NICHOLAS HOELL

DATA SCIENCE

MMF1922 COURSE NOTES

UNIVERSITY OF TORONTO

PUBLISHED BY UNIVERSITY OF TORONTO

http://www.math.toronto.edu/nhoell/MMF1922

First printing, October 2018

Contents

Probability Theory: An Overview 5

Exercises 41

Estimation and Inference 43

Exercises 53

Bibliography 55

Probability Theory: An Overview

Contents

A Cautionary Tale 6

Probability Spaces 7

Random Variables 10

Distributions and Densities 12

Useful Distributions 14

Random Vectors 17

Independence 18

Conditional Probability 19

Some Statistics 22

Entropy and Information 26

Relative Entropy 29

Some Inequalities 31

Conditional Expectation 33

Characteristic Functions 35

Some Limit Laws 37

Exercises **41**

This chapter provides an accelerated overview of fundamental concepts in the theory of probability. It is of a highly theoretical nature. If this is your first exposure to advanced probability theory, you may find it quite challenging. Excellent references are [1] and [2]

[1] William Feller. *An Introduction to Probability Theory and its Applications*, volume 1. Wiley, 1968

[2] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 4 edition, 2010

A Cautionary Tale

To motivate the necessity for caution and clarity regarding things dealing with randomness, we consider a cautionary tale in the form of the following question.

Example 1

QUESTION: Consider a circle in the plane with a 2 inch radius. Select a chord of this circle at random. What's the probability that this randomly selected chord will hit an inner circle whose radius is 1 inch? I'm going to give three separate ways of thinking of this

problem.
Reasonable Answer #1: Chords are uniquely (modulo those passing through the centre...) determined by the location of heir midpoints. Thus,

$$\begin{aligned}\mathbb{P}(\text{hitting inner circle}) &= \frac{\text{area of inner circle}}{\text{area of outer circle}} \\ &= \frac{1}{4}\end{aligned}$$

In the above, I've used the standard convention of denoting a probability with the symbol $\mathbb{P}$.

Reasonable Answer #2: Circles have an obvious symmetry. We may as well then simply consider vertically aligned chords. The diameter of the outer circle is 4 inches and the chord will intersect the inner circle if and only if the chord falls in the 2 inch diameter region. From this then we have

$$\begin{aligned}\mathbb{P}(\text{hitting inner circle}) &= \frac{2inches}{4inches} \\ &= \frac{1}{2}\end{aligned}$$

Reasonable Answer #3: We use symmetry again. Assume that the chord intersects the left edge of the larger circle subtending an angle $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ with the horizontal. By trigonometry we have

$$\begin{aligned}\mathbb{P}(\text{hitting inner circle}) &= \mathbb{P}(\theta \in [-\frac{\pi}{6}, \frac{\pi}{6}]) \\ &= \frac{\frac{2\pi}{6}}{\pi} \\ &= \frac{1}{3}\end{aligned}$$

Our naive intuition led, in the prior example, to three *completely different* answers to what sounded like an innocuous geometry question. Namely, we obtained

$$\mathbb{P} = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}$$

all with reasonable-sounding arguments. The reason for this mess lies in the carelessness with which we used the word “random" in the posing of the question. We never really made clear what we meant by ‘selecting a chord of this circle at random". Each such occurrence of a chord in the example is a **random event** and each of the three suggested answers is a different attempt to capture the random event in a **random variable**. We’re going to try to tighten up our understanding of randomness in order to have unambiguous answers to questions like the above.

Probability Spaces

Probability theory is a mathematical tool allowing us to draw insight from experiements. The results of experiments are **events**. Of course, a single experiment may result in several events. For instance, throwing two die can consist in the following events

- 3 & 3
- two odd faces, or
- sum of tosses = 6

none of which are mutually exclusive outcomes. The above are examples of what are called **compound events** as they decompose neatly into single events. For instance the “sum of tosses = 6" event decomposes into the single events $(1,5)$, $(2,4)$, $(3,3)$, $4,2)$ and $(5,1)$. To take another easy example, consider temperature represented in the variable x. Then each value of x represents a simple event whereas “ the temperature is in the fifties" is represented by the compound event

$$50 \le x < 60$$

When dealing with probability, we make the following definition.

DEFINITION 1: SAMPLE SPACE

The **Sample Space**, Ω, of an experiment is the set of all discoverable outcomes. Namely $\Omega \quad = \quad \{\text{all events}\}$. The elements of Ω are often called **sample points**.

EXAMPLE 2

We toss a fair coin 3 times. Then

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

And the event A = “ two or more heads" corresponds to the first four elements in Ω, the event $B =$ “just one tail" corresponds to the events HHT, THH, HTH and $A \cap B$ is also an event in Ω, namely, $A \cap B = \{HHT, THH\}$

Notice that every thinkable outcome of the experiment in question corresponds to one and only one event. It’s important to emphasize a sort of obvious point: *it only makes sense to discuss an event A if it’s clear for every outcome of the experiment in question whether or not A has occurred*. We then ought to be able to determine whether A or A^c has occurred. Further more, we should be able to know, given any two elements of Ω, whether $A \cup B$ or $A \cap B$ has occurred. These naive considerations lead us to the more sophisticated definition of a probability space.

Definition 2: Probability Space

A Probability Space is a tuple $(\Omega, \mathcal{F}, \mathbb{P})$ where

- Ω is a sample space of events under consideration
- $\mathcal{F}$ is a σ**-algebra** of subsets of Ω, and
- $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ is a **probability measure** on the above σ-algebra

A σ-algebra $\mathcal{F}$ of Ω is a collection of subsets of Ω such that

- $\Omega \in \mathcal{F}$
- $A \in \mathcal{F} \implies A^c \in \mathcal{F}$, and
- if $A_1, A_2, ... \in \mathcal{F}$ then $\cup_i A_i \in \mathcal{F}$

The last condition above is generally phrased “closure under countable unions". The probability measure $\mathbb{P}$ is a function from $\mathcal{F}$ to the unit interval $[0, 1]$ satisfying:

- $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$
- $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}(U_i A_i) = \sum_i \mathbb{P}(A_i)$ for all A_i satisfying $A_i \cap A_j = \emptyset, i \neq j$

The concept of σ-algebra may seem overly-technical and mysterious to you right now, but that should be overcome with familiarity. These σ-algebras will become a more essential part of your studies in courses on martingales. See, e.g., the excellent little book by Williams [3]. Mostly σ-algebras should be thought of as “discernable information", as it is only elements of these which we can sensibly assign probabilities in a consistent way to. They are divisions of sample space, namely a list of outcomes, for which we can determine whether or not an event has occurred.

[3] David Williams. *Probability with Martingales*. Cambridge University Press

Notice, as a consequence of the above, that

$$\begin{aligned} 1 &= \mathbb{P}(\Omega) \\ &= \mathbb{P}(A \cup A^c) \\ &= \mathbb{P}(A) + \mathbb{P}(A^c) \end{aligned}$$

so we have $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ gives the probability of the complementary outcome to event A. As well, if $A \subset B$, then $B = A \cup (A^c \cap B)$ provides a disjoint union decomposition of B. Then

$$\begin{aligned} \mathbb{P}(B) &= \mathbb{P}(A \cup (A^c \cap B)) \\ &= \mathbb{P}(A) + \mathbb{P}(A^c \cap B) \\ &\geq \mathbb{P}(A) \end{aligned}$$

So that $A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$. In this sense the function $\mathbb{P}$ helps to gives sizes to the sets of $\mathcal{F}$. We can extend the preceding to infinite collections of sets.

Proposition 1

The function $\mathbb{P}$ is countably subadditive. In other words,

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

for all $A_i \in \mathcal{F}$

Proof

The sets $\tilde{A}_i = A_i \setminus (\cap_{j<i} A_j)$ are disjoint. Then, $\mathbb{P}(\cup_i A_i) = \mathbb{P}(\cup_i \tilde{A}_i) = \sum_i \mathbb{P}(\tilde{A}_i)$. But, since $\tilde{A}_i \subset A_i$ we have $\sum_i \mathbb{P}(\tilde{A}_i) \leq \sum_i \mathbb{P}(A_i)$. □

We close with an important common case of probability spaces.

Example 3

There are **discrete probability spaces**. These occur when Ω is an at most countable set, $\mathcal{F} = \mathcal{P}(\Omega)$, the powerset of the sample space. Then we define

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega)$$

where $\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1$. For Ω a finite set, we often use $\mathbb{P}(\omega) = \frac{1}{|\Omega|}$, distributing probability equally among events.

Example 4

Consider rolling two dice. Then this is a discrete probability space with $\Omega = \{1, 2, 3, 4, 5, 6\}^2$ and σ-algebra given by all possible subsets of Ω. Then, as in the above, we define $\mathbb{P}(A) = \frac{|A|}{36}$.

Random Variables

We want to be able to deal with data (numbers) coming from a random event. So the numbers will be encoding randomness for us. Given the numbers which encode the randomness, we still want to be able to distinguish events. Consider the following example.

