectly (the likelihood of the observed sequence is 1). The data confirms the theory that was tailored to fit it. We have all reason, however, to be skeptical of this “ad hoc” (or “just so”) hypothesis. It will completely fail to describe a new sequence of coin tosses. This is a very extreme example but it is all too easy when confirmation is the yardstick to fall into the trap of constructing an elaborate theory to match the data that’s currently under study. “Just so” theories are always confirmed by the data they are meant to explain but confirmation gives us no reason to believe the theory to be predictive of new data.

One of the most famous philosophers of science, Karl Popper, argued that induction should not be understood as confirming our theories at all but instead as trying to *falsify* them. In his seminal work *The Logic of Scientific Discovery*, he points out that while the inductive argument is a fallacy, an argument of the type

One observed A_1 is **not** $B \Rightarrow$ The claim “All A ’s are B ’s” is false

is deductively correct – even one counterexample proves the general statement incorrect. Scientific hypotheses rarely come in such a neat deductive form, so the idea of falsification is more involved. What makes science, in Popper’s view, is our continued genuine effort to refute our own theories. He writes: “Repeated observations and experiments function in science as tests of our conjectures or hypotheses, i.e., as attempted refutations.” (Popper, *Conjectures and Refutations*, 1962, p.53)

It is important that our efforts at refutations are genuine. We learn nothing if our experiments have no chance at actually rejecting our theory.

Example 5.1. In her book, Mayo describes a case in which researchers developed a model for individualized chemotherapy [Potti et al., 2006]. The model essentially sought to predict which chemotherapy was most promising based on a patient’s characteristics. When evaluating their model, however, the researchers removed more than half of the samples, keeping only those that best agreed with their model [Mayo, 2018, p. 6]. The published data looked convincing, but when other researchers, excited by these results, tried to reproduce them, they failed [Coombes et al., 2007, Baggerly and Coombes, 2009]. The problem is in the logic of confirmation: a procedure seeking to validate a model by reporting only the results that agreed with it is not a genuine attempt at refutation.

Following Popper, Deborah Mayo (who is both a philosopher of science and statistics), emphasizes the importance of *severe testing*. The test to which we put our hypothesis has to be severe. She brings forward the following requirement [Mayo, 2018, p. 14]:

Principle (Strong Severity). *We have evidence for a claim C just to the extent it survives stringent scrutiny.* If C passes a test that was highly capable of finding flaws or discrepancies from C , and yet none or few are found, then the passing result x , is evidence for C .

Specifically, Popper imagines science to progress in the following way: We let a scientific hypothesis guide us to a novel, that is, not yet tested, prediction. Even if this hypothesis is deterministic, our prediction will most likely be a statistical hypothesis; there are always uncertainties in the form of knowledge-gaps or imperfect measurements. This statistical hypothesis, Popper calls the *falsifying* hypothesis – a hypothesis inferred to falsify some claim. As an example, think of the problem of inferring that an anomaly (a deviation from our theory-based expectation) is actually real by falsifying it being “due to chance”. We learn by putting this falsifying hypothesis to a severe test.

Example 5.2. Einstein’s general theory of relativity predicted that the Sun would bend starlight by about 1.75 arcseconds at its edge, roughly twice the Newtonian prediction, due its mass bending spacetime. During the solar eclipse of 1919, astronomers measured the apparent displacement of stars near the Sun and found results favoring Einstein’s prediction [Dyson et al., 1920]. This experiment was seen as a genuine chance of refuting the theory.

Following Popper and Imre Lakatos, Mayo condenses this reasoning into the following requirement [Mayo, 2018, p. 92]:

Principle (Severity Requirement). For data to warrant a hypothesis H requires not just that

(S-1) H agrees with the data (H passes the test), but also

(S-2) with high probability, H would not have passed the test so well, were H false.

Something subtle has happened: Carnap attempted to use probabilities to quantify how much data corroborates a hypothesis by looking at how probable the data makes the hypothesis. For Mayo’s error-statistical perspective, probabilities are not about hypotheses at all! They are about the properties of our scientific methods. To be a severe test of our hypothesis, our test needs to be highly likely to rule out our hypothesis to be false. Probability becomes a tool of gauging the severity of our tests.

It is this shift to a focus on the methodology of falsification that inspired Lakatos to coin the term “Methodological falsificationism” (Lakatos 1970, pg. 106).

5.2 A Lady Tasting Tea: Milk or Tea First?

In the late 1920s, Ronald Fisher, who we have already met, was enjoying some of his friends’ and colleagues’ company. When tea was brought to their outdoor table, one lady – now famous in statistics circles – Muriel Bristol-Roach refused. She had seen that milk had been poured first and was insistent that this ruined the tea. Fisher and her future husband William Roach doubted her ability to discern whether tea or milk had been poured first. But she was willing to put it to the test! [Salsburg, 2001, Chapter 1]

Fisher devised a clever strategy to test her: They would prepare 8 cups – 4 with milk being poured first and 4 with tea being poured first – and would present them to Muriel in a completely random order.

If Muriel was able to pick out the four cups where tea had been poured first, they would be impressed and would be pressed to believe her about her ability.

But suppose, she does not get all of them right. How should we decide if we should be impressed or if the result could have come about by her just guessing?

Fisher’s insight was that her result could easily be compared to what we should expect if she was guessing. This hypothesis that her result is purely due to chance is called the *null hypothesis*. The following table shows the probability distribution of her correctly guessing under this null hypothesis.

