Probability Theory

Jacob Denson

October 22, 2017

Table Of Contents

Chapter 1

Foundations

These notes outline the basics of probability theory, the mathematical framework which allows us to interpret the statement that we are *80% more likely* to develop lung disease if you are smoker rather than if you are a non-smoker, or that there is a *50-50 chance* of rain on Saturday? These statements seem intuitive, but there a couple difficulties. On Saturday, it will either rain, or not rain, so how can there be a 'chance' of either event happening at all? To a mathematician, these statements have a rigorous interpretation in the language of measure theory. To a natural sciences, probability theory is viewed in a different vein. In this chapter, we will explore the two major interpretations of probability theory in real life, each of which use the same underlying mathematical theory to make judgements about the world. After exploring these interpretations, we will make axiomatic definitions of probability (which hold regardless of which interpretation you have), and explore the basic consequences of the assumptions.

1.1 Frequentist Probability

Classical probability theory was developed according to the intuitions of what is now known as the frequentist school of probability theory, and is the simplest interpretation of probability to understand. It is easiest to understand from the point of view of a scientist. Suppose you are repeatedly performing some well-controlled experiment, in the sense that you do not expect the outcome of the experiment to change drastically between trials. Even under rigorously controlled conditions, the experiment will not always result in the same outcome. Slight experimental error results in slight changes in the outcome of the experiment. Nonetheless, some outcomes will occur more frequently than others. Let us perform an experiment as often as desired, obtaining an infinite sequence of outcomes

$$\omega_1, \omega_2, \omega_3, \dots$$

Let D be a certain question, or *proposition* about the outcome of the experiment (for instance, D may ask whether a flipped coin lands heads up when flipping a coin repeatedly). Mathematically, we can represent the proposition as a subset of the set of all outcomes in an experiment – the outcomes for which the proposition is true. We define the *relative frequency* of D being true in n trials by the equation

$$P_n(D) := \frac{\#\{k \leqslant n : \omega_k \in D\}}{n}$$

The key assumption of the frequentist school of probability is that, if our experiments are suitably controlled, then regardless of the specific sequence of measured outcomes, our relative frequencies will always converge to a well defined invariant ratio, which we define to be the probability of a certain event:

$$\mathbf{P}(D) := \lim_{n \to \infty} P_n(D)$$

Let's explore some consequences of this doctrine. First, $0 \leqslant P_n(D) \leqslant 1$ is true for any n, so that for any proposition D,

$$0 \leqslant \mathbf{P}(D) \leqslant 1 \tag{1.1}$$

If we let Ω denote the set of all possible outcomes to the experiment (a proposition true for all outcomes of the experiment), then

$$P_n(\Omega) = \frac{\#\{k \leqslant n : \omega_k \in \Omega\}}{n} = \frac{\#\{1, 2, \dots, n\}}{n} = 1$$

Thus we conclude

$$\mathbf{P}(\Omega) = 1 \tag{1.2}$$

If $A_1, A_2, \dots$ is a sequence of disjoint propositions (no more than one outcome is true for each outcome of the experiment), then

$$P_n\left(\bigcup_i A_i\right) = \frac{\#\{k \leqslant n : \omega_k \in \bigcup A_i\}}{n} = \frac{\sum_i \#\{k \leqslant n : \omega_k \in A_i\}}{n} = \sum_i P_n(A_i)$$

Hence,

$$\mathbf{P}\left(\bigcup_i A_i\right) = \sum_i \mathbf{P}(A_i) \tag{1.3}$$

This equation still holds for an arbitrary family of disjoint propositions, provided we interpret the sum of the propositions as the supremum of all finite sums. There is no real generality here, because only countably many disjoint propositions can be true in the sequence of experimental outcomes (for only one proposition can be true for each of the experiments), hence the probability of only countably many propositions is nonzero. Rules (1.1), (1.3), and (1.2) turn out to be sufficient to describe all the mathematically important rules of frequentist probability. What's more, we can use these rules to *prove* that the probability of a sequence of controlled experiments eventually settles down, which justifies the thought process of the frequentist school in the first place.

1.2 Bayesian Probability

The frequentist school is sufficient to use probability theory to model scientific experiments, but in everyday life we make a much greater use of probabilistic language. If you turn on the news, it's common to hear that "there is an 80% chance of downpour this evening". It is difficult to interpret this result in the frequentist definition of probability. Even if we see each night's temperament as an experimental trial, it is hard to convince yourself that these experiments are controlled enough to converge to a probabilistic result. The Bayesian school of probability redefines probability theory to be attuned to a person's individual beliefs, so that we can interpret "there is an 80% chance of downpour this evening" as an individual's belief that they think it will rain this evening rather than not rain, and that they're willing to bet on it as well!

You might argue that, if probability is a personal belief in an unknown event, we can choose probabilities however we want, and this would break down the logical structural required for a mathematical theory of probability. However, the probabilities that the Bayesian school studies are forced to be 'logically consistant'. Consistancy can be formulated in various ways, but my favourite is the dutch book method, developed by the Italian probabilist Bruno de Finetti; if you assign to a certain unknown event D a probability $\mathbf{P}(D)$, then you are willing to make a bet at $[\mathbf{P}(D) : 1-\mathbf{P}(D)]$ odds, playing the following game: If D occurs, you win $1-\mathbf{P}(D)$ dollars, but if D does not occur, you have to pay up $\mathbf{P}(D)$ dollars. You *must* also be willing to play the game where you lose $1-\mathbf{P}(D)$ dollars if D occurs, and gain $\mathbf{P}(D)$ dollars if D does not occur, so that you think the bets are 'fair' to both sides. For instance, you might be willing to bet a dollar against a dollar that a coin will turn up heads, which is $[1 : 1] = [1/2 : 1/2]$ odds, so we would assign the probability that a coin will turn up heads as $1/2$, because then we win or lose an equal amount depending on the outcome of the bet. A person's probability function is inconsistant if it possible to make a series of bets that will guarantee a profit regardless of the outcome; this is known as a dutch book.