Example 5

We consider a coin toss. In this case $\Omega = \{\omega \mid \omega \in \{H, T\}\}$. We will use $\mathcal{F} = \mathcal{P}(\Omega)$ the powerset of the sample space, as our σ-algebra. We can define a map $X : \mathcal{F} \to \{0, 1\}$ via

$$X(H) = 0, \quad X(T) = 1$$

Then observing the values of X allows us to determine whether a heads or tails occurred.

On our probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we have probabilities of events. As the previous example sort of indicated, random variables are a way for us to be able to assign probabilities to numerical data. There are, however, some technical obstructions precluding a naive assignment of probabilities to, say, $\mathbb{R}$. We deal with these now.

For a topological space S, we define $\mathcal{B}(S)$ to be the so-called “Borel field" of S given by $\sigma(\text{open sets of } S)$, the σ-algebra generated by the open sets of S. Namely, it’s a σ-algebra adapted to the topological structure of S and is the *smallest* σ-algebra containing the open sets of S (in the sense that all others with that property have $\mathcal{B}(S)$ as a subset). We often will just use $\mathcal{B} \doteq \mathcal{B}(\mathbb{R})$. Even without invoking subtle mathematics like the Axiom of Choice, one can show that $\mathcal{B} \neq \mathbb{R}$ since $\mathbb{R}$ can contain some very complicated sets not living in $\mathcal{B}$.

One common construction of $\mathcal{B}$ involves the following. We set $\pi(\mathbb{R}) = \{(-\infty, x] \mid x \in \mathbb{R}\}$. Then we consider the σ-algebra generated by this collection of subsets of $\mathbb{R}$. It turns out that

$$\mathcal{B} = \sigma(\pi(\mathbb{R}))$$

It’s this construction of $\mathcal{B}$ which is most helpful to us in our consideration of random variables.

Definition 3: Random Variable

A **random variable** $X : \mathcal{F} \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function such that

$$X^{-1}(B) \in \mathcal{F}$$

holds for all Borel sets $B \subset \mathcal{B}$. Such an X as above is said to be a “$\mathcal{F}$-measurable" function.

In the above, the condition $X^{-1}(B) \in \mathcal{F}$ is a common shorthand for

$\{\omega \mid X(\omega) \in B\} \in \mathcal{F}$. What this means is that the preimage of Borel subsets should be distinguishable, namely, things we can assign probabilities to. After all, we're using the function X as a way to measure probabilities on spaces like $\mathbb{R}$ which are inheriting their randomness from the probability space. So we want to be able to assign probabilities to the interesting subsets of $\mathbb{R}$.

Example 6

Consider the trivial algebra $\mathcal{F} = \{\emptyset, \Omega\}$. We can only construct discrete random variables on this space. Namely

$$\{\omega \mid X(\omega) = c\} \iff \omega = \Omega$$

In other words, X must be constant. The only random variables we can construct on the trivial algebra are constants.

Important: In the preceding example the sample space Ω need not be discrete, only the random variable must be. Random variables say more about the σ-algebras than about the underlying sample space.

Example 7

Consider $\mathcal{F} = \mathcal{P}(\Omega)$. In this case *any* function $X : \mathcal{F} \to \mathbb{R}$ is a random variable since $\{\omega \mid X(\omega) = c\}$ is clearly a subset of Ω and therefore is in $\mathcal{F}$.

The next example strikes a balance in between the two preceding ones.

Example 8

Consider the measurement of a stock on two different days. The price can move up or down, which we represent as tuples. Namely,

$$\Omega = \{\omega_1 = (u,u).\omega_2 = (u,d), \omega_3 = (d,u), \omega_4 = (d,d)\}$$

Now we construct a σ-algebra. Let $A = \{\omega_1, \omega_2\}$ and define

$$\mathcal{F} = \{A, A^c, \emptyset, \Omega\}$$

namely, $\mathcal{F} = \sigma(A)$, the σ-algebra generated by the set A.
Now consider two functions X, Y defined on Ω via

$$X(\omega_1) = X(\omega_2) = 1.5, \quad X(\omega_3) = X(\omega_4) = .5$$

and

$$Y(w_1) = 1.5^2, Y(\omega_2) = Y(\omega_3) = .75, Y(\omega_4) = .5^2$$

Notice that

$$\{\omega \mid X(\omega) = 1.5\} = \{X = 1.5\} = \{\omega_1, \omega_2\} = A \in \mathcal{F}$$

and

$$\{X = .5\} = A^c \in \mathcal{F}$$

so therefore X is $\mathcal{F}$-measurable and necessarily defines a random variable. On the other hand, $\{Y = .75\} = \{\omega_2, \omega_3\} \notin \mathcal{F}$ so Y is *not* a random variable.

Distributions and Densities

Random variables implicitly define **distribution functions**

$$F_X(x) \doteq \mathbb{P}(\{\omega \mid X(\omega) \leq x\})$$

Sometimes also written $F_X(x) = \mathbb{P}(X \leq x)$. Be sure you understand the difference and are ok with either. The distribution function is also called the **cumulative distribution function** and abbreviated cdf.

Of course, random variables then also induce probability measures (also called a **probability distribution** in this setting) on $\mathbb{R}$, via

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(\{\omega \mid X(\omega) \in B\}), \quad \forall B \in \mathcal{B}$$

The above makes $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ into a probability space in its own right. These considerations are encapsulated in the following representation

$$(\Omega, \mathcal{F}, \mathbb{P}) \overset{X}{\underset{X^{-1}}{\longleftrightarrow}} (\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$$

Random variables are determined by the values they take on, and this is governed by their distribution functions. So, to talk about a random variable *is* to talk about a distribution function.

Proposition 2

Let F be a distribution function for a random variable X. Then

1. F is non-decreasing
2. $\lim_{x\uparrow\infty} F(x) = 1$ and $\lim_{x\downarrow-\infty} F(x) = 0$
3. $\lim_{x\downarrow y} F(x) = F(y)$, F is right-continuous
4. $\lim_{x\uparrow y} F(x) = \mathbb{P}(X < y)$ and
5. $\mathbb{P}(X = x) = F(x) - F(x-)$

Proof

To prove the first claim we notice that $x \leq y \implies \{X \leq x\} \subset \{X \leq y\}$. So therefore $\mathbb{P}(X \leq x) \leq \mathbb{P}(X \leq y)$ proving the non-decrease of the cdf.

The second claim follows trivially from

$$\lim_{x\uparrow\infty}\{X \leq x\} = \Omega, \qquad \lim_{x\downarrow-\infty}\{X \leq x\} = \emptyset$$

For the right-continuity of the cdf observe that $y \downarrow x \implies \{X \leq y\} \downarrow \{X \leq x\}$. For the fourth claim notice that $y \uparrow x \implies \{X \leq y\} \uparrow \{X < x\}$. Notice the strict inequality here. Finally, $\mathbb{P}(X = x) = \mathbb{P}(X \leq x) - \mathbb{P}(X < x) = F(x) - F(x-)$. □

As it happens, the first three conditions given in the previous proposition rigidly determine distribution functions in the sense that if a function satisfies the first three conditions then it happens to be a cdf for some random variable.[4].

Proposition 3

Suppose that $F(x)$ satisfies

1. F is non-decreasing
2. $\lim_{x\uparrow\infty} F(x) = 1$ and $\lim_{x\downarrow-\infty} F(x) = 0$
3. $\lim_{x\downarrow y} F(x) = F(y)$, F is right-continuous

Then $F(x) = F_X(x)$ for some random variable X.

Proof

Let $\Omega = (0,1)$, $\mathcal{F} = (B)$ and $\mathbb{P}$ denote the Lebesgue measure on the unit interval. Define

$$X(\omega) = \sup\{y \mid F(y) < \omega\}$$

We show that X is a random variable, namely that it is $\mathcal{F}$-measurable. For this, notice that

$$\omega \leq F(x) \implies X(\omega) \leq x$$

If $w > F(x)$ the from the right-continuity of F we have that there exists an $\epsilon > 0$ such that $F(x+\epsilon) < \omega$ and $X(\omega) \geq x + \epsilon > x$. Therefore

$$w \leq F(x) \iff X(\omega) \leq x$$

So $\{\omega \mid X(\omega) \leq x\} = \{w \mid \omega \leq F(x)\}$. Thus

$$\begin{aligned}\mathbb{P}(X \leq x) &= \mathbb{P}(\omega \leq F(x) \\ &= \mathbb{P}([0, F(x)]) \quad \text{in Lebesgue measure} \\ &= F(x)\end{aligned}$$

□

[4] The following proof uses Lebesgue measure. Lebesgue measure is a standard measure on intervals of $\mathbb{R}$ given by $\mathbb{P}([a,b]) = b - a$ on closed intervals and vanishing on individual points.

A commonly used tool in probability, related to cdf, is that of a density function which we now introduce.

Definition 4: Probability Mass and Density

- Consider the induced distribution $\mathbb{P}_X$ of random variable X. If there exists an at most countable subset $B = \{x_1, x_2, ...\} \subset \mathbb{R}$ such that $\mathbb{P}(B) = 1$ then X is a **discrete random variable**. In this case we define
$$p(x) \doteq \mathbb{P}(X = x)$$
to be the **probability mass function** (pmf) of X.