Table 2: Tea-Tasting Distribution Assuming the Null Hypothesis

Success count	Combinations of selection	Number of Combinations
0	oooo	$1 \times 1 = 1$ $\binom{4}{0} \times \binom{4}{4}$
1	ooox, ooxo, oxoo, xooo	$4 \times 4 = 16$ $\binom{4}{1} \times \binom{4}{3}$
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	$6 \times 6 = 36$ $\binom{4}{2} \times \binom{4}{2}$
3	oxxx, xoxx, xxox, xxxo	$4 \times 4 = 16$ $\binom{4}{3} \times \binom{4}{1}$
4	xxxx	$1 \times 1 = 1$ $\binom{4}{4} \times \binom{4}{0}$
Total		70

Under the null hypothesis³¹, the number of correct guesses is distributed according to the hypergeometric distribution, that is, the random variable X = the number of correct guesses follows $X \sim \text{Hypergeometric}(N = 8, K = 4, n = 4)$. The probability of getting, say, at least 3 out of the 4 guesses correct, if one were guessing, is $P(X \geq 3) = \frac{16+1}{70} \approx 24.3\%$.

Even though he set out to account for the possibility of Muriel being fallible in her discerning ability, his experiment was too small to reject the hypothesis of her guessing if she even got one guess wrong. Fisher was lucky, however: Muriel actually discerned all 8 cups correctly. This is very unlikely under the null hypothesis, $P(X = 4) = \frac{1}{70} \approx 1.43\%$. They had to give it to her: pouring milk or tea first made a discernible difference to her.

³¹Fisher did not specify an alternative hypothesis. There is a well-documented disagreement between Fisher and Neyman about how hypothesis testing should be conceptualized – and if alternative hypotheses should be stated. If Fisher had wanted, here the alternative could be stated as a noncentral hypergeometric distribution with odds-ratios unequal to 1.

Fisher would later discuss the benefits of larger experiments. We will briefly discuss how to gauge necessary sample sizes for convincing tests in the next section.

5.3 Are Sex Ratios Proof Of Divine Providence?

In 1710, physician John Arbuthnot studied the number of boys and girls born in London [Arbuthnot, 1710]. Every single year, for 82 consecutive years, boys outnumbered girls (Figure 6). To Arbuthnot, this seemed extremely fortuitous, as boys had a higher mortality rate growing up and the excess of baby boys led to a more balanced sex ratio in the population later in life. But maybe nothing mysterious is going on here: could it be that the chances of being born a boy or a girl are actually the same and this excess is just due to chance?

To figure this out, Arbuthnot reasoned as follows. Suppose that each kid being born has an equal probability of being born a boy or a girl; let's call this our null hypothesis H_0 . In any given year, the probability of having an excess number of boys is (at most) $1/2$. We can see this by reasoning from symmetry; either sex is equally likely to outnumber the other, and there may also be a tie. Writing M_i and F_i for the two counts in year i , the probability of observing an excess of boys in 82 consecutive years satisfies

$$P_{H_0}(M_i > F_i \text{ for all } i = 1, \dots, 82) = \prod_{i=1}^{82} P_{H_0}(M_i > F_i) \leq \left(\frac{1}{2}\right)^{82}.$$

This is extraordinarily small. Arbuthnot takes this as evidence for divine providence. It is God, in his wisdom, who compensates for the higher mortality rates of men with a slight excess of boy births.

From a modern vantage point, does Arbuthnot's reasoning about divine providence hold up? The answer is "probably not." His calculation gives us reason to doubt that boys and girls are born with equal probability, but it does not establish divine providence [Stigler, 2016]. Go back to the previous section and spend some time thinking through why, from the perspective of severe testing, Arbuthnot did not do a convincing job of establishing divine providence.

The problem is that he convincingly showed that the observations are extremely unlikely under the null H_0 , but it is not clear if we should prefer the specific alternative explanation H_1 of divine providence instead. In the next section, we turn to the Neyman-Pearson framework for hypothesis testing. There, testing is framed explicitly as a decision problem of choosing between a null and an alternative hypothesis given observations. This will make the role of the alternative hypothesis explicit.

5.4 Neyman-Pearson Hypothesis Testing

Parts of this section still need to be written ³².

³²We are missing: A conceptual introduction to the Neyman-Pearson hypothesis testing framework. This should cover its relation to decision theory. Additionally, we will mention Lakatos' perspective that the "[t]he Neyman-Pearson approach rests completely on methodological falsificationism" [Lakatos,

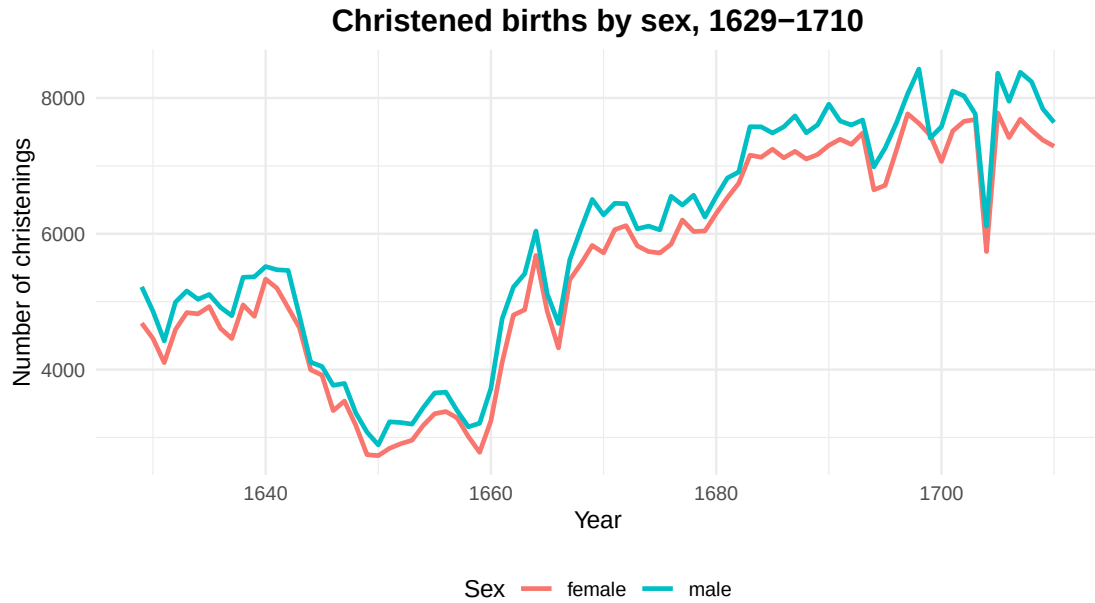


Figure 6: Arbuthnot’s data presented as a time series (1629–1710) of christened births by sex.