Here's an example of how the Dutch book method can be employed to obtain general rules of probability. We claim that for any event D, $0 \leqslant \mathbf{P}(D) \leqslant 1$. If a person believed that $\mathbf{P}(D) < 0$, then I could make a bet that person that D occured, and I would make money regardless of the outcome. Similar results occur from betting against D if $\mathbf{P}(D) > 1$. It can be shown, via similar arguments, that (1.1), (1.2), and (1.3) hold for any logically consistant Bayesian choice of probabilities (Definetti would have only allowed finitely many bets at once, which means that he would only accent (1.3) for finite sums, but here we allow countably many bets to be made at once – allowing limit operations is too useful to ignore!). What this means is that, regardless of whether you think that probabilities are a measure of 'degrees of belief' in an event happening, or the experiment frequencies of an experiment, then you still believe in the same laws of probability. Regardless of which philosophy you agree with, the fundamental principles of probability theory remain the same. We shall take the three laws we derived, and use it to make a rigorous model so that we can avoid future philosophical controversies, and this is where mathematical probability theory takes its form.

1.3 Axioms of Probability

Mathematically ,rigorous probability theory is defined under the banner of measure theory. The framework enables us to avoid some paradoxes which can be found if we aren't careful when analyzing experiments with infinitely many outcomes. Note, however, that the focus of probability theory is on events and random quantities (what we will soon refer to as random variables), rather than on focusing on a particular measure space under questions. Probability theorists focus on studying these *concepts*, and the framework provides the formality to understand these concepts. A **probability space** is a measure space Ω with a positive measure $\mathbf{P}$ such that $\mathbf{P}(\Omega) = 1$. To the non-initiated, this means that there is a function $\mathbf{P}$, mapping from subsets of X to numbers in $[0,1]$, satisfying the properties (1.1), (1.2), and (1.3) defined above, except that only subsets of Ω in a given σ *algebra* may have there probability measured. That is, we can only guarantee that we can measure the probability of Ω itself, that if we can measure the probability of some event E, then we can measure the probability of E^c, and that if we can measure the probability of $E_1, E_2, \dots,$ then we can measure the probability of $\bigcup E_i$. The reason for this is that, in certain more complex spaces, especially those occuring in infinite spaces, paradoxes occur if we try and measure the probability of all subsets of the space at once. Ω is known as the **sample space**, and $\mathbf{P}$ is known as the **probability distribution** or **probability measure**. We interpret Ω as the space of outcomes to some random phenomena, and $\mathbf{P}$ measuring the likelyhood of each outcome happening.

Example. *Suppose we flip a coin. There is a certain chance of flipping a heads, or flipping a tails. Since the coin is essentially symmetric, we should expect that the chance of a heads is as equally likely as a chance of tails. We can encode the set of outcomes in the sample space $\{H, T\}$, and then model the probability distribution as $\mathbf{P}(H) = \mathbf{P}(T) = 1/2$. More generally, if we have a finite sample space S, we can put a distribution on S which considers all points equally known as the* **uniform distribution**, *with distribution $\mathbf{P}(s) = 1/|S|$.*

Example. *If $\omega \in \Omega$ is fixed, the* **point mass distribution** *δ_ω at ω is the probability distribution defined by*

$$\delta_\omega(E) = \begin{cases} 1 & \omega \in E \\ 0 & \omega \notin E \end{cases}$$

The distribution represents an event where an outcome is certain to occur.

Example. *If Ω is a countable set, then every probability measure every subset of Ω (by completion of the underlying set), and we can view the probability* *measure as a member of the set*

$$\left\{ v \in [0,1]^\Omega : \sum_{\omega \in \Omega} v(\omega) = 1 \right\}$$

This can be viewed as a subset of the convex subset of the unit ball in $l^\infty(\Omega)$, and there is some interesting linear analysis in the theory of such sets. Most importantly, the σ algebra plays no real role in the theory until later on, and need not be introduced until we study probability distributions over stranger spaces, like the uncountable unit interval $[0,1]$.

The first immediately obvious fact from the axioms is that

$$\mathbf{P}(E^c) = 1 - \mathbf{P}(E)$$

since E and E^c are disjoint events whose union is Ω. A similar discussion shows that

$$\mathbf{P}(E \cup F) = \mathbf{P}(E) + \mathbf{P}(F) - \mathbf{P}(E \cap F)$$

because $E \cup F$ can be written as the union of the three disjoint events $E \cap F$, $E \cap F^c$, and $E^c \cap F$, and

$$\mathbf{P}(E) = \mathbf{P}(E \cap F) + \mathbf{P}(E \cap F^c) \qquad \mathbf{P}(F) = \mathbf{P}(E \cap F) \cup \mathbf{P}(E^c \cap F)$$