- If X is a **continuous random variable** with differentiable distribution function $F_X(x)$, set $p_X(x) = F'_X(x)$. We call $p_X(x)$ the **probability density function** (pdf) of X.

A few remarks are in order.

- *Par abus de langage* we shall normally just refer to pmfs and pdfs as, simply pdfs, or even more simply, “densities".

- Sometimes we omit writing the random variable X in the density in question and abbreviate $p_X(x)$ as simply $p(x)$. The random variable for which the function is a density should be clear from context.

- We should think or densities, pmfs particularly so, as
$$p(x_i) = \lim_{n \nearrow \infty} \frac{\#\{\text{times } x_i \text{ occurred in } n \text{ trials}\}}{n}$$
Namely, it’s the relative frequency or *idealized histogram* of outcomes from experimental investigation. The above shouldn’t be taken too seriously in the case of continuous distributions but is a good intuition to nurture.

- Given a pdf, $p(x)$, for a continuous variable X, we can then define
$$\mathbb{P}(X \in (a, b)) = \int_a^b p(x)dx$$
This useful formula gives us the probability of observing values of our random variable on any interval of the real line.

- We often write $X \sim p(x)$ to indicate that the random variable X has density given by the function p, namely that X is “p distributed".

Useful Distributions

Here we cover some of the commonly encountered distributions relevant for our work.

1. **Bernoulli Distribution**. Let $X \in \{0, 1\}$ with $\mathbb{P}(X = 1) = p$. Then

$$p_X(x) = p^x(1-p)^{1-x}$$

represents the "probability of success" in a so-called "Bernoulli trial". We denote this parameterized distribution by $B(p)$ and write $X \sim B(p)$.

2. **Binomial Distribution:** Let X denote the number of successes in n Bernoulli trials where the probability of a single success is p. You can think of this as being the number of total heads in n tosses of a coin whose probability of getting a single head is p. Then $X = \sum_{i=1}^{n} Y_i$ for $Y_i \sim B(p)$. Then we have

$$\begin{aligned}\mathbb{P}(X = x) &= \mathbb{P}(x\text{successes out of } n \text{ trials})\\ &= \mathbb{P}(x \text{ of } n \text{ of the } Y\text{'s are 1 with probability } p)\\ &= \binom{n}{x} p^x(1-p)^{n-x}\end{aligned}$$

We say that $X \sim B(n, p)$ and say that X is binomially distributed. Notice that $B(p) = B(1, p)$.

3. **Poisson Distribution:** We say that X is Poisson-distributed (with parameter λ) if

$$p_X(k, \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k = 0, 1, 2, ...$$

This represents the probability of observing k events within a fixed-time interval.

The Poisson distribution is related to the Binomial distribution. Here we sketch the relationship. Consider p to be the (smallish) chance of success in a Bernoulli trial and suppose that $n >> 1$. Then define $\lambda = np$ which should be medium-sized, based on the assumptions on n and p. Then, as we saw before, $B(n, p)(k) = \binom{n}{k} p^k(1-p)^{n-k}$. So, in particular, we have

$$B(n, p)(0) = (1-p)^n = (1 - \frac{\lambda}{n})^n \xrightarrow{n\to\infty} p_{poisson}(0; \lambda) = e^{-\lambda}$$

Next,

$$\begin{aligned}B(n, p)(1) &= np(1-p)^{n-1}\\ &= \frac{\lambda}{1-p} B(n, p)(0)\\ &= \frac{\lambda}{1-\frac{\lambda}{n}} B(n, p)(0)\\ &\xrightarrow{n\to\infty} \lambda e^{-\lambda}\\ &= p_{poisson}(1; \lambda)\end{aligned}$$

Similarly,

$$\begin{aligned}
B(n,p)(2) &= \frac{n(n-1)p^2}{2(1-p)^2} B(n,p)(0) \\
&\xrightarrow{n\to\infty} \frac{\lambda^2}{2} e^{-\lambda} \\
&= p_{poisson}(2;\lambda)
\end{aligned}$$

From which we see the trend that

$$p_{poisson}(k;\lambda) \approx B(n, \frac{\lambda}{n})(k), \qquad n \text{ large}$$

Though we see this by working through formulas the reasoning is perhaps mysterious. We'll try to clear it up a bit.

Let's suppose that we have an experiment whose conditions remain constant in time (if you like, you can imagine the experiment being clicks on a geiger counter while observing a radioactive substance). Split the unit time of the experiment into n intervals of time each of length $\frac{1}{n}$. We suppose that non-overlapping time intervals have the property that the number of events occurring in one interval reveals nothing about the number of events occurring in any other interval. We will count as a success whether a particular subinterval has *at least* one event take place in it. We define a density function indexed by the number of subintervals we've got under consideration

$$p_n = \mathbb{P}(\text{success})$$

which, necessarily, is the same for each subinterval. The probability of observing k successes over the entire duration of the experiment ought to be $B(n,p_n)(k)$ since we've got a Bernoulli trial for each subinterval. Of course, as n gets really big we expect a vanishing probability of a success on any subinterval.

Next, we split each interval in half. Notice that this gives us

$$p_n = 2p_{2n} - p_{2n}^2$$

since an event happening in a window of length $\frac{1}{n}$ should be twice as likely as an event happening in a window of half the length, except sometimes there will be events that happen only in the first (or second) half of the now-split subinterval. So the subtraction allows for us to not mistakenly overcount those multiple events that happen on one half of an interval and aren't evenly spread out. The above then implies that $p_n < 2p_{2n}$, so in particular we have that $np_n < 2np_{2n}$ which means that np_n is monotonically increasing (but bounded above). We define $\lambda \doteq \lim_{n\nearrow\infty} np_n$. Then we have $B(n, \frac{\lambda}{n})(k) \xrightarrow{n\uparrow\infty} p_{poisson}(k;\lambda)$ as previously claimed.

As a follow-up on the above, we remark that $e^{-\lambda t}\frac{(\lambda t)^k}{k!}$ is the probability of finding k successes in an interval of length t. So we can scale the length of the Poisson distributed variables accordingly.

4. **Normal Distribution:** X is said to be normally distributed, of Gaussian distributed, if

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \mathcal{N}(\mu, \sigma^2)(x)$$

with $\sigma \neq 0$. This distribution, for reasons to be made clear later, is ubiquitous in probability, statistics, physics, mathematics, and data science. In some sense it is the canonical distribution for a random variable. We remark that Gaussian distributed random variables exhibit a nice scaling property whereby if

$$X \sim \mathcal{N}(\mu, \sigma^2) \implies \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1) \tag{0.0.1}$$

We call $\mathcal{N}(0,1)$ - distributed random variables “Standard Normal" variables.

5. **Cauchy Distribution:** X is said to be standard Cauchy distributed if

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

More generally the Cauchy Distribution function is $Ca(x_0,\gamma)(x) = \frac{1}{\pi(1+(\frac{x-x_0}{\gamma})^2)}$.

6. **Logistic Distribution:** We say that $X \sim LG(\mu, s)$ is logistically distributed when $LG(\mu, s) = \frac{e^{-\frac{x-\mu}{s}}}{s(1+e^{-\frac{x-\mu}{s}})^2}$ The Logistic distribution can be defined in terms of the **logistic function** $\sigma(z) \doteq \frac{1}{1+e^{-z}}$ which we will encounter later on when doing regression and discussing machine learning.

7. **Laplace Distribution:** X is Laplace distributed if $X \sim Lap(\mu, b) = \frac{1}{2b}e^{\frac{|x-b|}{b}}$ for $b > 0$. This is sometimes called a double exponential distribution.

8. **Exponential Distribution:** X is exponentially distributed if $X \sim Exp(a,b)$. The function $Exp(a,b) = \frac{1}{a}e^{-\frac{x-b}{a}}\mathbb{1}_{x>b}$. Here we’ve used the indicator function, $\mathbb{1}_A(x)$ which takes the value 1 on $x \in A$ and zero otherwise.

9. **Uniform Distribution:** We say that X is uniformly distributed on $[a,b]$ when $X \sim \mathcal{U}(a,b)$ with $\mathcal{U}(a,b)(x) = \frac{\mathbb{1}_{[a,b]}}{b-a}$.

Random Vectors

We begin with a definition.

