Probability Theory

Jacob Denson

October 16, 2016

Table Of Contents

Chapter 1

Foundations

These notes outline the basics of probability theory, the framework which allows us to communicate that we are 80% more likely to develop lung disease if you are a smoker rather than an average person, or that there is a 50/50 chance of rain today? But what is probability? What do we mean by these probabilistic statements? To a mathematicians, there need not be a rigorous interpretation – probability theory is just a subfield of measure theory. To a natural scientist, 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 the scientific experiment. Suppose you are repeatedly performing some event, in a manner which is controlled well enough that the outcome of the experiment should be the same. Even under rigorously controlled conditions however, the experiment will not always result in the same outcome – we instead have a range of outcomes

which we may observe from a single result in an 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

$$\omega_1, \omega_2, \omega_3, \ldots$$

Let D be a proposition decidable from the outcome of the experiment (for instance D may represent whether a flipped coin lands heads up or heads down in the experiment of flipping a coin). Mathematically, a proposition is just a subset of the set of all outcomes in an experiment – the outcomes for which the proposition is true. We may then define the relative frequency of this proposition being true in n trials to be

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

The key assumption of the frequentist school 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 $0 \leqslant \mathbf{P}(D) \leqslant 1$. If we let Ω denote the set of all possible outcomes to the experiment (a proposition true for all outcomes of the experiment), then

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

Thus $\mathbf{P}(\Omega) = 1$. If $A_1, A_2, \ldots$ 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, $\mathbf{P}(\bigcup_i A_i) = \sum_i \mathbf{P}(A_i)$. This will be true 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. Mathematically, this can also be explained by measure theoretic considerations.

1.2 Bayesian Probability

The frequentist school is sufficient to use probability theory to model scientific experiments, but our own use of probability is much more general. 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 as a frequentist. 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 belief.

One problem with the Bayesian interpretation of probability theory is that there is no way to go out into the world and 'learn probabilities' – it's all based on a person's individual interpretation. The only constraint we have on the choice of probabilities is that they are '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 D a probability $\mathbf{P}(D)$, then you are willing to play the following game: If D does not occur, you lose $\mathbf{P}(D)$ dollars, but if D occurs, you win $1-\mathbf{P}(D)$ 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 odes not occur, so that you think the bets are 'fair' to both sides. 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: 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 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 the probability of the certain event is one, and if $\{A_i\}$ is a countable collection of disjoint events, then $\mathbf{P}(\bigcup_i A_i) = \sum_i \mathbf{P}(A_i)$ (Definetti would have only accepted this statement for finite collections of

events. Here, we allow one to make a countable number of bets at once, rather than only finitely many at any point - Allowing limit operations is too useful to ignore!)

What we have shown is that consistant degrees of belief in the Bayesian system have similar properties of experimental frequencies to a frequentist. Regardless of which philosophy you agree with, you will eventually have to agree on the same fundamental principles of probability theory. Neither of these systems rest of mathematical foundations, so we need to make a rigorous model, from which we can avoid the philosophical controversies that arise. Just as the game of chess does not have to be about knights and castles, the game of probability theory does not have to be about frequencies nor degrees of belief, but can be played from the basic assumptions which define the theory. Note, however, that the interpretations of probabilities have significant effects on how one performs statistical inference.

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 an infinite sample space, and also enables us to rigorously define certain subtle objects in the more advanced theory. A **probability space** is a measure space X with a positive measure μ such that $\mu(X) = 1$. X is known as the **sample space** and μ is known as the **probability distribution** or **probability measure**. We interpret X as the space of outcomes to some random phenomena, and μ measuring the likelyhood of each outcome happening. An arbitrary probability measure μ is often denoted $\mathbf{P}$, since we often only need to talk about one distribution (or the distributions are defined such that we can always determine which measure we are talking about with the notation $\mathbf{P}$), but this is not the only notation!

If X is countable, then a probability measure can be viewed a vector $v \geqslant 0$ in $\mathbf{R}\cdot X$ such that $\sum_{x\in X} v(x) = 1$, because the σ algebra of the measure space plays no real role in the theory. Uncountable examples are much more difficult to construct requiring deep results in measure theory, which we assume to get to the more advanced probability theory.