This process can be generalized to unions of finitely many events. We have

$$\mathbf{P}(E \cup F \cup G) = \mathbf{P}(E) + \mathbf{P}(F) + \mathbf{P}(G) - \mathbf{P}(E \cap F) - \mathbf{P}(E \cap G) - \mathbf{P}(F \cap G) + \mathbf{P}(E \cap F \cap G)$$

which can be reasoned by looking at the number of times each element of $E \cup F \cup G$ is 'counted' on the right hand side. In general, we have the inclusion-exclusion principle

$$\mathbf{P}\left(\bigcup_{k=1}^n E_k \right) = \sum_{S \subset \{1,\dots,k\}} (-1)^{|S|} \mathbf{P}\left(\bigcap_{k \in S} E_k \right)$$

This can be proven by a clumsy inductive calculation. More interestingly, we often want to calculate the probability of an infinite union of sets E_k occuring. The inclusion-exclusion principle can be taken 'in the limit' to conclude that

$$\mathbf{P}\left(\bigcup_{k=1}^\infty E_k \right) = \lim_{n \to \infty} \mathbf{P}\left(\bigcup_{k=1}^n E_k \right) = \sum_{\substack{S \subset \mathbf{N} \\ |S| < \infty}} (-1)^{|S|} \mathbf{P}\left(\bigcap_{k \in S} E_k \right)$$

where the sum on the right is taken as the limit of the partial sums where $S \subset \{1, \dots, n\}$ (the sum rarely convergences absolutely, so it is important to take the limit in this order).

The inclusion-exclusion formula can be tricky to calculate in real examples, so we often rely on estimates to upper bound or lower the probability of a particular event occuring. The trivial **union bound**

$$\mathbf{P}\left(\bigcup E_i\right) \leqslant \sum \mathbf{P}(E_i)$$

can be applied. This is a good inequality to apply if the E_i are 'nearly disjoint' (the bound is shockingly bad if all the E_i are equal to one another), or if the event E_k occur with such a small probability that we don't mind the resultant inaccuracy.

Another useful fact to consider is that $\mathbf{P}(E_k) \to \mathbf{P}(E)$ if the sets E_k 'tend to' E in one form of another. If the E_k are an increasing sequence whose union is E, then we can certainly conclude $\mathbf{P}(E_k) \to \mathbf{P}(E)$. Similarily, if E_k is a decreasing sequence whose intersection is E, then $\mathbf{P}(E_k) \to \mathbf{P}(E)$. To obtain general results, we say that $E_k \to E$ if $\limsup E_k = \liminf E_k = E$, where

$$\limsup_{k \to \infty} E_k = \bigcap_{n=1}^{\infty} \bigcup_{k \geqslant n} E_k = \{\omega : \omega \in E_k \text{ for infinitely many } k\}$$

$$\liminf_{k \to \infty} E_k = \bigcup_{n=1}^{\infty} \bigcap_{k \geqslant n} E_k = \{\omega : \omega \in E_k \text{ for sufficiently large } k\}$$

We can then conclude that $\mathbf{P}(E_i) \to \mathbf{P}(E)$, since once can show

$$\limsup \mathbf{P}(E_k) \leqslant \mathbf{P}\left(\limsup E_k\right)$$

$$\liminf \mathbf{P}(E_k) \geqslant \mathbf{P}\left(\liminf E_k\right)$$

so we can apply the squeeze theorem. This already enables us to prove a very interesting theorem which can guarantee an event can 'never occur'.

Lemma 1.1 (Borel-Cantelli Lemma)**.** *If $E_1, E_2, \dots$ is a sequence of events with $\sum \mathbf{P}(E_k) < \infty$, then $\mathbf{P}\left(\limsup E_k\right) = 0$. Thus none of the events E_k can happen 'infinitely often'.-*

Proof. Note that because

$$\mathbf{P}\left(\bigcup_{k \geqslant n} E_k\right) \leqslant \sum_{k \geqslant n} \mathbf{P}(E_k)$$

for any $\varepsilon > 0$ we can find an N such that for $n \geqslant N$, $\mathbf{P}\left(\bigcup_{k \geqslant n} E_k\right) < \varepsilon$. But for any n, $\limsup E_k \subset \bigcup_{k \geqslant n} E_k$, and so we conclude $\mathbf{P}(\limsup E_k) < \varepsilon$. We then let $\varepsilon \to 0$ to conclude $\mathbf{P}(\limsup E_k) = 0$. □

1.4 Conditional Probabilities

In the Bayesian interpretation of probability theory, it is natural for probabilities to change over time as more information is gained about the system in question. That is, given that we know some proposition F holds over the sample space, we obtain a new probability distribution over Ω, denoted $\mathbf{P}(\cdot|F)$, which represents the ratio of winnings from the bet which is only played out if F occurs. That is

- You win $1 - \mathbf{P}(E|F)$ dollars if E occurs, and F occurs.
- You lose $\mathbf{P}(E|F)$ dollars if E does not occur, and F occurs.
- No money exchanges hands if F does not occur.