Definition 5: Random Vector

Let X_i, $i = 1, ..., n$ be random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$ is a **random vector**. It has **joint distribution function** given by

$$F_{X_1,...,X_n}(x_1, ..., x_n) = \mathbb{P}(X_1 \leq x_1, ..., X_n \leq x_n)$$

The main takeaway from the definition of a random vector is that it doesn't make sense, a priori, to describe components by themselves. One should think of the entire vector as a random object. On the other hand, one way to think about the way the components of a random vector distribute themselves is via **marginal distributions**. These are defined as follows. If $\{i_1, ..., i_k\} \subset \{1, 2, ..., n\}$ with $i_1 < i_2 < \cdots < i_k$ then

$$F(x_{i_1}, ..., x_{i_k}) \doteq F_{X_1,...,X_n}(\infty, \infty, ..., x_{i_1}, \infty, ..., x_{i_2}, \infty, ..., \infty)$$

is the **marginal distribution function** of $(X_{i_1}, ..., X_{i_k})^T$. Probability density functions and probability mass functions are defined analogously to how they were in the single variable case. Namely, if $F_{X_1,...,X_n}$ is differentiable then

$$F_{X_1,...,X_n}(x_1, ..., x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} \underbrace{p(x_1, ..., x_n)}_{\text{Joint pdf of } X_1,...,X_n} \; dx_1 \cdots dx_n$$

Independence

Two events, A and B, are said to be **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. You can think of this as similar to the informal notion of independence meaning “not influenced by". This is because if the two events have nothing really to do with each other then calculating the chances that both happen is simply doing combinatorics on the chances of each individually happening. For each of the multitude of ways A could happen, we'd have a multitude of ways for B to happen and therefore we'd be multiplying the ways (probabilities). We can, of course, generalize this notion to multiple events. Namely, A_i, for $i = 1, ..., n$ are independent events if and only if

$$\mathbb{P}(\cap_i A_i) = \Pi_i \mathbb{P}(A_i)$$

We can extend the notion of independent events to the more powerful notion of **independent random variables**.

DEFINITION 6: INDEPENDENT RANDOM VARIABLES

Random variables $X_1, ..., X_n$ with joint distribution function $F_{X_1,...,X_n}$ are **independent** if

$$F_{X_1,...,X_n}(x_1, ..., x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

i.e. the joint distribution factors into a product of marginals.

We denote that variables X and Y are independent with the symbol $X \perp\!\!\!\perp Y$. If the joint distribution function admits a density $p_{X_1,....,X_n}$, then the variables $X_1, ..., X_n$ are independent if and only if

$$p_{X_1,....,X_n}(x_1, ..., x_n) = \Pi_{i=1}^{n} p_{X_i}(x_i)$$

EXAMPLE 9

Let Σ be an $n \times n$ symmetric, positive-definite matrix (spd, for short) and $\mu \in \mathbb{R}^n$. Then we say that $X = (X_1, ..., X_n)^T$ is n-dimensional Gaussian distributed, denoted $X \sim \mathcal{N}(\mu, \Sigma)$ if the distribution function for X is given by

$$\frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

We call μ the **mean vector** of X and Σ the **covariance matrix** of X. In this case, independence of the components of X is governed by the details of Σ^{-1}. In particular, the entries of X will be independent if and only if Σ is diagonal.

Conditional Probability

Asking about event A *given knowledge of event* B is, in general, different from asking about A. We denote this kind of probability as describing a **conditional** event, $A|B$, read “A given B". We measure probability of conditional events by

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Clearly then, if $A \perp\!\!\!\perp B$ then $\mathbb{P}(A|B) = \mathbb{P}(A)$. We like to use the intuition that conditionals represent causal chains $B \to A$, because the conditional probability tells us about how event B can change the probability of event A.

This leads quite naturally to conditional random variables in the obvious way. Importantly for us, we denote the conditional pdf of two random variables by

$$p(x|y) = \frac{p(x, y)}{p(y)}, \qquad p(y) \neq 0$$

We now work an example typifying the use of conditionals.

Example 10

In a class, it is found that 30% of women received an A, whereas only 25% of men did. The class is %60 female. A student is selected at random (uniformly) and found to have received an A. What's the probability this student is female?

We're gong to introduce some variables. We let A denote the event of getting an A, W denote the student being female, and M denote the student being male. We *want* $\mathbb{P}(W|A)$. For this

$$\begin{aligned}\mathbb{P}(W|A) &= \frac{\mathbb{P}(W,A)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(W,A)}{\mathbb{P}(A|W)\mathbb{P}(W)+\mathbb{P}(A|M)\mathbb{P}(M)}\end{aligned}$$

Here, we've used $\mathbb{P}(A) = \mathbb{P}(A|W)\mathbb{P}(W)+\mathbb{P}(A|M)\mathbb{P}(M)$ which follows directly from the marginal $\mathbb{P}(A) = \mathbb{P}(A,M)+\mathbb{P}(A,W)$. Next, we see

$$\mathbb{P}(W|A) = \frac{\mathbb{P}(A|W)\mathbb{P}(W)}{\mathbb{P}(A|W)\mathbb{P}(W)+\mathbb{P}(A|M)\mathbb{P}(M)} = \frac{.3\times .6}{.3\times .6+.25\times .4} \approx 64\%$$

Put differently, knowing the student received an A helped to gain predictive power over the sex of the student by an increase of close to 10%. In this way, conditionals help gain predictive and explanatory power.

Example 11

Suppose we are interested in screening for deadly Disease X. It is known that about .3% of people carry Disease X. Medical researchers developed a test for Disease X which is quite good. The test they've developed carries only a 1% chance of false positives or false negatives. This is quite good in terms of modern standards of reliability in medical screening. We select a person from the population at random (uniformly) and give this person the test for Disease X. The test comes back positive. What's the probability this person has Disease X?

To begin, I'm going to create some variables. I denote by $X+$ and $X-$ the event that the person does or does not have Disease X, respectively. I also denote by TP and TN whether the person *tests* positive or negative for the disease, respectively. Then we are interested in finding $\mathbb{P}(X+|TP)$. For this, we have

$$\begin{aligned}\mathbb{P}(X+|TP) &= \frac{\mathbb{P}(X+,TP)}{\mathbb{P}(TP)} \\ &= \frac{\mathbb{P}(TP|X+)\mathbb{P}(X+)}{\mathbb{P}(TP|X+)\mathbb{P}(X+)+\mathbb{P}(TP|X-)\mathbb{P}(X-)} \\ &= \frac{.99\times .003}{.99\times .003+.01\times .997} \\ &\approx 26\%\end{aligned}$$

In other words, despite the reliable test indicating the person had the disease, they are about three times more likely to not have the disease than to have it.

The above is a classic case of the use of conditionals leading to somewhat counterintuitive results. [5] We close this section by noticing that from $\mathbb{P}(A|B) = \frac{\mathbb{P}(A,B)}{\mathbb{P}(B)}$ we have $\mathbb{P}(A,B) = \mathbb{P}(A|B)\mathbb{P}(B)$. But $\mathbb{P}(A,B) = \mathbb{P}(B,A)$, so playing the same trick but with the variables swapped gives $\mathbb{P}(A,B) = \mathbb{P}(B|A)\mathbb{P}(A)$. Equating the two leads to

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

a result known as **Bayes Rule** after philosopher Thomas Bayes. A common usage of Bayes rule is with events H and D denoting hypothesis being true and data, respectively. Then, Bayes rule would read

$$\mathbb{P}(\text{hypothesis given the data}) = \frac{\mathbb{P}(\text{data given the hypothesis})\mathbb{P}(\text{hypothesis})}{\mathbb{P}(\text{data})}$$

Namely, the likelihood of a given hypothesis being true is determined by how likely the hypothesis is to produce the data in question, multiplied by the relative probabilities of hypothesis and data in absence of other considerations. In this view, the *a priori* probability of the hypothesis, $\mathbb{P}(\text{hypothesis})$ can be see to often hold a decisive role in adjudicating the likelihood of a given hypothesis being true. Put differently, *more reasonable hypothesis should be treated more credibly* and *less reasonable hypotheses should demand stronger evidence*. When moving to random variables Bayes rule takes on the form

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}, \qquad p(x) \neq 0$$

It is in this form that Bayes rule will become powerful for us in modelling and inference later on. And, in Bayesian contexts $p(y|x)$ is called the *posterior* or *a posteriori* distribution, $p(y)$ is called the *prior* or *a priori* distribution and $p(x|y)$ is called the *likelihood function* (as it determined the likelihood of observing the data if the hypothesis were true in our earlier example). We often won't be concerned much with the numerical factor $p(x)$

[5] Please, do *NOT* misinterpret this as calling into question the reliability or utility of medical tests. This only has implications for cautioning against *naive interpretation* of results of such tests. Put differently, results like this are why doctors are reluctant to test for disease in the absence of symptoms. The culprit in this example, after all, was the very low prevalence of the disease in the public. Ignoring prevalence is a mistake known as the *base-rate fallacy* and is a common cause of people overestimating the likelihoods of rare outcomes.

in our examples later on as we will be looking to perform optimization of $p(y|x)$ with x fixed, so $p(x)$ would just cause an unimportant overall global scaling.

Some Statistics

Consider a discrete random variable $X : \mathcal{F} \to \mathbb{R}$ on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose we consider n samples of the random variable (known also as “draws"). Then we look at the average of our sample,

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

where X_i is the $i'th$ instance of a draw of X. Of course, the value of this becomes

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{l} x_i n_i$$

where $n_i = \#\{X = x_i \text{ in } n \text{ trials}\}$. This sum is simply $\sum_{i=1}^{l} x_i \frac{n_i}{n}$. And, we observed before, in our discussion of pmfs, that $\lim_{n \nearrow \infty} \frac{n_i}{n} = p_X(x_i)$. Therefore, we see that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \sum_{x_i \text{ outcomes}} x_i p_X(x_i)$$

Put differently, in the limit of large samples, the mean of the sample approaches an weighted sum of the samples, where the weight is the relative frequency of each outcome. This weighted sum will play a fundamental role in our undrstanding of random variables.