We begin with the formal treatment of the Neyman-Pearson framework for hypothesis testing.

Table 3: Two types of errors in hypothesis testing

		Decision	
		Accept H_0	Reject H_0
Truth	H_0	Correct decision	Type I Error
	H_1	Type II Error	Correct decision

Definition 5.1. A hypothesis is any statement about a population parameter.

Definition 5.2. In the Neyman-Pearson hypothesis testing framework, the two complementary hypotheses are called the null hypothesis (H_0) and the alternative hypothesis (H_1).

In general, we have $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$.

1970, p. 109]. We can then turn back and try to understand Neyman-Pearson Hypothesis Testing from the perspective of Mayo’s severe testing. We will also have to add a discussion of the conceptual background of p-values, why they are often misunderstood as a measure of evidence, and their relationship to the replication crisis.

Example 5.3. Suppose we are testing a novel blood pressure medication and we want to know its efficacy. To do so, we may set up a randomized control trial comparing the new medication to the best state-of-the-art alternative. One arm of the trial – the control arm – receives the old medication, the treatment arm receives the new medication. Denote with θ the average difference in patients’ blood pressure between the treatment- and control group. We may be interested in testing $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$ to find out if the new medication is superior to the old.

Definition 5.3. A hypothesis test is a rule that specifies

1. For which sample values the decision is made to accept H_0 .
2. for which sample values H_0 is rejected and H_1 is accepted.

The subset of the sample space for which H_0 is rejected is called the rejection region R . The complement of this region is called the acceptance region.

In testing, it’s usual to control the Type I error.

Definition 5.4. For $0 \leq \alpha \leq 1$, a test is a size α test if

$$\sup_{\theta \in \Theta_0} P_\theta(X \in R) = \alpha.$$

Definition 5.5. For $0 \leq \alpha \leq 1$, a test is a level α test if

$$\sup_{\theta \in \Theta_0} P_\theta(X \in R) \leq \alpha.$$

Example 5.4. Suppose we have a random sample $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma)$ with known standard deviation but unknown mean μ . We may be interested in testing the following null hypothesis:

$$H_0: \mu = \mu_0$$

We have already seen that $\bar{X}$ is a good estimator for the mean μ , so the following test statistic recommends³³ itself:

$$Z = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma}$$

For this test, we have the following rejection region for a level α test. Here z_α is the upper-tail standard normal cutoff, so $1 - \Phi(z_\alpha) = \alpha$, where Φ is the standard normal cdf.

³³We will not discuss how to derive tests and the question of finding optimal tests. For a detailed treatment, see Lehmann and Romano [2022, Chapters 3–6].

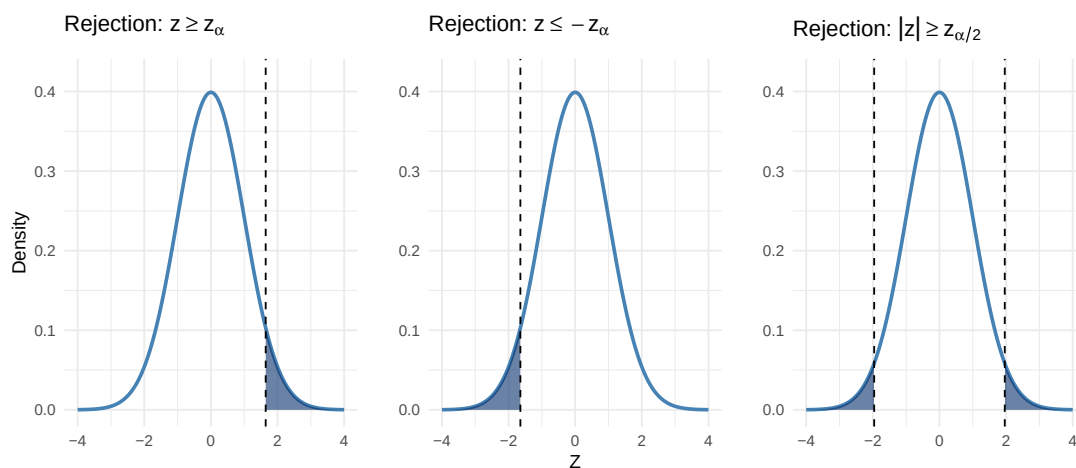


Figure 7: Rejection Regions for different alternative hypotheses for the Z-test.

Alternative hypothesis	Rejection region for level α test
$H_1: \mu > \mu_0$	$Z \geq z_\alpha$
$H_1: \mu < \mu_0$	$Z \leq -z_\alpha$
$H_1: \mu \neq \mu_0$	either $Z \geq z_{\alpha/2}$ or $Z \leq -z_{\alpha/2}$

Example 5.5. Again, suppose we have a random sample $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma)$ with unknown mean μ . But now, we do not assume that σ is known. We may still be interested in testing the following null hypothesis:

$$H_0: \mu = \mu_0.$$

Instead of the Z -score, we now use

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

as a test-statistic, where $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is the sample variance (and $s = \sqrt{s^2}$ is the sample standard deviation).

Under H_0 , this test-statistic follows the Student's t -distribution with $n - 1$ degrees of freedom. For a level α test, we have the following rejection regions, where $t_{n-1, \alpha}$ is the upper-tail cutoff with probability α for this distribution.