It then follows from a dutch book argument that

$$\mathbf{P}(F)\mathbf{P}(E|F) = \mathbf{P}(E \cap F)$$

TODO: Fillin this argument. In the emperical interpretation, $\mathbf{P}(E|F)$ is the ratio of times that E is true in experiments, where we only count experiments in which F also occurs. That is, we define $\mathbf{P}(E|F)$ as the limit of the ratios

$$P_n(E|F) = \frac{\#\{k \leqslant n : \omega_k \in E, \omega_k \in F\}}{\#\{k \leqslant n : \omega_k \in F\}}$$

But it is easy to calculate, by dividing the numerator and denominator by n, that $P_n(E|F) = P_n(E \cap F)/P_n(F)$, so by taking limits, we find

$$\mathbf{P}(E|F) = \lim_{n \to \infty} \frac{P_n(E \cap F)}{P_n(F)} = \frac{\mathbf{P}(E \cap F)}{\mathbf{P}(F)}$$

which gives us the formula $\mathbf{P}(F)\mathbf{P}(E|F) = \mathbf{P}(E \cap F)$. We must of course assume that $\mathbf{P}(F) \geqslant 0$, since overwise we are almost certain that F will never occur, and then we can almost guarantee that the limit of the values $P_n(E|F)$ does not exist.

Thus we have motivation to define conditional probabilities by the formula $\mathbf{P}(F)\mathbf{P}(E|F) = \mathbf{P}(E \cap F)$, provided that $\mathbf{P}(F) > 0$. It enables us to model the information gained by restricting our knowledge to a particular subset of sample space. In particular, we can use the definition to identify events which contain information 'useless' to learning about another event. We say two events E and F are independant if $\mathbf{P}(E \cap F) = \mathbf{P}(E)\mathbf{P}(F)$, or, provided $\mathbf{P}(F) > 0$, $\mathbf{P}(E|F) = \mathbf{P}(F)$; knowledge of F gives us no foothold over knowledge of the likelihood of E.

Example. *The Monty Hall problem is an incredible example of how paradoxical probability theory can seem. We are on a gameshow. Suppose there are three doors in front of you. A car (brand new!) is placed uniformly randomly behind one of the doors. After we pick a door (the first door, for instance), the gameshow host then opens the second door, which you didn't pick, revealing the car isn't behind the door. It is important to note that he picked randomly from the remaining doors which you didn't pick and don't have a car behind them. What is the chance that the door you picked has the brand new car? You likely would think the two doors have a 50-50 chance of containing the car given this info, but you'd be wrong. Let $X \in \{1,2,3\}$ denote the door chosen uniformly at random where the car lies, and let $Y \in \{1,2,3\}$ denote the door that the host randomly chose to open. We know $Y \neq 1$, because the gameshow host would never open the door we picked; that would give the game away! If $X = 1$, then Y is picked from $\{2,3\}$ with uniform possibility. However, if $X = 2$, something interesting occurs – the gameshow is forced to open door number 3, because that's the only door that (he thinks) won't give any information to the player, and similarily, if $X = 3$, then $Y = 2$. Now we know that since X is chosen uniformly at random $\mathbf{P}(X = k) = 1/3$ for each k. Similarily, we know that Y is then chosen uniformly at random from $\{2,3\}$, given that $X = 1$, so assuming X and Y are independent, we conclude*

$$\mathbf{P}(X = 1, Y = 2) = \mathbf{P}(X = 1)\mathbf{P}(Y = 2) = 1/6$$

$$\mathbf{P}(X = 1, Y = 3) = 1/6$$

But we also know that if $X = 2$, then $Y = 3$, so

$$\mathbf{P}(X = 2, Y = 3) = \mathbf{P}(X = 2) = 1/3$$

$$\mathbf{P}(X = 3, Y = 2) = \mathbf{P}(X = 3) = 1/3$$

It follows that

$$\begin{aligned}&\mathbf{P}(\textit{door } 1 \textit{ has a car}|\textit{door } 2 \textit{ was opened})\\&\qquad = \frac{\mathbf{P}(\textit{door } 1 \textit{ has a car}, \textit{door } 2 \textit{ was opened})}{\mathbf{P}(\textit{door } 2 \textit{ was opened})}\\&\qquad = \frac{\mathbf{P}(\{(X = 1, Y = 2)\})}{\mathbf{P}(\{(X = 1, Y = 2), (X = 3, Y = 2)\})} = \frac{1/6}{1/6 + 1/3} = 1/3\end{aligned}$$

This means we should definitely change our minds about which door we were going to pick! The argument above causes a great media uproar when it was published in 1990 in a popular magazine, because of how convincing the fallacious argument below is. The total number of possibilities is

$$(X = 1, Y = 2), (X = 1, Y = 3), (X = 2, Y = 3), (X = 3, Y = 2)$$

and the car seems to be in door one half of the possibilities. However, these events do not have the same probability of occuring. However, if the host changes his strategy, the conditional probabilities fall more in line with intuition – if the host always picks door number 2 to open if door number 1 was picked and had the car behind it, then the two remaining doors have an equal chance of being picked.

We end this chapter with a final probability rule which is important in statistical analysis. If B is partitioned into a finite sequence of disjoint events $A_1, \dots, A_n$, then we have the formula $\mathbf{P}(B) = \sum_i \mathbf{P}(B|A_i)\mathbf{P}(A_i)$. This easily gives us Bayes rule

$$\mathbf{P}(A_j|B) = \frac{\mathbf{P}(B|A_j)}{\sum_i \mathbf{P}(B|A_i)\mathbf{P}(A_i)}$$

If we view A_j as a particular hypothesis from the set of all hypotheses, and B as some obtained data, then Bayes rule enables us to compute the probability that A_j is the true hypothesis from the probability that B is the data generated given the hypothesis is true. This is incredibly important if you can interpret these probabilities correctly (if you are a Bayesian), but not so useful if you are an empiricist (in which case we assume there is a 'true' result we are attempting to estimate from trials, so there is no probability distribution over the correctness hypothesis, other than perhaps a point mass, in which case Bayes rule gives us no information). We reiterate that Bayes rule is a theorem of probability theory, so is true in any interpretation, but can be used by Bayesians in a much more applicable way to their statistical analysis.