DEFINITION 7: EXPECTATION OPERATOR

Let X be a random variable on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We define a linear functional $\mathbb{E} : X \to \mathbb{R}$, called the **expectation operator** via

- $\mathbb{E}X = \sum_{x_i} x_i \mathbb{P}(X = x_i)$ when X is a discrete random variable
- $\mathbb{E}X = \int_{\mathbb{R}} x p_X(x) dx$ when X is a continuous random variable

The result $\mathbb{E}X$, in either case, is called the **expected value of** X (or *mean value*), which we often denote by $\mu = \mathbb{E}X$.

EXAMPLE 12

We consider $X \sim B(\theta)$. Then

$$\begin{aligned} \mathbb{E}X &= \sum_{x_i} x_i p(x_i) \\ &= \sum_{x_i \in \{0,1\}} x_i \theta^{x_i} (1-\theta)^{1-\mathbf{x}_i} & = \theta \end{aligned}$$

We want to be able to understand how the values of a random variable distribute themselves, and a natural way to look into this is by studying how they disperse from their expected value. For this we use the L^2 distance, as in the following definition.

DEFINITION 8: VARIANCE

Let X be a random variable on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The **variance** of X is defined by

$$var(X) \doteq \mathbb{E}[(X - \mathbb{E}X)^2]$$

This is often denoted σ_X^2.

Of course, for a continuous variable, the formula for variance reduces to

$$var(X) = \int_{\mathbb{R}} (x - \mu_X)^2 p_X(x) dx$$

The variance measures the *average spread of the values* that X takes on. Notice that $var(cX) = c^2 var(X)$ for any constant c. Moreover

$$\begin{aligned} var(X) &= \mathbb{E}(X - \mu_X)^2 \\ &= \mathbb{E}(X^2 - 2X\mu_X + \mu_X^2) \\ &= \mathbb{E}[X^2] - 2\mu_X \mathbb{E}[X] + \mu_x^2 \\ &= \mathbb{E}[X^2] - (\mathbb{E}X)^2 \end{aligned}$$

yielding a commonly used form for calculating the variance in terms of the *second moment* $\mathbb{E}[X^2]$.

Another often-useful statistic for us is the **standard deviation** of a random variable, which is the number $\sigma_X \doteq \sqrt{var(X)}$. You should verfiy that $\sigma_{cX} = |c|\sigma_X$.

Example 13

We again consider the case of $X \sim B(\theta)$. Then

$$\begin{aligned} var(X) &= \mathbb{E}(X-\theta)^2 \\ &= \sum_{x_i \in \{0,1\}} (x_i-\theta)^2 \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^2(1-\theta) + (1-\theta)^2\theta \\ &= \theta(1-\theta) \end{aligned}$$

Thus, as well, we have the standard deviation of a Bernoulli-distributed variable is given by $\sigma_X = \sqrt{\theta(1-\theta)}$. From this, for a fair coin ($\theta = \frac{1}{2}$), we get the standard deviation in the outcome is given by $\frac{1}{2}$.

When a vector valued random variable varies, it can vary from its mean in a more complicated way, since the are now multiple dimensions in which it can move. The following definition is our generalization of variance to this kind of situation.

Definition 9: Covariance Matrix

Let $X = (X-1, ..., X_n)^T$ be a random vector. Then

$$\mathbb{E}((X-\mathbb{E}X) \otimes (X-\mathbb{E}X))_{ij} \doteq cov(X_i, X_j)$$

are the entries of the **covariance matrix** of X.

Notice, in the above, that the covariance matrix is necessarily a symmetric matrix. Moreover, given two random variables, X and Y, we have

$$cov(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$$

gives the **covariance** between the variables X and Y. Notice the covariance is a weighted sum (or integral for continuous variables). The sum gets larger when the summand is positive. This happens when both X and Y have propensity to stray from their mean values *at the same time and in same direction*. If they simultaneously stray from their mean values but in the opposite direction, then we expect a negative summand. In this way, the covariance gives us a measure of how well synchronized the variables X and Y are to each other; it tells us how X and Y “move together". Related to the covariance of two variables X and Y is a normalized version called the **correlation coefficient**, ρ_{XY} given by

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

This statistic can be more informative because it blurs out the fact that certain numerical values will be large or small and focusses on simply how well “in sync" the two variables are to each other. Namely, it shouldn’t matter if both variables stray 10 units or 10^{10} units from their mean in the same way, provided they are actually doing so in lockstep. The correlation coefficient normalizes for this.

Suppose that we have two independent random variables X and Y with respective means μ_X and μ_Y. Notice, first off that we have, from independence and Fubini’s theorem of multivariable calculus,

$$\begin{aligned} \mathbb{E}XY &= \int_{\mathbb{R}^2} xyp_{XY}(x,y)dxdy \\ &= \int_{\mathbb{R}^2} xyp_X(x)p_Y(,y)dxdy \\ &= \int_{\mathbb{R}} xp_X(x)dx \int_{\mathbb{R}} yp_Y(y)dy \\ &= \mathbb{E}X\mathbb{E}Y \end{aligned}$$

Then,

$$\begin{aligned} cov(X,Y) &= \mathbb{E}(X-\mu_X)(Y-\mu_Y) \\ &= \mathbb{E}[XY] - \mu_Y\mathbb{E}X - \mu_X\mathbb{E}Y + \mu_X\mu_Y \\ &= \mathbb{E}X\mathbb{E}Y - \mu_X\mu_Y \\ &= 0 \end{aligned}$$

In other words **independent variables are uncorrelated**. We leave as a simple exercise the fact that $var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$. But, given this fact, we then see that

$$X \perp\!\!\!\perp Y \implies var(X+Y) = var(X) + var(Y)$$

In fact, it’s the above feature of variance that makes it so useful in practice; variance is additive over independent variates.

Theorem 10

Let $X_1, ..., X_n$ be random variables on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with finite variances $\sigma_i^2 < \infty$. Define $S_n = \sum_i X_i$. Then

$$var(S_n) = \sum_i \sigma_i^2 + 2\sum_{j<k} cov(X_j, X_k)$$

Proof

We have $\mu_i = \mathbb{E}X_i$ and then $m = \mathbb{E}S_n = \sum_i \mu_i$. Then $S_n - m =$

$\sum_i (X_i - \mu_i)$. Therefore

$$\begin{aligned}(S_n - m)^2 &= \sum_i (X_i - \mu_i) \sum_j (X_j - \mu_j) \\ &= \sum_i (X_i - \mu_i)^2 + \sum_{i \neq j} (X_i - \mu_i)(X_j - \mu_j) \\ &= \sum_i (X_i - \mu_i)^2 + 2 \sum_{i<j} (X_i - \mu_i)(X_j - \mu_j)\end{aligned}$$

Taking expectations of both sides yields the result. □

Therefore, we get the following corollary.

Corollary 1

Let $X_1, ..., X_n$ be uncorrelated random variables. Then

$$var(\sum_i X_i) = \sum var(X_i)$$

Entropy and Information

Much of this section is influenced by the wonderful series "A Human's Guide to Words" by Yudkowsky[6]. Consider now a random variable $X \in \{X_1, ..., X_8\}$, where each X_i is equally likely. If you wish, X can represent the outcome of a toss of a fair eight-sided die. We'd like to know which state X is in by querying someone who has tossed the die beyond our view. I get to query by asking only yes/no questions. So, for instance, my strategy for asking could be based on the following division

$$\{\underbrace{X_1, X_2, X_3, X_4}_{0 \text{ (no)}}, \underbrace{X_5, X_6, X_7, X_8}_{1 \text{ (yes)}}\}$$

This corresponds to me asking the tosser "did the die come up in either of X_1, X_2, X_3, X_4?". Suppose I'm allowed to ask more questions. Then I can split the outcomes I'm querying on further. One such division would be

$$\{\underbrace{X_1, X_2}_{00}, \underbrace{X_3, X_4}_{01}, \underbrace{X_5, X_6}_{10}, \underbrace{X_7, X_8}_{11}\}$$

These are placeholders for queries such as "Was the outcome in the first four states?" followed by, say, "Good, now is it in the first half of that subset of four states?", and the like. In this way we may end up settling on the following **encoding scheme**

$$X_1 : 001, X_2 : 010, X_3 : 011, X_4 : 100$$
$$X_5 : 101, X_6 : 110, X_7 : 111, X_8 : 000$$

[6] Elizier Yudkowsky. https://wiki.lesswrong.com/wiki/A_Human's_Guide_to_Words

Which allows us to figure out the outcome of the die-toss in at most three binary questions. We say, in this instance, that *X has entropy* of 3 bits. In this sense, the entropy measures the average number of binary queries it takes to determine the value of a random variable. As another example, consider a random variable Y such that

$$\begin{aligned}
\mathbb{P}(Y = Y_1) &= \frac{1}{2} \\
\mathbb{P}(Y = Y_2) &= \mathbb{P}(Y = Y_3) = \frac{1}{8} \\
\mathbb{P}(Y = Y_4) &= \frac{1}{4}
\end{aligned}$$