Alternative hypothesis	Rejection region for a level α test
$H_1: \mu > \mu_0$	$t \geq t_{n-1, \alpha}$
$H_1: \mu < \mu_0$	$t \leq -t_{n-1, \alpha}$
$H_1: \mu \neq \mu_0$	either $t \geq t_{n-1, \alpha/2}$ or $t \leq -t_{n-1, \alpha/2}$

Definition 5.6. The power function of a test is defined by $\beta(\theta) = P_\theta(X \in R)$. For $\theta \in \Theta_0^c$, the Type II error probability is $1 - \beta(\theta)$.

Example 5.6 (Normal power function). Let $X_1, \dots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma)$ with known population standard deviation σ . Suppose we want to test $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. We have seen that the test that rejects the null H_0 when $\sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \geq z_\alpha$ is a size α test. The power function of this test is given by

$$\begin{aligned} \beta(\mu) &= P_\mu \left(\sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \geq z_\alpha \right) \\ &= P_\mu \left(\sqrt{n} \frac{\bar{X} - \mu}{\sigma} \geq z_\alpha + \sqrt{n} \frac{\mu_0 - \mu}{\sigma} \right) \\ &= P_\mu \left(Z \geq z_\alpha + \sqrt{n} \frac{\mu_0 - \mu}{\sigma} \right), \end{aligned} \tag{50}$$

where $Z \sim \mathcal{N}(0, 1)$ is standard normal. Notice that the power function $\beta(\mu)$ for the Z test is an increasing function for μ with

$$\lim_{\mu \rightarrow -\infty} \beta(\mu) = 0, \quad \lim_{\mu \rightarrow \infty} \beta(\mu) = 1 \quad \text{and} \quad \beta(\mu_0) = \alpha.$$

The following table summarizes the power functions for the Z-test for various alternative hypotheses.

Alternative hypothesis	Power $\beta(\mu')$ for a level α test
$H_1 : \mu > \mu_0$	$1 - \Phi \left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \right)$
$H_1 : \mu < \mu_0$	$\Phi \left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \right)$
$H_1 : \mu \neq \mu_0$	$1 - \Phi \left(z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \right) + \Phi \left(-z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \right)$

Example 5.7. The power function for the Z test is a function of the sample size. We can thus ask how large our sample needs to be to achieve a specific power $\beta(\mu') \geq \beta$ against an alternative value μ' at a given significance level α .

$$n = \begin{cases} \left[\frac{\sigma(z_\alpha - z_\beta)}{\mu_0 - \mu'} \right]^2 & \text{for a one-tailed (upper or lower) test} \\ \left[\frac{\sigma(z_{\alpha/2} - z_\beta)}{\mu_0 - \mu'} \right]^2 & \text{for a two-tailed test (an approximate solution)} \end{cases}$$

Round n up to the next integer.

5.5 P-Values

Definition 5.7. A p-value $p(X)$ is a test statistic satisfying $0 \leq p(x) \leq 1$ for every sample point x . Small values of $p(X)$ give evidence that H_1 is true. A p-value is valid if for every $\theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$, we have

$$P_\theta(p(X) \leq \alpha) \leq \alpha.$$

If $p(X)$ is a valid p-value, it is, by construction, easy to build a level α test based on $p(X)$. The test that rejects H_0 if and only if $p(X) \leq \alpha$ is automatically a level α test.

Theorem 5.1. Let $W(X)$ be a test statistic such that large values support the alternative H_1 . For each sample point x , define

$$p(x) = \sup_{\theta \in \Theta_0} P_\theta(W(X) \geq W(x)).$$

Then $p(X)$ is a valid p-value.

Example 5.8 (Two-sided normal p-value). Suppose $X_1, \dots, X_n$ are a sample from $\mathcal{N}(\mu, \sigma)$ with unknown mean μ and standard deviation σ . We want to test $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$. For the observed value $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$, the p-value is

$$p(x) = P(|T_{n-1}| \geq |t|) = 2P(T_{n-1} \geq |t|),$$

where T_{n-1} follows a t-distribution with $n - 1$ degrees of freedom.

Example 5.9 (One-sided normal p-value). Again, consider the same sample, but now suppose we want to test $H_0: \mu \leq \mu_0$ versus $H_1: \mu > \mu_0$. The supremum in the theorem occurs at $\mu = \mu_0$, so the p-value is

$$p(x) = P(T_{n-1} \geq t).$$

5.6 Crisis in Science and Fallacies of Rejection

A core concern for any empirical investigation is the solidity of its results. The degree to which the empirical result constitutes a fact depends on how well it generalizes (across situations) and how durable it is (across time). This makes replicability, the possibility of repeating the investigation and arriving at the same conclusion, a core pillar of the empirical sciences. Replicability is part of the self-understanding of the modern sciences – only if you can repeat my investigation and find a similar result, can we hope to claim any degree of objectivity.

In the last decade, starting in the early 2010's, a crisis rocked the empirical sciences. During this *replication crisis*, implicating fields from psychology and medicine to the social and even natural sciences [Baker, 2016], it was uncovered that many insights we had thought to have gained did not survive the scrutiny of attempted replication. There are many infamous examples [e.g., Ritchie et al., 2012, Ranehill et al., 2015]. In 2011, the field of psychology mounted an impressive and self-critical effort in which 100 articles

from leading psychology journals were chosen at random ... only around a third of them could be replicated [Open Science Collaboration, 2015]!

This came as a shock to the scientific community at large. While the initial uproar has largely subsided, the crisis continues. We live in an altered scientific world, one of heightened skepticism towards empirical results. With this comes the need for more critical discussions of proper empirical and especially statistical practice.

Many factors contribute and coalesce to make empirical research so brittle. To get a fuller picture of (and eventually solve) this crisis, we need to understand the sociology of research, the material conditions of researchers (think of “publish or perish”) and the way we teach empirical methods and statistics.

All of these we need to leave for others to spell out. Here we focus instead on one aspect related to statistical hypothesis testing that played (a major) role in the replication crisis: fallacies of Rejection in null hypothesis significance testing.