1.5 Kolmogorov's Zero-One Law

s

Chapter 2

Random Variables

The formality of probability theory is ironic, because even though we require the theory of measures and real analysis to place the foundations of the theory, in the probabilistic way of thinking we try to eschew as much of this foundation as possible; studying properties of random variables which aren't 'independent' of the sample space considered is avoided. As a rough approximation, if $T : X \to Y$ is a surjective measure preserving map between probability spaces (X is an extension of the space Y, allowing more outcomes), then the random variable $Y \circ T$ is considered the 'same' as the random variable Y, and the concepts studied in probability theory should be preserved under this extension. As we reach further and further into statistical theory, samples spaces will soon become a near distant memory.

The irony of introducing the sample space is unfortunate, because while the space is in the background, in the probabilistic way of thinking about problems we try and eschew the sample space as much as possible.

2.1 Expectation

Theorem 2.1. *For any $X \geqslant 0$,*

$$\mathbf{E}[X] = \int \mathbf{P}(X \geqslant x) dx$$

Proof. Applying Fubini's theorem,

$$\begin{aligned}\int_0^\infty \mathbf{P}(X \geqslant x)dx &= \int_0^\infty \int_x^\infty d\mathbf{P}_*(y)\ dx \\ &= \int_0^\infty \int_0^y dx\ d\mathbf{P}_*(y) \\ &= \int_0^\infty y d\mathbf{P}_*(y) = \mathbf{E}[X]\end{aligned}$$

□

Chapter 3

Inequalities

It is often to calculate explicitly the probability values of a certain random variable, but it often suffices to bound the values, especially when discussing the convergence of certain variables.

The most important inequality bounds the chance that a probability will deviate from the mean. if X has mean $\mu < \infty$, then the Lebesgue integral calculates

$$\mu = \int X d\mathbf{P}$$

as the supremum of step functions. In particular, if we take the step function $x\mathbf{I}(X \geqslant x) \leqslant X$, then we find

Theorem 3.1 (Markov's Inequality)**.** *If X has finite mean μ, then*

$$\mathbf{P}(X \geqslant x) \leqslant \frac{\mathbf{E}(X)}{x}$$

The bound is trivial, and is therefore very rough. Nonetheless, it suffices for many purposes. One can obtain better estimates by taking a more detailed step function bounded by X, but the payoff isn't normally that great. We obtain a somewhat sharper estimate if X has a finite variance σ.

Theorem 3.2 (Chebyshev's Inequality)**.** *If X has mean μ and variance σ^2, then*

$$\mathbf{P}(|X - \mu| \geqslant x) \leqslant \frac{\sigma^2}{x^2}$$

If $Z = (X - \mu)/\sigma$, then

$$\mathbf{P}(|Z| \geqslant x) \leqslant \frac{1}{x^2}$$

Proof. Applying Markov's inequality, we find

$$\mathbf{P}(|X-\mu| \geqslant x) = \mathbf{P}(|X-\mu|^2 \geqslant x^2) \leqslant \frac{\mathbf{E}|X-\mu|^2}{x^2} = \frac{\sigma^2}{x^2}$$

We obtain the inequality for Z by carrying out coefficents and applying Chebyshev's inequality. □

We can continue this process. When X has an n'th moment, then

$$\mathbf{P}(|X-\mu| \geqslant x) \leqslant \frac{\mathbf{E}|X-\mu|^n}{x^n}$$

which shows that the existence of moments guarantees the decay of X. It is often difficult to calculate high degree moments, however, so this inequality does not occur as often.

Example. *Let $X_1, \dots, X_n \sim Ber(p)$ by independant and identically distribution, where p is an unknown value. A good way to estimate p is via the random variable*

$$\widehat{p} = \frac{X_1 + \dots + X_n}{n}$$

which has a binomial distribution. We measure the utility of $\widehat{p}$ by minimizing the probability that $\widehat{p}$ deviates far from the mean. That is, $\mathbf{P}(|\widehat{p} - p| \geqslant x)$ is small for large values of x. We find $\widehat{p}$ has mean p and variance $p(1-p)/n$, so we may apply Chebyshev's inequality to conclude

$$\mathbf{P}(|\widehat{p} - p| \geqslant x) \leqslant \frac{p(1-p)}{nx^2} \leqslant \frac{1}{4nx^2}$$

so even for the worst possible choice of p, we still obtain inversely linear decay; not great, but still enough to guarantee that $\widehat{p}$ converges in distribution to the point mass measure at p as $n \to \infty$. This is the weak law of large numbers for the Bernoulli distribution.

Measure theory gives us general bounds, which are just special results of more general inequalities. We have the Cauchy Schwarz inequality, which says $\mathbf{E}(XY) \leqslant \sqrt{\mathbf{E}(X^2)\mathbf{E}(Y^2)}$, and Jensen's inequality, which says that is f is convex, then $f(\mathbf{E}(X)) \leqslant \mathbf{E}(f(X))$. If f is concave, $f(\mathbf{E}(X)) \geqslant \mathbf{E}(f(X))$. In particular, Jensen's inequality shows

$$\mathbf{E}(X^2) \geqslant [\mathbf{E}(X)]^2 \qquad \mathbf{E}(1/X) \geqslant 1/\mathbf{E}(X) \qquad \mathbf{E}(\log x) \leqslant \log \mathbf{E}(X)$$

which is used in the more advanced theory to obtain deeper inequalities.