In probability theory, much more than measure theory, many partic-

ular examples occur across the theory. If we have a certain name F for the example (which we call a distribution), we write $\mu \sim F$ if μ is equal to this distribution. We don't need all the examples below right away, but its good to check that a few are probability distributions for practice.

Example. *If X is finite, then we can define the uniform distribution or normalized counting measure $\mu \sim Uni(X)$ by letting $\mu(A) = |A|/|X|$. Analysis of the probabilities on these spaces thus reduces to combinatorial methods. If $X = \{Heads, Tails\}$, then the uniform distribution on X models the resulting distribution from a fair coin flip.*

Example. *If $x \in X$ is fixed, the point mass distribution $\delta_x \sim PM(x)$ at x is the probability distribution defined by*

$$\delta_x(A) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

The distribution represents an event where a certain outcome is certain.

Example. *The uniform distribution works well in modelling a fair coin, but what if we are flipping an unfair coin? The Bernoulli distribution assigns a probability of p to getting a heads, and $1 - p$ to getting a tails, for some parameter p. We write $\mu \sim Ber(p)$ for this distribution. We normally rename the elements of the sample space by letting $Heads = 1$ and $Tails = 0$, so that the probability distribution counts the number of heads and tails from one coin flip. This permits an easy generalization to the Binomial distribution $Bin(n, p)$, which counts the number of times we get a head in n coin flips from a $Ber(p)$ distribution. It is defined on $\mathbf{Z}$ by letting*

$$\mathbf{P}(m) = \binom{n}{m} p^m (1-p)^{n-m}$$

for $0 \leqslant m \leqslant n$, and $\mathbf{P}(m) = 0$ otherwise. If $\mu \sim Bin(n, p)$, then we have a representation $\mu = \nu * \cdots * \nu$ as an n-ary convolution, where $\nu \sim Ber(p)$. In general, if μ and ν are probability measures on a group G, then $\mu * \nu$ is a probability measure, since

$$(\mu * \nu)(X) = \int_{X^2} d\mu(x)\, d\nu(y) = \int_X d\mu(x) \int_X d\nu(Y) = 1$$

which represents a certain 'sum' over the two distributions, especially in the case $G = \mathbf{Z}$.

Example. *The Geometric distribution on* $\mathbf{N}$ *gives us the probability distribution of the number of flips required to get a heads from a weighted coin. If* $\mu \sim Geo(p)$*, then*

$$\mu(m) = p(1-p)^m$$

The infinite sum over all $m \geqslant 0$ *is easily evaluated to be one, so that the distribution truly is a probability measure.*

Example. *The Poisson distribution on* $\mathbf{N}$ *is defined, for* $\mu \sim Poisson(\lambda)$*, by letting*

$$\mu(n) = \lambda^n \frac{e^{-\lambda}}{n!}$$

This distribution is usually used to count rare, independant events, like for radioactive decay and traffic accidents. If $\mu \sim Poisson(\lambda)$ *and* $\nu \sim Poisson(\gamma)$*, then* $\mu * \nu \sim Poisson(\lambda + \gamma)$*.*

To obtain measures on uncountable sample spaces, we rely on integration theory.

Example. *If* $f \geqslant 0$ *is a measurable function on a measure space* X *with measure* μ *with* $\|f\|_1 = 1$*, then we may define a probability measure* ν *on* X *by letting*

$$\nu(A) = \int_A f \, d\mu$$

The measure ν *is often denoted* $f \, d\mu$ *because of the integral formula we use to obtain it.*

The method in the last example gives us a number of canonical distributions, giving us a wide number of examples. A theorem of measure theory (the Radon-Nikodym theorem) tells us why almost all the important probability distributions can be given by integration (or summation, those measure theory tells us these are essentially the same processes).