Often research proceeds as follows. Data is collected to study some substantive hypothesis H^* . This substantive hypothesis might be a specific causal story of how variables in your data influence each other. We then construct a null hypothesis H_0 of no effect that would falsify our substantive hypothesis H^* . For example, we may ask, what would the data look like if the variables of interest were independent? This null hypothesis is constructed such that we can calculate (or simulate) the sampling distribution of some test statistic t . This test statistic t is designed to pick up on violations of H_0 (alternatives H to H_0 remain implicit in the choice of t , they “lurk in the undergrowth” as famous statistician Cox said [Mayo and Cox, 2006, p. 81]). With all these ingredients, we calculate a p-value $p_{obs} = P_{H_0}(t \geq t(y_{obs}))$. If this p-value is very small (say 0.05 or 0.001), we reject the null hypothesis H_0 .

Nothing problematic has happened so far, it is all about how we interpret this p-value! P-values are frequently misinterpreted and misused. Deborah Mayo lists the following frequent *fallacies of rejection* [Mayo, 2018, p. 94]:

- The reported p-value is spurious, i.e., it does not describe the actual probability of seeing a test statistic as extreme if the null hypothesis was true. This happens, for example, if the probabilistic model we use does not fit the data. P-values are also spurious when we test multiple hypotheses without accounting for multiple testing.
- The reported statistical result is genuine but it is an isolated statistical fluctuation. An isolated statistical result is not, by itself, indicative of a genuine experimental phenomenon.
- This is probably the most important fallacy: There may be evidence of a genuine (reproducible) statistical effect but the substantive interpretation is unwarranted. We may, with convincing power, conclude H when rejecting H_0 but our statistical alternative H may not warrant concluding the substantive hypothesis H^* .

Keeping these fallacies in mind when interpreting your own p-values may save you from contributing to the pile of non-replicable findings!

6 Decomposing Variation: Linear Regression

6.1 Galton, Decomposing Variation and Regression to the Mean

In 1859, Charles Darwin published *On the Origin of Species*. His book would not only go on to found evolutionary biology but would, through some twists and turns, also lead to the development of one of statistics' most central tools – linear regression.

We begin the story of linear regression with Charles Darwin's cousin, Francis Galton. Galton thought he had uncovered a paradox in Darwin's theory of evolution: While Darwin's theory explained how species changed over long periods of time, his theory seemed to contradict the apparent stability of populations in the short term. Consider, as Galton did, a population's height distribution. If height is hereditary, Galton argued, tall parents should have tall offspring. Integral to Darwin's theory are random deviations that are then subject to natural selection. Following this logic, some of the tall parents' offspring should be even taller than them. If this is correct, we should see a population's height distribution become wider and wider even over just a few generations. This was in contradiction to the stability of populations observed. Galton set out to find an explanation.

While not a brilliant mathematician, Galton was clever and inventive. His journey to an explanation began with the development of the quincunx. This vertical board is a physical analogy for the central limit theorem. The board is filled with interleaved rows of pegs on which metal beads dropped from the top bounce either left or right. All these small deflections (think of them as small "errors") aggregate such that the beads, when collected at the bottom, form a normal distribution. This distribution of beads served, in Galton's imagination, as an analogy to the distribution of heights observed in society. As discussed before (Section 2), Quetelet had already established that heights (and many other characteristics) followed a normal distribution. Where Quetelet saw the normal distribution as a way to reason about "a constant cause", to speak about what unifies the seeming diversity (his "average man"), Galton³⁴ instead – in what would turn out to be a crucial step for the application of statistics to the social world – turned to explain exactly this variation. He asked what factors influenced this variation seen in the distribution of heights (in modern statistical language, we would speak of the standard deviation) and why heritability of height does not lead to an increase in this variation from generation to generation.

Having built a physical model producing a normal distribution, he began studying this variation. His first step was the following: Instead of letting the beads fall the entire length of the quincunx, he stopped them halfway. This still leads to a normal distribution but of smaller standard deviation. Think of this as the distribution of heights of parents (or for ease of simplicity, the height of one of the parents).³⁵ When we now release the beads from only one specific parental height, we get a smaller distribution (in the number of total beads) at the bottom of the quincunx that, crucially, has the same deviation as the (entire) parents' distribution on the mid-level. If we, however, release

³⁴probably influenced by his eugenicist politics.

³⁵which Galton, to probably no one's surprise, fixed to be the fathers'.

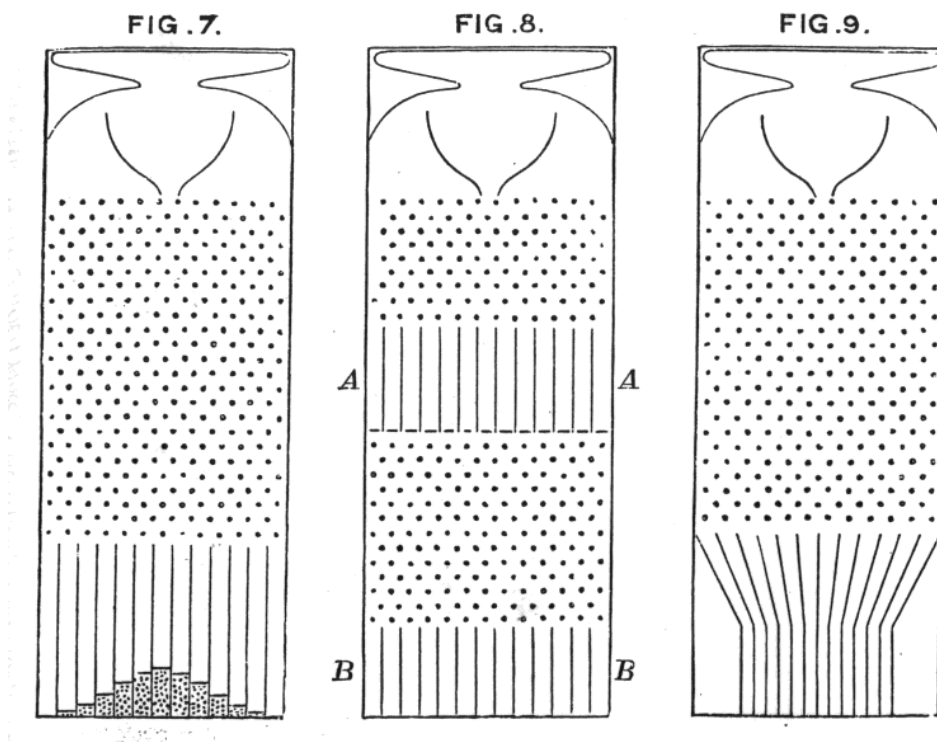


Figure 8: Galton's quincunx. The middle diagram shows the two-stage version: beads collect at A before passing through a second stage to B . The right diagram adds converging channels. Source: Galton [1889, p. 63, Figs. 7–9].