Hoeffding's inequality is similar to Markov's inequality, but is generally much sharper. It therefore has a more complicated formula.

Theorem 3.3 (Hoeffding's Inequality). *Let $X_1, \dots, X_n$ be centrally distributed i.i.d random variables, with $a_i \leqslant X_i \leqslant b_i$, then for any $t > 0$,*

$$\mathbf{P}\left(\sum X_i \geqslant x\right) \leqslant e^{-tx} \prod_{i=1}^{n} e^{t^2(b_i - a_i)^2/8}$$

Proof. For any $t > 0$, Markov's inequality implies

$$\begin{aligned} \mathbf{P}\left(\sum X_i \geqslant x\right) &= \mathbf{P}\left(t \sum X_i \geqslant tx\right) \\ &= \mathbf{P}\left(e^{t \sum X_i} \geqslant e^{tx}\right) \\ &\leqslant e^{-tx} \prod \mathbf{E}[e^{tX_i}] \end{aligned}$$

We can write

$$X_i = \Lambda a_i + (1 - \Lambda) b_i$$

for some function $0 \leqslant \Lambda \leqslant 1$, and by applying the convexity of the exponential function, we find

$$e^{tX_i} \leqslant \Lambda e^{ta_i} + (1 - \Lambda) e^{tb_i}$$

Hence

$$\mathbf{E}(e^{tX_i}) \leqslant \mathbf{E}(\Lambda) e^{ta_i} + (1 - \mathbf{E}(\Lambda)) e^{tb_i}$$

Now we may explicitly calculate $\Lambda = (X_i - a_i)/(b_i - a_i)$, so that

$$\mathbf{E}(e^{tX_i}) \leqslant \frac{a_i}{a_i - b_i} e^{ta_i} + \frac{b_i}{b_i - a_i} e^{tb_i} = e^{F(t(b_i - a_i))}$$

Where $F(x) = -\lambda x + \log(1 - \lambda + \lambda e^x)$, where $\lambda = a_i/(a_i - b_i)$. Note that $F(0) = F'(0) = 0$, and $F''(x) \leqslant 1/4$ for $x > 0$, so that by Taylor's theorem, there is $y \in (0, x)$ such that

$$F(x) = \frac{x^2}{2} g''(y) \leqslant \frac{x^2}{8}$$

Hence $\mathbf{E}(e^{tX_i}) \leqslant e^{t^2(b_i - a_i)^2/8}$, and this completes the proof. □

Example. *If $\widehat{p} \sim Bin(n,p)$, and we take $X_i = (\widehat{p} - p)/n$, then $\mathbf{E}(X_i) = 0$, and $-p/n \leqslant X_i \leqslant 1/n - p/n$, and since $\sum X_i = \widehat{p} - p$, we find*

$$\mathbf{P}(\widehat{p} - p \geqslant x) \leqslant e^{t^2/8n - tx}$$

For $t = 4nx$, we find $\mathbf{P}(\widehat{p} - p \geqslant x) \leqslant e^{-2nx^2}$. By symmetry, we can calculate the absolute deviance as $\mathbf{P}(|\widehat{p} - p| \geqslant x) \leqslant 2e^{-2nx^2}$. This gives us a much sharper rate of convergence than our last result.

3.1 Subgaussian Random Variables

Hoeffding's inequality only applies to bounded random variables. In the general case, we can't apply the inequality (which relies on the bounded intervals to use convexity), and Chebyshev's inequality often does not suffice. We should still obtain fast tail decay in most circumstances, say, for instance a Gaussian distribution with variance σ^2. Calculating, we find

$$\begin{aligned} \mathbf{P}(X - \mu \geqslant y) &= \int_y^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx \\ &\leqslant \frac{1}{y\sqrt{2\pi\sigma^2}} \int_y^\infty x e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \frac{\sigma e^{-\frac{y^2}{2\sigma^2}}}{y\sqrt{2\pi}} \end{aligned}$$

This quantity is almost always better than Chebyshev's inequality, since the ratio x/y, which measures the inaccuracy of our inequality, is nullified by the exponential function. We can find similar equalities for random variables which are 'bounded' by normal distributions.

We shall say a random variable X is σ^2-**subgaussian** if for all $\lambda \in \mathbf{R}$,

$$\mathbf{E}[e^{\lambda X}] \leqslant e^{\lambda^2 \sigma^2 / 2}$$

where we assume $e^{\lambda X}$ is integrable for all λ. Pointwise, we have

$$e^{\lambda X} = \sum_{k=0}^\infty \frac{\lambda^k X^k}{k!} \leqslant \sum_{k=0}^\infty \frac{\lambda^{2k} \sigma^{2k}}{2^k k!} = e^{\lambda^2 \sigma^2 / 2}$$

since $e^{|\lambda||X|}$ is integrable ($|X| = X^+ + X^-$, and the Cauchy Schwarz equality implies)

$$\mathbf{E}[e^{|\lambda|X^+} e^{|\lambda|X^-}]^2 \leqslant \mathbf{E}[e^{2|\lambda|X^+}]\mathbf{E}[e^{2\lambda X^-}] < \infty$$