The entropy in this instance can be found by using the following encoding

$$Y_1 : 1, \quad Y_2 : 01, \quad Y_3 : 001, \quad Y_4 : 000$$

In this way, half time we need only ask a single binary question (“is the first digit a one?"), a quarter of the time we’ll need to ask two binary questions, and an eighth of the time three queries are required (in two possible outcomes). Therefore

$$\begin{aligned}
\mathbb{E}[\text{number of queries}] &= \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 3 \times \frac{1}{8} \\
&= 1.75
\end{aligned}$$

This is of course, just $-\mathbb{E}[\log_2(p_Y)]$ as

$$\begin{aligned}
\mathbb{E}[\log_2(p_Y)] &= \mathbb{P}(Y = Y_1)\log_2 \mathbb{P}(Y = Y_1) + \mathbb{P}(Y = Y_2)\log_2 \mathbb{P}(Y = Y_2) \\
&\quad + \mathbb{P}(Y = Y_3)\log_2 \mathbb{P}(Y = Y_3) + \mathbb{P}(Y = Y_4)\log_2 \mathbb{P}(Y = Y_4) \\
&= \frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{4}\log_2(\frac{1}{4}) + \frac{2}{8}\log_2(\frac{1}{8}) \\
&= -(\frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 6 \times \frac{1}{8}) \\
&= -1.75
\end{aligned}$$

In agreement with the earlier calculation. Notice that on average we expect to only have to ask 1.75 questions, which is lower than the naive estimate of 2 which we would expect were all outcome states equally likely. So knowing the probability isn’t evenly distributed helped us come up with an improved encoding scheme which reduced the number of queries we’d require on average.

Notice, in the prior example, that we used the following heuristic in our encoding scheme: probable events carry less information. In other words *probable events should have shorter codes and therefore require fewer guesses*. Of course, intuitively that’s true. The more probable an outcome is the less uncertainty an instance of it occurring resolves.[7] The number of guesses we expect to make is a proxy for what we might call the “unpredictability" of outcomes.

[7] Notice that this is the reason why commonly used words happen to be short, e.g. “the", “a", “but", “time", “if", etc. Words are coding concepts and the more frequently used words ought to have shorter codes. Although natural language may naively *seem* arbitrary, it’s a general pattern that languages with this heuristic are the ones that ended up sticking around.

Definition 11: Entropy of a Random Variable

Let $X \sim p_X$ be a random variable on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then

$$H(X) \doteq -\mathbb{E}[\ln p_X]$$

is the **entropy** of X. When $p_X = 0$ we use $p_X \ln p_X = 0$. Entropy is measured in “nats".

The literature also often uses $H(X) \doteq -\mathbb{E}[\log_2 p_X]$ as we were in our introductory examples, in which case the entropy is measured in “bits" rather than nats. We alternate between using bits and nats.

Example 14

We will calculate the base-2 entropy of a Bernoulli variable $X \sim B(p)$.

$$\begin{aligned} -H(X) &= -\mathbb{E}[\log_2 p_X] \\ &= \mathbb{P}(X=0)\log_2 \mathbb{P}(X=0) + \mathbb{P}(X=1)\log_2 \mathbb{P}(X=1) \\ &= (1-p)\log_2(1-p) + p\log_2 p \end{aligned}$$

i.e. the entropy is $H(X) = (p-1)\log_2(1-p) - p\log_2 p$. Notice that $H(X)$ is then 0 for $p = 0, 1$. In other words *certain outcomes carry zero entropy*. If we wanted to maximize the entropy we would take derivative in p, set to zero and find that $p = \frac{1}{2}$ causes maximum entropy of 1. In other words *fair coins carry the maximum entropy*.

We consider now a *pair* of variables X and Y with the following properties: X takes one of 8 possible outcomes with equal likelihood and Y takes one of 4 probabilities with equal probability. We ask what is the entropy of the system (X, Y)? On the one hand we now that $H(X) = 3$ and $H(Y) = 2$ which would suggest that $H(X, Y) = 3 + 2 - 5$. Yet, suppose that X and Y are non-independent. For instance, what if it were known that upon any observation of the pair of uniformly distributed variables X and Y we saw that either both X and Y were odd or both were even? This implies that, once we determine if, say, $X = X_5$, which can be done in at most 3 queries, we have left to determine whether $Y = Y_1$ or $Y = Y_3$, which can be answered in a single query. In other words (X, Y) can be determined in at most 4 queries. This reasoning is clearly independent of whichever value of X we happen to observe. So, in fact, $H(X, Y) = 4 < H(X) + H(Y)$. Somehow the degree of *shared information* among the two variables served to decrease the total amount of information from what it might have naively been. We've implicitly been using the **joint entropy** of a pair of random variables (X, Y) as

$$H(X, Y) = -\mathbb{E}[\log p_{XY}]$$

or $H(X,Y) = -\sum_{x,y} p_{XY}(x,y) \log p_{XY}(x,y)$. What we care about then is the degree to which variables can share information, as in the preceding example, in such a way as to lower the entropy from what it might have been in the case where the variables are independent.

DEFINITION 12: MUTUAL INFORMATION

Given random variables X, Y the **mutual information** between X and Y is given by

$$I(X,Y) \doteq H(X) + H(Y) - H(X,Y)$$

In the case of the example considered in the above we see that $I(X,Y) = 4$. As well, it can be shown (see the exercises) that the mutual information between variables is zero when and only when the variables are independent, in congruity with the informal notion of shared information.

Relative Entropy

Consider the following situation, as described in the classic work by Kullback[8]: we're given draws of a random variable X from an unknown distribution and we wish to decide which, among two competing distributions, the variable X is distributed according to. Namely, we have

$$\text{Either} \quad X \sim p_X \qquad \text{or} \qquad X \sim q_X$$

We let H_i denote the associated *hypotheses*, i.e. H_i shall denote the hypothesis that $X \sim f_i$ where $f_1 = p_X$ and $f_2 = q_X$. We shall now try to weight the evidence, provided to us by samples of X, in favour of each distribution. Bayes rule says that

$$\begin{aligned}
\mathbb{P}(H_i \mid x) &= \frac{\mathbb{P}(x \mid H_i)\mathbb{P}(H_i)}{\mathbb{P}(x)} \\
&= \frac{\mathbb{P}(x \mid H_i)\mathbb{P}(H_i)}{\mathbb{P}(x \mid H_1)\mathbb{P}(H_1) + \mathbb{P}(x \mid H_2)\mathbb{P}(H_2)} \\
&= \frac{\mathbb{P}(x \mid H_i)\mathbb{P}(H_i)}{p_X(x)\mathbb{P}(H_1) + q_X(x)\mathbb{P}(H_2)} \\
&= \frac{\frac{\mathbb{P}(x|H_i)\mathbb{P}(H_i)}{q_X(x)\mathbb{P}(H_2)}}{1 + \frac{p_X(x)}{q_X(x)}\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_2)}}
\end{aligned}$$

From this we get

$$\mathbb{P}(H_1 \mid x) = \frac{\frac{p_X(x)}{q_X(x)}\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_2)}}{1 + \frac{p_X(x)}{q_X(x)}\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_2)}} \qquad P(H_2 \mid x) = \frac{1}{1 + \frac{p_X(x)}{q_X(x)}\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_2)}}$$

and therefore

$$\frac{\mathbb{P}(H_1 \mid x)}{\mathbb{P}(H_2 \mid x)} = \frac{p_X(x)}{q_X(x)}\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_2)}$$

[8] Solomon Kullback. *Information Theory and Statistics*. Dover Publications Inc., 1978

from which we can calculate the “log-odds" given by

$$\log(\frac{p_X}{q_X}) = \log(\frac{\mathbb{P}(H_1 \mid x)}{\mathbb{P}(H_2 \mid x)}) - \log(\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_2)})$$

The right hand side of the above gives the difference, in log-odds, of the hypothesis before and after measurement of $X = x$. Namely, it tells how our odds for a given hypothesis are affected by a value obtained in a draw from X. In this way the right hand side tells us *how the information of observing $X = x$ can be used to discern distributions*. In this way $\log(\frac{p_X}{q_X})$ tell us about the information encoded in draws of X, which would allow us to help decide whether X was distributed according to p_X or q_X, deciding in favour of H_1 versus H_2. This leads to the following definition

Definition 13: Relative Entropy

Let X be a random variable and let p, q be two density functions. Then

$$\mathbb{E}_{X\sim p}[\log \frac{p}{q}] = \int_{\mathbb{R}} p(x) \log(\frac{p(x)}{q(x)})dx$$

is the **relative entropy** between the distributions p, q. This is often called the **Kullback-Leibler divergence** and denoted

$$D_{KL}(p \parallel q) \doteq \mathbb{E}_{X\sim p}[\log \frac{p}{q}]$$

A few remarks are in order.