all beads from the middle-stage (representing parents' height), the resulting distribution at the bottom is wider (that is, has larger standard deviation) than the (entire) parents' distribution. He had physically built an analogy to what we now understand as the law of total variance (a conditional variance decomposition), illustrated by the middle diagram in Figure 8. Let h_o be the height of the offspring and let h_p be the height of the parents. We can then decompose the variance of the offspring's height as follows

$$\text{Var}(h_o) = \mathbb{E}[\text{Var}(h_o|h_p)] + \text{Var}(\mathbb{E}[h_o|h_p]).$$

Galton had found that the conditional variance of the children's heights (that is the width of the smaller distributions if beads are only released from one parental height) was equal to the width of the parents' distribution. That is, $\mathbb{E}[\text{Var}(h_o|h_p)] = \text{Var}(h_p)$. He then argued that for the standard deviation of the children $\text{Var}(h_o)$ to be equal to the parents', the other part in the variance decomposition $\text{Var}(\mathbb{E}[h_o|h_p])$ had to be small. As an example, suppose height was perfectly heritable in the sense that children are (on average) the same height as their parents, that is, $\mathbb{E}[h_o|h_p] = h_p$. In this case, the variance of the children's distribution would be twice that of the parents. That is, from generation to generation, the height distribution would become wider in terms of

the standard deviation by a factor of about 1.4. Galton's first thought was that there needed to be a biological explanation that narrowed the distribution back down ensuring that parents' and children's distributions are of the same width. He even imagined a quincunx that at the bottom had diagonal chutes (see the right diagram in Figure 8) to emulate this biological factor that ensured stability of the height distribution across generations. A biological explanation was not particularly likely, however. What factor could have such a fine-tuned effect? It turns out the explanation wasn't biological at all, it would come from statistics.

The insight that the explanation could not be biological (or causal at all) came to Galton by comparing heights of brothers. Here he found the same phenomenon: fixing one brother's height, the other brother was *on average* closer to the center of the distribution. Clearly, because of the symmetry of the situation, there could be no causal explanation for this.

The solution comes by *decomposing* the height into two components, a heritable part ν and a random part ε :

$$h = \nu + \varepsilon.$$

So why is a tall brother's brother *on average* shorter than the other? Both share the heritable part ν . If one of them is exceptionally tall, not only his heritable part contributes to his height but his random part does so too. That is, for a tall brother ε is large too. Most likely, his brother's random part will not be as large. A tall brother's brother will be on average shorter than he is.

Herein lies Galton's solution to the heritability paradox he uncovered in Darwin's theory of evolution. If height (or any other heritable trait) is composed of a heritable portion and a random portion, offspring from an individual far from the center of the distribution will be (on average) closer to the mean. This phenomenon is coined *regression to the mean*.³⁶ The method of linear regression (which we will cover in Section 6.3) borrows its name from this phenomenon, even though the method's usefulness goes beyond the example which gave birth to it and its name.

6.2 Pearson and Correlation

The early 20th century was a tumultuous time. During a time of many upheavals, so too changed the scientific view of the world. Where a few decades before, inspired by Newton's success of describing the heavens, the world seemed governed by eternal and deterministic laws, an appreciation for statistical descriptions started to blossom [Salsburg, 2001]. Many insights contributed to this shift: In the natural sciences, statistical mechanics and quantum mechanics revealed that even physics had to resort to stochastic descriptions. Another contribution was the slow but steady advance of statistics into the social sciences. Ever since Bernoulli in the 17th century, there was hope probability would help illuminate the functioning of society. While Laplace is often credited for one of the first successful applications of probability to matters of state – he estimated the

³⁶Galton called it regression to mediocrity.