Thus we may apply the dominated convergence theorem to conclude

$$\mathbf{E}[e^{\lambda X}] = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \mathbf{E}[X^k]$$

and for any λ,

$$\sum_{k=0}^{\infty} \frac{\mu^k}{k!} \mathbf{E}[X^k] \leqslant \sum_{k=0}^{\infty} \frac{\mu^{2k}\sigma^{2k}}{2^k k!}$$

If $\mathbf{E}[X] > 0$, we may subtract by one and divide by $\mu\mathbf{E}[X]$ to conclude that for $\mu > 0$

$$1 + \mu\frac{\mathbf{E}[X^2]}{2\mathbf{E}[X]} + \dots \leqslant \mu\frac{\sigma^2}{2} + \dots$$

and if we take $\mu \to 0$, we obtain $1 \leqslant 0$, a contradiction. If $\mathbf{E}[X] < 0$, the same equation holds for $\mu < 0$, so we must have $\mathbf{E}[X] = 0$. Similarily, the bound $\mathbf{V}[X] \leqslant \sigma^2$ is obtained by comparing coefficients.

Example. *If X is a symmetric Bernoulli random variable with*

$$\mathbf{P}(X = -1) = \mathbf{P}(X = -1) = 1/2$$

We have

$$\mathbf{E}[e^{\lambda X}] = \frac{e^{\lambda} + e^{-\lambda}}{2} = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \leqslant \sum_{k=0}^{\infty} \frac{(\lambda^2)^k}{2^k k!} = e^{\lambda^2/2}$$

so X is a 1 *subgaussian random variable.*

Example. *If X is uniformly distributed on $[-n, n]$, then*

$$\mathbf{E}[X^k] = \int_{-n}^{n} \frac{x^k}{2n} dx = \frac{n^{k+1} - (-n)^{k+1}}{(k+1)2n} = \begin{cases} \frac{n^k}{k+1} & k \textit{ even} \\ 0 & k \textit{ odd} \end{cases}$$

So

$$\mathbf{E}[e^{\lambda X}] = \sum_{k=0}^{\infty} \frac{n^{2k}\lambda^{2k}}{(2k+1)(2k)!} \leqslant \sum_{k=0}^{\infty} \frac{n^{2k}\lambda^{2k}}{2^k k!} = e^{n^2\lambda^2/2}$$

so X is n-subgaussian.

Example. *In general, if a centrally distributed random variable X satisfies $|X| \leqslant M$ almost surely, then X is M^2 subgaussian. Assume without loss of generality that $M = 1$. Set $Y = X + 1$, and*

$$f(t) = \frac{e^{2t} + 1}{2} - \mathbf{E}(e^{tY})$$

Since $\mathbf{E}(Y) = 1$,

$$f'(t) = \mathbf{E}(Y[e^{2t} - e^{tY}])$$

since $Y \leqslant 2$ almost surely, $f'(t) \geqslant 0$, and so f is increasing. In particular, $f(0) = 1 - 1 = 0$, so that for $t \geqslant 0$, -

$$\mathbf{E}(e^{tX}) = e^{-t}\,\mathbf{E}(e^{tY}) \leqslant \frac{e^t + e^{-t}}{2} \leqslant e^{t^2/2}$$

Since we can perform the same argument for $-X$, we see that X is 1 subgaussian.

The set of subgaussian random variables form a vector space. If X is a σ^2 subgaussian random variable, then cX is $(c\sigma)^2$ subgaussian. If Y is τ^2 subgaussian, then, using the Hölder inequality, we find that for $p^{-1} + q^{-1} = 1$,

$$\mathbf{E}[e^{\lambda(X+Y)}] = \mathbf{E}[e^{\lambda X} e^{\lambda Y}] \leqslant \mathbf{E}[e^{p\lambda X}]^{p^{-1}} \mathbf{E}[e^{q\lambda X}]^{q^{-1}} \leqslant e^{\frac{\lambda^2\sigma^2}{2}} e^{\frac{\tau^2 q}{2}} = e^{\frac{\lambda^2}{2}(p\sigma^2 + q\tau^2)}$$

This value is minimized for $p = 1 + \tau/\sigma$, where

$$p\sigma^2 + q\tau^2 = \sigma^2 + 2\tau\sigma + \tau^2 = (\sigma + \tau)^2$$

so $X + Y$ is $(\sigma + \tau)^2$ subgaussian. If X and Y are independant, then we actually have $X + Y$ a $\sigma^2 + \tau^2$ subgaussian variable. We can even make the set of subgaussian random variables into a Banach space, under the norm

$$\sigma(X) = \inf\{\sigma \geqslant 0 : X \text{ is } \sigma^2 \text{ subgaussian}\}$$

By continuity, X is a $\sigma(X)$ subgaussian variable. The main reason for studying subgaussian random variables is that we obtain very good tail bounds for the distribution.

Theorem 3.4. *If X is σ^2-subgaussian, then $\mathbf{P}(X \geqslant x) \leqslant e^{-x^2/2\sigma^2}$.*

Proof. Using Markov's inequality,

$$\begin{aligned}\mathbf{P}(X \geqslant x) &= \mathbf{P}(e^{\lambda X} \geqslant e^{\lambda x}) \\ &\leqslant \mathbf{E}[e^{\lambda X}]e^{-\lambda x} \\ &\leqslant e^{(\lambda^2\sigma^2/2)-\lambda x}\end{aligned}$$

The value of λ which minizes this quantity $\lambda = x/\sigma^2$, which gives us the bound in question. □

The exponential decay of tails is exactly what specifies a subgaussian random variable. To prove this, note that if $\mathbf{P}(X \geqslant x) \leqslant e^{-x^2/2\sigma^2}$ holds, though we do not prove this.