Example. *We would like to have a probability distribution which models picking a point 'uniformly' from an interval* $[a, b]$*. What we really mean is that the probability measure is translation invariant, in the sense that* $\mathbf{P}(A + t) = \mathbf{P}(A)$*, for all* A *and* t *for which this should be reasonably defined. What we are looking for is exactly the normalized Lebesgue measure restricted to* $[0, 1]$*, which is rigorously defined in measure theory. We take the lebesgue measure* dx*, and then normalize to obtain a probability measure* $dx/(b-a) \sim Uni(a, b)$*. It seem*

paradoxical that the probability that any particular point is chosen is zero, yet in an experiment we must choose a particular point. However, to obtain probabilities we take the limit of ratios over an infinite number of experiments, and it is entirely possible for the limit of the ratios to converge to zero, even if a point is chosen in some experiment.

Example. *Define a function* $f : \mathbf{R} \to \mathbf{R}$ *by the formula*

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

the integral of this function over $\mathbf{R}$ *is one, and is proved in multivariate calculus courses. The probability measure generated by* f *is known as the* **normal distribution** *with mean* μ *and standard deviation* σ*. We write* $\mathbf{P} \sim N(\mu, \sigma^2)$ *if* $\mathbf{P}$ *for such a measure.*

Example. *The exponential distribution* $Exp(\beta)$ *is defined on the positive real numbers by the measure induced by the lebesgue measure, and the function*

$$f(x) = \frac{e^{-x/\beta}}{\beta}$$

It measures the waiting times between rare events that occur at continuous times.

Example. *The Gamma distribution is defined with respect to the remarkable function known as the Gamma function*

$$\Gamma(x) = \int_0^\infty y^{x-1} e^{-y} dy$$

The Gamma distribution $\Gamma(\alpha, \beta)$ *with positive parameters* α *and* β *is induced from the Lebesgue measure by the function*

$$f(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}$$

We have $Exp(\beta) = \Gamma(1, \beta)$*.*

Example. *The Beta distribution* $Beta(\alpha, \beta)$ *for positive* $\alpha, \beta > 0$ *is induced by the function*

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

It is a continuous version of the Binomial distribution.

Example. *The t distribution with ν degrees of freedom, denoted t_ν, if it is the measure taken by*

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{(1+x^2/\nu)^{(\nu+1)/2}}$$

It is similar to the normal distribution, but has thicker tails. In fact, as $\nu \to \infty$, the distributions corresponding to t_ν look more and more like the normal distribution. The Cauchy distribution is the t distribution with $\nu = 1$, defined by

$$f(x) = \frac{1}{\pi(1+x^2)}$$

Example. *The χ^2 distribution with p degrees of freedom, denoted χ^2_p, has distribution induced on the positive real numbers by*

$$f(x) = \frac{x^{(p/2)-1} e^{-x/2}}{\Gamma(p/2) 2^{p/2}}$$

if $\mu \sim N(0,1)$, and ν is a measure defined on the positive real numbers by letting $\nu(A) = \mu(B)$, where $B = \{x \in \mathbf{R} : x^2 \in A\}$, then the p-adic convolution $\nu * \cdots * \nu \sim \chi^2_p$.

We have seen a whole smorgasbord of probability measures, which easily suffices to motivate the general theory. We will see these measures arise naturally in the right context, but it is also useful to have these measures to use for intuition in the theory.

The first immediately obvious fact from the axioms of a probability space X is that $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$, since A and A^c are disjoint events whose union is the whole space X. Another useful fact is that if a sequence of events A_i converges to A, in the sense that $\limsup A_i = \liminf A_i = A$, where

$$\limsup A_i = \bigcap_{i=1}^{\infty} \bigcup_{j \geqslant i} A_j \qquad \liminf A_i = \bigcup_{i=1}^{\infty} \bigcap_{j \geqslant i} A_j$$

then $\mathbf{P}(A_i) \to \mathbf{P}(A)$. In particular, this holds if the A_i are monotone increasing or decreasing. As in all measure spaces where the measures of the sets are finite

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$$

Which is a simple extension of the disjoint sum of probabilities over the space.

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 B holds over the sample space, we obtain a new probability distribution over X, denoted $\mathbf{P}(D|B)$, which represents the ratio of winnings from the bet which is only played out if B occurs. That is

- You win $1 - \mathbf{P}(D|B)$ dollars if D occurs, and B occurs.
- You lose $\mathbf{P}(D|B)$ dollars if D does not occur, and B occurs.
- You do not lose or win money if B does not occur.