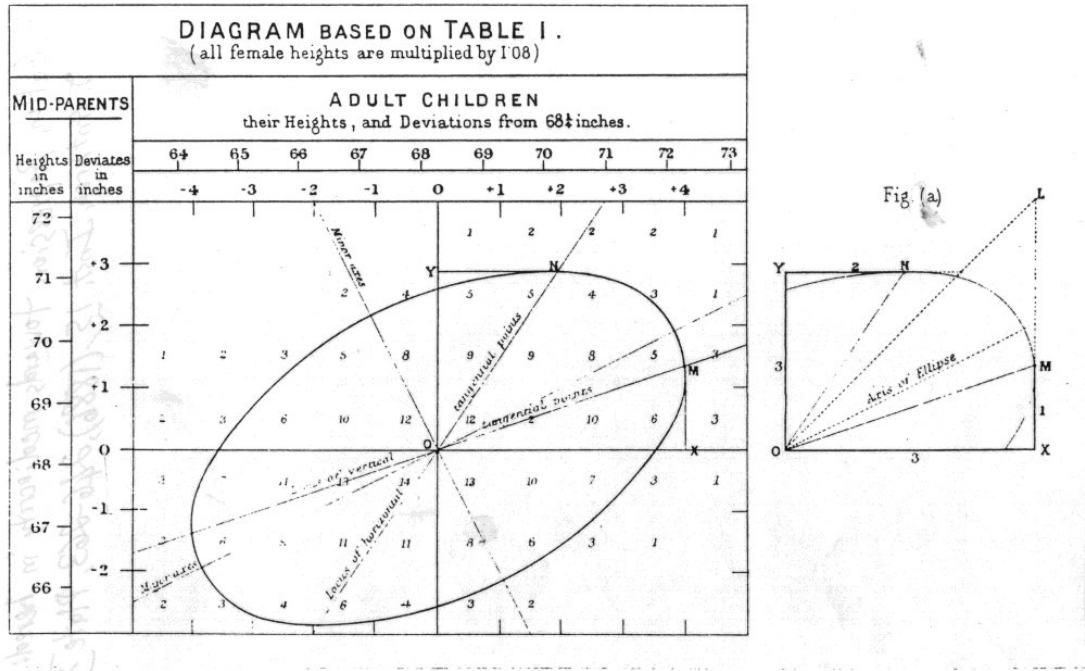


Figure 9: Galton's original correlation diagram: adult children's heights on the horizontal axis and mid-parental heights on the vertical axis. The ellipse is a contour of equal frequency; the two regression lines differ from its major axis. Source: Galton [1886, Plate X].

French population in 1802 based on extrapolating the ratio of birthrates to people from a few sampled areas to the entirety of the country – it took a few more decades and Quetelet's life-long dedication to convince the minds of his time that statistics could help explain society. Where before the human realm was a chaotic stage for individuals' free will, with the advent of the nation state and large scale population data, Quetelet managed to show that even the human world revealed order and stability. This order, which first appeared in the form of the surprising constancy of birth, death, marriage and crime rates, crucially, only emerged on the level of the *aggregate*. The laws of society seemed intrinsically statistical.

Where before, in a purely deterministic world, everything seemed to be necessitated by its cause, Karl Pearson tried to free science from these shackles of such a narrow notion of causality. In a statistical world, we cannot hope to uncover perfect causes that necessarily – and without fail – lead to an effect, but we also don't need to. To guide our actions, we only need to find stable (and meaningful) relationships between variables. These relationships can be weaker than those of necessity and still be meaningful.

Definition 6.1 (Bivariate Normal). A pair of random variables (X, Y) has a bivariate normal distribution with means $\mu_X, \mu_Y \in \mathbb{R}$, standard deviations $\sigma_X, \sigma_Y > 0$ and

correlation $-1 < \rho < 1$ if its joint density is

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} \right]} \quad (51)$$

for $x, y \in \mathbb{R}$.

Definition 6.2 (Pearson's correlation coefficient). For a pair of random variables (X, Y) , (Pearson's) correlation coefficient is given by

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}. \quad (52)$$

Pearson's correlation coefficient captures how two random variables (linearly) vary with each other. It ranges from -1 to 1, where at both of these two extremes there is a perfect linear relationship between the variables; knowing one lets you exactly determine the other. There are some pitfalls working with the correlation coefficient ρ because it only captures linear dependence. Most importantly, a correlation of 0 means that there is no linear relationship between the variables – we cannot conclude that they are independent.³⁷

Example 6.1 (Anscombe's quartet). Figure 10 shows four datasets constructed by Anscombe [1973] with almost identical means, variances and correlation coefficients (about 0.82). Yet the relationships look very different!

The history of ideas is a winding road. While Pearson tried to free science from the suffocating grip of causality (understood as necessity), much of the recent development of statistics has been focused on uncovering when we can meaningfully think of correlations as causal. We should not think of this as a step back but rather as a continuation of what Pearson set out to do. We will discuss this in more detail in Section 7.

6.3 Linear Regression: Disentangling Variation

To be written.

7 Causal Inference

To be written.

References

Ernest W Adams. *On Rational Betting Systems*. Archiv für mathematische Logik and Grundlagenforschung, 6: 7–29., 1962.

³⁷A correlation coefficient of 0 is only equivalent to independence for jointly normal random variables.

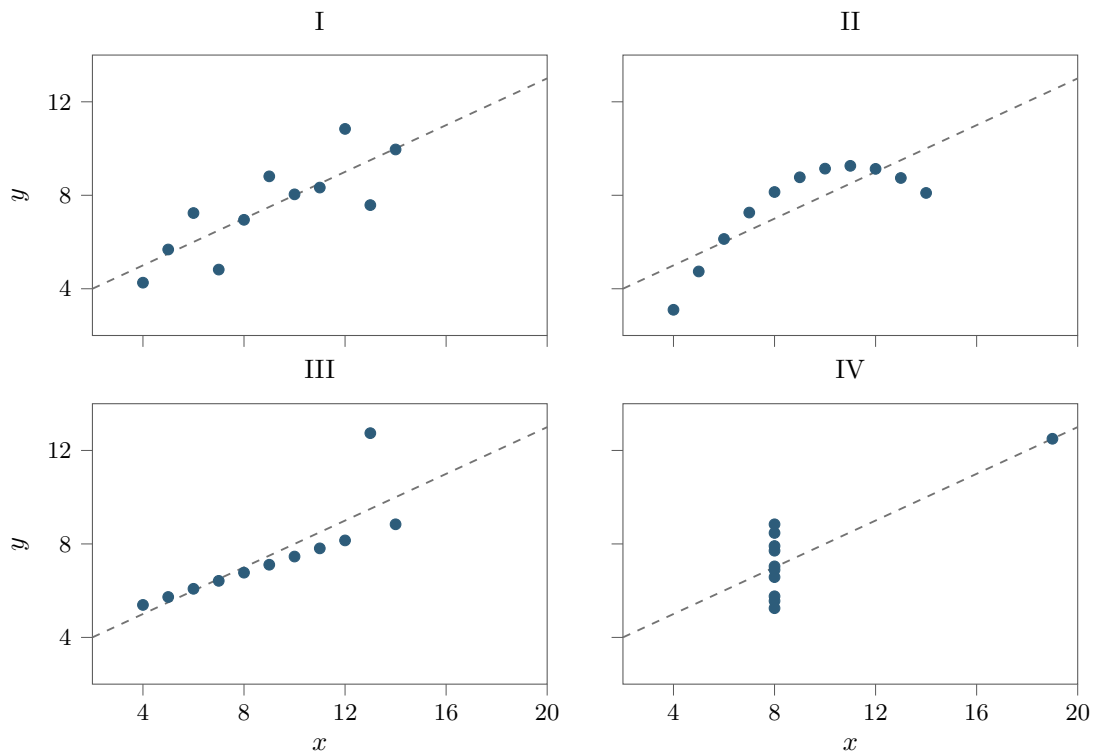


Figure 10: Anscombe's quartet. The lines show the least-squares fits.