Chapter 4

Existence Theorems

In certain fields of probability theory, we wish to discuss collections of random variables defined over the same sample space. For instance, given a sequence $\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_n$ of probability distributions defined over a space Y, we may want to talk about a sequence of independent random variables $X_i : \Omega \to Y$, such that $\mathbf{P}(X_i \in U) = \mathbf{P}_i(U)$. The construction here is simple; we take $\Omega = Y^n$, let $X_i = \pi_i$ be the projection on the i'th variable, and let $\mathbf{P}$ be the probability measure induced by

$$\mathbf{P}(U_1 \times U_2 \dots \times U_n) = \mathbf{P}_1(U_1)\mathbf{P}_2(U_2)\dots\mathbf{P}_n(U_n)$$

The construction here is simple because we have finitely many distributions, but the problem becomes much harder when we need to talk about an infinite family of distributions $\mathbf{P}_i$, or when we need to talk about non-independent random variables, with some specified relationships between the variables. The problem is to show there exists a sample space Ω 'big enough' for the random variables to all be defined on the space.

Chapter 5

Entropy

Let μ be a probability distribution. We would like to measure the expected 'amount of information' contained in the distribution – in essence, the average information entropy of μ. It was Claude Shannon who found the correct formula to measure this.

Shannon considered the problem of efficient information transfer. Suppose there was a channel of communication between two friends A and B. The friends have agreed on a standard dictionary X of possible messages, along with a probability distribution μ over the dictionary, and we would like to encode these messages into bits, in such a way that the average length of the message is smallest. We then define this to be the information entropy of μ. Shannon showed that if μ is discrete with probabilities $p_1, \dots, p_n$, then the entropy can be calculated as

$$H(\mu) = \sum p_n \log_2 \left(\frac{1}{p_n} \right)$$

where the entropy is measured in bits, we can define the entropy in terms of the natural logarithm, in which case the entropy is said to be measured in nats. We assume that $p_i \log 1/p_i = 0$ for $p_i = 0$, which makes sense by the continuity of $x \log(1/x)$.

The entropy of a distribution also tells us

Now suppose that we were attempting to optimize a message with respect to a discrete distribution μ, and we instead encounter a distribution ν. Then the policy we have used for messages will be less optimal than if we had known that ν was the distribution in the first place. We define the relative difference in information between μ and ν as the difference between the encoding of ν with respect to μ, and the encoding of μ with respect to μ. This is not a linearly ordered relation, ν does not possess more information than μ, just different information. If μ takes probabilities p_i and ν takes relative probabilities q_i, the difference in information is calculated to be

$$D(\mu,\nu) = \sum p_i \log(1/q_i) - \sum p_i \log(1/p_i) = \sum p_i \log(p_i/q_i)$$

This is known as the **Kullback Leibler distance** between μ and ν.

Now suppose we are viewing independent samples $X_1, \dots, X_n$, but we do not know where the samples are drawn from μ or ν. The larger $D(\mu,\nu)$ is, the less time we should take to make an accurate decision that the distribution is μ or ν. Indeed, if $p_i > 0$ and $q_i = 0$, then $D(\mu,\nu) = \infty$, and we can conclude with certainty that the distribution is μ if we ever view the outcome corresponding to p_i.

It is necessary to define the 'entropy' of an arbitrary distribution, but it is then not clear how to interpret the entropy, since an encoding of uncountably many values will always have an infinite expected number of bits. However, we can defined the relative entropy by performing a discretization; Let μ and ν be distributions on some sample space X. Consider function $f : X \to \{1,\dots,n\}$, and define

$$D(\mu,\nu) = \sup_f D(f_*\mu, f_*\nu)$$

where f_* pushes measures on X onto discrete measures on $\{1,\dots,n\}$. For a fixed f, $D(\mu,\nu)$ upper bounds the difference in information we expect to see over a particular discretization. One can then calculate that

$$D(\mu,\nu) = \begin{cases} \infty & : \mu \not\ll \nu \\ \int \log(\frac{d\mu}{d\nu}) d\mu & : \mu \ll \nu \end{cases}$$

The relative entropies of well known distributions are easy to compute. Normal distributions, for instance, have

$$D(N(\mu_1,\sigma^2), N(\mu_2,\sigma^2)) = (\mu_1 - \mu_2)^2/2\sigma^2$$

For Bernoulli distributions, we have

$$D(B(p),B(q)) = p\log(p/q) + (1-p)\log\left(\frac{1-p}{1-q}\right)$$

Which is true except perhaps at boundary conditions.

The Kullback Leibler distance gives us certain bounds which are essential to information theoretic lower bounds. The bound is useful, for it relates the probabilities of distributions by the difference in information contained within.

Theorem 5.1 (The High Probability Pinsker Bound)**.** *If μ and ν are probability measures on the same space X, and $U \subset X$ is measurable, then*

$$\mu(A) + \nu(A^c) \geqslant \frac{1}{2}e^{-D(\mu,\nu)}$$

Suppose we have a decision procedure which attempts to distinguish between events in probability distributions. If we choose an event A upon which the decision procedure fails to make the correct decision on the measure μ, and A^c measures the decision to fail under the measure ν, then the bound above shows the decision procedure cannot work reliably on both μ and ν.

Bibliography

[1] Larry Wasserman, *All of Statistics*

[2] Walter Rudin, *Real and Complex Analysis*