- If the measures induced by p and q agree up to sets of measure zero then $D_{KL}(p \parallel q) = 0$. This means that the great agreement in induced measures implies there’s no valuable information allowing one to discern the true distribution of X.

- If X and Y are independent with respect to both p and q then

$$D_{KL}(p_{XY} \parallel q_{XY}) = D_{KL}(p_X \parallel q_X) + D_{KL}(p_Y \parallel q_Y)$$

- Although $D_{KL}(p \parallel q)$ can be thought of as a kind of distance between distributions, it strictly speaking isn’t. For one thing, it isn’t symmetric, namely $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$

As an example of the utility of the relative entropy, we consider the following situation: Suppose that we are given data $X \sim p_X$ where we don’t know p_X. We’d like to be able to find a nice approximation or estimate, q of the true data-generating distribution. This amounts to finding a q which is “as close as possible" to p_X, where closeness will be determined by the KL-divergence between them. We’ll denote by Q a space of distributions from which we will check various q’s. Then we may use

$$q^* = \arg\min_{q\in Q} D_{KL}(p \parallel q)$$

Of course, we could also use

$$\tilde{q}^* = \arg\min_{q \in \mathcal{Q}} D_{KL}(q \parallel p)$$

And, in general, q^* and $\tilde{q}^*$ will not agree. To see why not, suppose that p was a sum of two Gaussians with equal variance but unequal means, and suppose that $\mathcal{Q}$ is equal to the set of all normal distributions. In other words, we're trying to find the best normal approximant to a bimodal sum of normals. Then q^* is obtained by averaging over both peaks (as the integral is calculated by integrating against the bimodal distribution) whereas $\tilde{q}^*$ can be computed by localizing around a single peak of p_X as the remaining mass is small in q. The choice of whether to use q^* or $\tilde{q}^*$ is one of design and practicality. We also remark that $D_{KL}(p \parallel q) = -H(X \sim p) - \int_{\mathbb{R}} p(x) \log q(x) dx$. In other words

$$-\mathbb{E}_{X \sim p}[\log q] = D_{KL}(p \parallel q) + H(X \sim p)$$

We use the above to define the **cross-entropy** of p with q, given by

$$H(p,q) \doteq -\mathbb{E}_{X \sim p}[\log q]$$

This becomes a commonly encountered quantity when we are performing estimation of densities and finding q^* as above, since

$$\begin{aligned} \arg\min_{q \in \mathcal{Q}} D_{KL}(p \parallel q) &= \arg\min_{q \in \mathcal{Q}} (-H(X \sim p) + H(p,q)) \\ &= \arg\min_{q \in \mathcal{Q}} H(p,q) \end{aligned}$$

In other words, minimizing the KL-divergence amounts to minimizing the cross-entropy between two densities.In practice, especially in machine learning applications, the cross-entropy is therefore commonly encountered as a **loss function** which is hoped to be minimized.

Some Inequalities

In the following theorem we present some of the most widely used inequalities in probability theory.

Theorem 14

Let X and Y be random variables on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then the following inequalities hold.

1. $\mathbb{E}|X+Y|^r \leq c_r(\mathbb{E}|X|^r + \mathbb{E}|Y|^r), r > 0$ with $c_r = 1$ for $r \leq 1$ and 2^{r-1} otherwise.

2. $\mathbb{E}|XY| \leq (\mathbb{E}|X|^r)^{\frac{1}{r}}(\mathbb{E}|Y|)^{\frac{1}{s}}$ for $r > 1$ and $\frac{1}{r} + \frac{1}{s} = 1$. This is **Hölder's inequality** and generalizes the Cauchy-Schwartz inequality which occurs for $r = s = \frac{1}{2}$.

3. $f(\mathbb{E}X) \leq \mathbb{E}f(X)$ for all convex functions f. This is known as **Jensen's inequality**.

4. Let ϕ be a non-decreasing, monotonic even function on $[0, \infty)$. Then, for all $a > 0$ we have

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[\phi(X)]}{\phi(a)}$$

This is known as **Markov's inequality**.

Proof

We leave the first claim to be proven by the reader. For Hölder's inequality, we first notice that, for all positive numbers a, b we have

$$\frac{a^r}{r} + \frac{b^s}{s} \geq ab$$

whenever $\frac{1}{r} + \frac{1}{s} = 1$. Now, set $A = (\mathbb{E}|X|^r)^{\frac{1}{r}}$ and $B = (\mathbb{E}|Y|)^{\frac{1}{s}}$. If $AB = 0$ there's nothing to prove as this forces $|XY| = 0$. Moreover, if either of A or B is infinite the inequality is obvious. Thus, we assume that $0 < AB < \infty$. Now with $a = \frac{|X|}{A}$ and $B = \frac{|Y|}{b}$ we get

$$\frac{|XY|}{AB} \leq \frac{|X|^r}{rA^r} + \frac{|Y|^s}{sB^s}$$

Taking $\mathbb{E}$ of both sides of the above inequality yields the result.
For Jensen's inequality we only give a sketch. For convex functions we know that $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$ holds for $\theta \in [0,1]$. Induction then gives that

$$f(\sum_i \theta_i x_i) \leq \sum_i \theta_i f(x_i)$$

for all finite sums $\sum_i \theta_i = 1, \theta_i \geq 0$. Proceeding in this way by viewing θ_i's as $p(x_i)$'s completes a sketch of the main idea. Finally, for Markov's inequality, we let $a > 0$. Then

$$\begin{aligned} \phi(a)\mathbb{1}_{\{|X| \geq a\}} &\leq \phi(|X|)\mathbb{1}_{|X| \geq a} \\ &\phi(|X|) \\ &\leq \phi(X) \end{aligned}$$

Thus

$$\begin{aligned} \mathbb{E}[\phi(a)\mathbb{1}_{\{|X| \geq a\}}] &\leq \phi(a)\mathbb{E}\mathbb{1}_{\{|X| \geq a\}} \\ &= \phi(a)\mathbb{P}(|X| \geq a) \\ &\leq \mathbb{E}\phi(X) \end{aligned}$$

as was to be shown. □

Conditional Expectation

In this section we present results in a quasi-rigourous way. Namely, to avoid too many technical detours, we give some very hand-wavy justifications of the main abstract constructions. The interested reader is advised to pick up a copy of the book by Billingsley [9].

Let X and Y be two random variables on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Given a conditional density function $p(x|y)$ on our variables we define $\mathbb{E}[X|Y = y]$ to be $\sum_x xp(x|y)$ in the case of discrete random variables or $\int_{\mathbb{R}} xp(x|y)dx$ in the case of continuous variables. This definition gives us the **conditional expectation** $\mathbb{E}[X|Y = y]$, as a function of the values the conditioned-upon variable can take. Similar definitions hold for things like $\mathbb{E}[f(X)|Y = y]$ or when dealing with vector-valued random variables.

In the definition given above it's clear that $h(y) = \mathbb{E}[X|Y = y]$ defines a function of y (notice we've integrated out on X's values). As it happens $h(Y)$ is a $\sigma(Y)$-measurable function. In other words, the conditional expectation is itself a random variable (measured with respect to the information contained in Y as dictated by it's σ-algebra). Let's investigate why this is so in a bit more detail.

Consider the σ-algebra generated by the random variable Y,

$$\sigma(Y) = \{Y^{-1}(B) \mid B \in \mathcal{B}\}$$

A theorem in mathematical analysis, the **Radon-Nikodym theorem** says the following: If two measures μ and ν on a space $(\Omega, \mathcal{F})$ satisfy[10] $\nu << \mu$ then there exists a function f such that

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

The function f is called the *Radon-Nikodym derivative* of ν with respect to μ, denoted $\frac{d\nu}{d\mu}$. Now, in the case at hand consider the measure

$$\mu_Y(A) \doteq \int_{Y^{-1}(A)} X d\mathbb{P}$$

Clearly, here we have that $\mathbb{P}(Y \in A) \implies \mu_Y(A) = 0$. Therefore $\mu_Y << \mathbb{P}$ and the Radon-Nikodym derivate $\frac{d\mathbb{P}}{d\mu_Y}$ exists. But that simply means that there is a function $(\frac{d\mathbb{P}}{d\mu_Y})$ let's call g with the property that

$$\int_{Y^{-1}(A)} X d\mathbb{P} = \int_A g d\mu_Y$$

This function g *is the conditional expectation* $\mathbb{E}[X|Y]$. In other words

$$\int_{Y^{-1}(A)} X d\mathbb{P} = \int_A \mathbb{E}[X|Y] d\mu_Y$$

Notice that, provided X is $\sigma(Y)$-measurable, we have $\mathbb{E}X = \mathbb{E}(\mathbb{E}[X|Y])$. This is also sometimes denoted

$$\mathbb{E}[X|\sigma(Y)]$$

[9] Patrick Billingsley. *Probability and Measure*. Wiley

[10] The notation $\nu << \mu$ means, simply, that $\mu(B) = 0 \implies \nu(B) = 0$ for all $B \in \mathcal{F}$

where the more general form $\mathbb{E}[X|\mathcal{F}]$ for a σ-field $\mathcal{F}$ is like our best guess of X given the information available in the σ-field $\mathcal{F}$.