F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, 1973. doi: 10.1080/00031305.1973.10478966.

John Arbuthnot. An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society of London*, 27:186–190, 1710. doi: 10.1098/rstl.1710.0011.

Keith A. Baggerly and Kevin R. Coombes. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*, 3(4):1309–1334, 2009. doi: 10.1214/09-AOAS291.

Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, 2016. doi: 10.1038/533452a.

Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. doi: 10.1098/rstl.1763.0053. Communicated by Richard Price.

James O Berger, Jose M Bernardo, and Dongchu Sun. *Objective Bayesian Inference*. WORLD SCIENTIFIC, 2024. doi: 10.1142/13640.

- George Casella and Roger Berger. *Statistical Inference (2nd edition)*. CRC Press, 2024.
- Kevin R. Coombes, Jing Wang, and Keith A. Baggerly. Microarrays: retracing steps. *Nature Medicine*, 13(11):1276–1277, 2007. doi: 10.1038/nm1107-1276b.
- Bruno de Finetti. *Theory of Probability (Volume 1)*. New York: John Wiley & Sons., 1990.
- A. De Morgan. *Formal Logic, or, The Calculus of Inference, Necessary and Probable*. London: Taylor and Walton., 1847.
- Alain Desrosières. *The Politics of Large Numbers*. Harvard University Press, 2002.
- Jay L. Devore. *Probability and Statistics for Engineering and the Sciences*. Cengage Learning, 9th edition, 2016. ISBN 9781305251809.
- Émile Durkheim. *Le suicide: etude de sociologie*. Collier-Macmillan, New York, 1897. English: Suicide: A Study in Sociology, 1951, Trans. John A. Spaulding and George Simpson.
- F. W. Dyson, A. S. Eddington, and C. Davidson. A determination of the deflection of light by the Sun’s gravitational field, from observations made at the total eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society of London. Series A*, 220:291–333, 1920. doi: 10.1098/rsta.1920.0009.
- William Feller. *An Introduction to Probability Theory and Its Applications, Volume 1, 3rd Edition*. Wiley Series in Probability and Statistics, 1991.
- Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 1922.
- Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886. doi: 10.2307/2841583.
- Francis Galton. *Natural Inheritance*. Macmillan, London, 1889. URL <https://galton.org/books/natural-inheritance/index.html>.
- Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 3rd edition, 2013.
- Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow, 2020. URL <https://arxiv.org/abs/2011.01808>.
- D. Kahneman, P. Slovic, and A. Tversky, editors. *Judgment Under Uncertainty. Heuristics and Biases*. Cambridge: Cambridge University Press, 1982.

- Imre Lakatos. Falsification and the methodology of scientific research programmes. In Imre Lakatos and Alan Musgrave, editors, *Criticism and the Growth of Knowledge*, pages 91–196. Cambridge University Press, Cambridge, 1970. doi: 10.1017/CBO9781139171434.009.
- E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, Cham, 4th edition, 2022. doi: 10.1007/978-3-030-70578-7.
- James Clerk Maxwell. Illustrations of the dynamical theory of gases. — part i: On the motions and collisions of perfectly elastic spheres. *Philosophical Magazine. Series 4. 19 (124): 19–32*, 1860.
- Deborah G. Mayo. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press, 2018. doi: 10.1017/9781107286184.
- Deborah G. Mayo and D. R. Cox. Frequentist statistics as a theory of inductive inference. In *2nd Lehmann Symposium – Optimality*, volume 49 of *Institute of Mathematical Statistics Lecture Notes–Monograph Series*, pages 77–97. Institute of Mathematical Statistics, 2006. doi: 10.1214/074921706000000400.
- Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015. doi: 10.1126/science.aac4716.
- Anil Potti, Holly K. Dressman, Andrea Bild, Richard F. Riedel, Gina Chan, Robyn Sayer, Janiel Cragun, Hope Cottrill, Michael J. Kelley, Rebecca Petersen, David Harpole, Jeffrey Marks, Andrew Berchuck, Geoffrey S. Ginsburg, Phillip Febbo, Johnathan Lancaster, and Joseph R. Nevins. Genomic signatures to guide the use of chemotherapeutics. *Nature Medicine*, 12(11):1294–1300, 2006. doi: 10.1038/nm1491. Retracted in 2011; retraction DOI: 10.1038/nm0111-135.
- Eva Ranehill, Anna Dreber, Magnus Johannesson, Susanne Leiberg, Sunhae Sul, and Roberto A. Weber. Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science*, 26(5): 653–656, 2015. doi: 10.1177/0956797614553946.
- Stuart J. Ritchie, Richard Wiseman, and Christopher C. French. Failing the future: Three unsuccessful attempts to replicate Bem’s ‘retroactive facilitation of recall’ effect. *PLoS ONE*, 7(3):e33423, 2012. doi: 10.1371/journal.pone.0033423.
- David Salsburg. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. W. H. Freeman, New York, 2001.
- James C. Scott. *Seeing Like a State*. Yale University Press, 1999.
- Stephen M. Stigler. *The History of Statistics*. Harvard University Press, 1990.
- Stephen M. Stigler. *The Seven Pillars of Statistical Wisdom*. Harvard University Press, Cambridge, MA, 2016. doi: 10.4159/9780674970199.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.