It then follows by a dutch bet argument that

$$\mathbf{P}(B)\,\mathbf{P}(D|B) = \mathbf{P}(B \cap D)$$

Suppose instead that $\mathbf{P}(B)\,\mathbf{P}(D|B) < \mathbf{P}(B \cap D)$. Bet on B occuring, and also bet against $B \cap D$ occuring. Then

- If B does not occur, we gain $\mathbf{P}(B \cap D)$ dollars.
- If $B \cap D$ occurs, we lose $1 - \mathbf{P}(B \cap D)$ dollars, and gain $1 - \mathbf{P}(B)\,\mathbf{P}(D|B)$ dollars.
- If $B \cap D^c$ occurs, we gain $\mathbf{P}(B \cap D)$ dollars, and lose $\mathbf{P}(B)\,\mathbf{P}(D\ B)$ dollars.

The inequality guarantees that we always make a profit on these bets. Similarily results happen if we assume the opposite inequality, so we must have equality.

In the empirical interpretation, $\mathbf{P}(D|B)$ is the ratio of times that D is true in experiments, where B also occurs. That is, we define $\mathbf{P}(D|B)$ as the limit of the ratios

$$\frac{\#\{k \leqslant n : \omega_k \in B\}}{n} \frac{\#\{k \leqslant n : \omega_k \in D, \omega_k \in B\}}{\#\{k \leqslant n : \omega_k \in B\}} = \frac{\#\{k \leqslant n : \omega_k \in D, \omega_k \in B\}}{n}$$

which gives us the formula $\mathbf{P}(B)\,\mathbf{P}(D|B) = \mathbf{P}(D \cap B)$. We must of course assume that $\mathbf{P}(B) \geqslant 0$, since overwise we are almost certain that B will

never occur, and we can therefore define $\mathbf{P}(D|B)$ arbitrarily (or not define it at all).

Thus we have motivation to define conditional probabilities by the formula $\mathbf{P}(B)\mathbf{P}(D|B) = \mathbf{P}(D \cap B)$. 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 sets which do not give us any information about a particular event. We say two events D and B are independant, denoted $D \amalg B$, if $\mathbf{P}(D|B) = \mathbf{P}(D)$, or equivalently, if $\mathbf{P}(D \cap B) = \mathbf{P}(D)\mathbf{P}(B)$; knowledge of B gives us no foothold over knowledge of the likelihood of D.

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 (brand new) car is placed uniformly randomly behind one of the doors. After we pick a door, the gameshow host then randomly opens one of the other doors which you didn't pick, revealing the car isn't behind the door (his intention). What is the chance that the door you picked has the brand new car? It we pick door i, then certainly*

$$\mathbf{P}(i \textit{ has a car}) = 1/3$$

Yet this implies, by symmetry, that for any $j \in \{2,3\}$,

$$\begin{aligned} 1/3 &= \sum_{j=1}^{3} \mathbf{P}(i \textit{ has a car}|\textit{door } j \textit{ is opened})\mathbf{P}(\textit{door } j \textit{ is opened}) \\ &= 2(1/6 + 1/3)\mathbf{P}(i \textit{ has a car}|\textit{door } j \textit{ is opened}) \end{aligned}$$

Dividing out, the probability that our door is correct is $1/3$, so we should definitely switch to obtain our best chance of success.

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 sample space of the problem can be described by the tuple (j,k), where j is where the car is, and k is a door we open. We can explicitly enumerate the sample space as

$$(1,2),(1,3),(2,3),(3,2)$$

and the car is seen to be in our door half of the time.

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 $A_1, \ldots, A_n$, 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). Note that Bayes rule applies in the probability theory of each rule of thought, but can be used by Bayesians in a much more applicable way to their statistical analysis.

Chapter 2

Random Variables

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

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

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 $\mu > 0$

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

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*

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

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 $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n$ 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

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

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 X and Y be random variables.

Bibliography

[1] Larry Wasserman, *All of Statistics*

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