Example 15

If X is $\mathcal{F}$-measurable then $\mathbb{E}[X|\mathcal{F}] = X$.

Example 16

If, for all $B \in \mathbb{R}$ and all $A \in \mathcal{F}$ the random variable X has the property that $\mathbb{P}(\{X \in B\} \cap A) = \mathbb{P}(X \in B)\mathbb{P}(A)$ holds, then we say that X is "independent of $\mathcal{F}$". When this is the case we have

$$\begin{aligned}\int_A X d\mathbb{P} &= \mathbb{E}[X\mathbb{1}_A] \\ &= \mathbb{E}X\mathbb{E}\mathbb{1}_A \\ &= \int_A \mathbb{E}X d\mathbb{P}\end{aligned}$$

and so $\mathbb{E}[X|\mathcal{F}] = \mathbb{E}X$. This means, essentially, that if you have no information about $\mathcal{F}$, your best guess for X is $\mathbb{E}X$.

Example 17

Q: On every day except for Sundays, a train leaves the station every s minutes. On Sundays it leaves every $2s$ minutes. You arrive at the station not knowing anything about what time or day it currently is. How long should you expect to wait until the next arrival?

A: For this, we set T = waiting time random variable. Then, we know that

$$T|workdays \sim p(t|workdays) = \frac{\mathbb{1}_{[0,s]}(t)}{s}$$

and

$$T|Sunday \sim p(t|Sunday) = \frac{\mathbb{1}_{[0,2s](t)}}{2s}$$

Then we have

$$\begin{aligned}\mathbb{E}[T|Workday] &= \int tp(t|workday)dt \\ &= \frac{1}{s}\int_0^s tdt \\ &= \frac{s}{2}\end{aligned}$$

And also by similar reasoning we have $\mathbb{E}T|Sunday] = 2s$. Now also $\mathbb{P}(workday) = \frac{6}{7}$ and $\mathbb{P}(Sunday) = \frac{1}{7}$. Thus, using marginals (make sure you can justify this!) we have

$$\begin{aligned}\mathbb{E}[T] &= \mathbb{E}[T|workday]\mathbb{P}(workday) + \mathbb{E}[T|Sunday]\mathbb{P}(Sunday)\\ &= \frac{s}{2}\frac{6}{7} + 2s\frac{1}{7}\\ &= \frac{5s}{7}\end{aligned}$$

You should expect to wait $\frac{5s}{7}$ for the next train.

Example 18

Q: I roll a die and it comes up even. What's the expected roll?

A: We have that $\mathbb{E}[roll|even] = \sum_{roll\in\{1,2,3,4,5,6\}} roll\mathbb{P}(roll|even)$. This is $1\times 0 + 2\times\frac{1}{3} + 3\times 0 + 4\times\frac{1}{3} + 5\times 0 + 6\times\frac{1}{3} = 4$.

Example 19

Q: I roll a die until 6 appears. What should I expect for the sum of all the rolls?

A: I'll call $S =$ sum of all the rolls. I will also label N as the number of rolls it takes to hit a 6. As well, I'll let X_i be the number obtained on the i'th roll. Then, notice that $\mathbb{E}[S|N = n]$ is easy to compute since $\mathbb{E}[S|N = n] = \mathbb{E}(X_1 + \cdots + X_n) = n\mathbb{E}[X_i] = \frac{7n}{2}$. Thus,

$$\begin{aligned}\mathbb{E}[S] &= \mathbb{E}(\mathbb{E}[S|N])\\ &= \sum_n \mathbb{E}[S|N = n]\mathbb{P}(N = n)\\ &= \frac{7}{2}\sum_n n\mathbb{P}(N = n)\\ &= \frac{7}{2}\mathbb{E}N\end{aligned}$$

Notice that $\mathbb{P}(N = n)$ should be $\mathbb{P}(N = n, N \neq n-1, N \neq n-2, ..., N \neq 1)$. But the outcomes of the tosses are clearly independent. Thus, the probabilities multiply and we get $\frac{1}{6}(\frac{5}{6})^{n-1}$. This gives then $\mathbb{E}N = \sum_{n\geq 1}\frac{1}{6}(\frac{5}{6})^{n-1}n = 6$. Therefore $\mathbb{E}S = \frac{7}{6}\times 6 = 7$. Here we've used the result (see exercises) that for $p \in (0,1)$ we have $\sum_{k\geq 1} kp(1-p)^{k-1} = \frac{1}{p}$.

Characteristic Functions

We define a **characteristic function** of a random variable to be (inverse) Fourier transform of the density function.

Definition 15: Characteristic Function

Given a random variable $X \sim p_X(x)$ we define the **characteristic function** of X to be

$$\phi(t) \doteq \mathbb{E}[e^{itX}]$$

sometimes denoted $\phi_X(t)$.

If you are familiar with properties or uses of the Fourier transform, you will know that the characteristic function therefore gives a spectral view of the behaviour of a distribution. In this sense, it plays a dual role to the “density-centric" view we’ve adopted thus far. We remark that

- $\phi(0) = 1$
- $\phi(-t) = \bar{\phi}(t)$
- $|\phi(t)| \leq 1$

holds for the characteristic function of any random variable.

Example 20

We consider a (variant of a) Bernoulli distributed random variable X such that $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{2}$. Then

$$\begin{aligned}\phi(t) &= \mathbb{E}[e^{itX}] \\ &= e^{-it}\mathbb{P}(X = -1) + e^{it}\mathbb{P}(X = 1) \\ &= \cos t\end{aligned}$$

Example 21

We consider a standard normal random variable, $X \sim \mathcal{N}(0, 1)$. Then

$$\begin{aligned}\phi(t) &= \int_{\mathbb{R}} e^{itx} \frac{1}{2\pi} e^{-\frac{x^2}{2}} dx \\ &= e^{-\frac{t^2}{2}} \int_{\mathbb{R}} \frac{1}{2\pi} e^{-\frac{(x-it)^2}{2}} dx \\ &= e^{-\frac{t^2}{2}}\end{aligned}$$

since $\int_{\mathbb{R}} \frac{1}{2\pi} e^{-\frac{(x-it)^2}{2}} dx$ is the integral of a density of a $\mathcal{N}(it, 1)$ variable. Put differently,

$$\phi_{\mathcal{N}(0,1)} = e^{-\frac{t^2}{2}}$$

In general we have that $\phi_{aX+b}(t) = e^{itb}\phi_X(at)$. In other words, characteristic function turns translations into unitary rotations and scaling in X to scaling in the characteristic variable. From these considerations it’s easy to see that

$$\phi_{\mathcal{N}(\mu,\sigma^2)}(t) = e^{i\mu t - \frac{\sigma^2 t^2}{2}}$$

In other words, normal distributions have the property that they look similar on the spectral side to their form on the density side.

Some Limit Laws

We begin this section by introducing some notions of what it means for a sequence of random variables to converge. The differing notions can be quite subtle.

Definition 16: Convergence of Sequence of Random Variables

Consider random variables $X_1, ..., X_n$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P}$. We say that

- The sequence *converges to X almost surely*, denoted $X_n \xrightarrow{a.s.} X$, if $\mathbb{P}(\{\omega \mid \lim_{n\nearrow\infty} X_n = X\}) = 1$
- The sequence *converges to X in p'th mean*, denoted $X_n \xrightarrow{L^p} X$, if $\lim_{n\nearrow\infty} \mathbb{E}|X_n - X|^p = 0$, for $p \geq 1$. This is what analysts would call *convergence in* L^p. $L^p(\Omega)$ can be thought of as the set $\{Y \mid \mathbb{E}|Y|^p < \infty\}$.
- The sequence *converges to X in probability*, denoted $X_n \xrightarrow{\mathbb{P}} X$, if, given any $\epsilon > 0$ we have $\lim_{n\nearrow\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$.
- The sequence *converges to X in distribution*, denoted $X_n \xrightarrow{d} X$ or $F_{X_n} \xrightarrow{d} F_X$ if $\lim_{n\nearrow\infty} F_{X_n}(x) = F_X(x)$ for all x at which F_X is continuous.

We note that the different notions are related in the following ways.

Theorem 17

Consider random variables $X_1, ..., X_n$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P}$. Then

1. $X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{\mathbb{P}} X$
2. $X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{d} X$
3. $X_n \xrightarrow{a.s} X \implies X_n \xrightarrow{\mathbb{P}} X$

Proof

To prove that convergence in the p'th mean implies convergence in probability, we use Markov's inequality with the function $\phi(x) = |x|^p$. Namely, we have that

$$\mathbb{P}(|X_n - X| > \epsilon) \leq \frac{\mathbb{E}|X_n - X|^p}{\epsilon^p}$$

Taking limits proves the assertion. Next, to show that convergence in probability is stronger that convergence in distribution we let $\epsilon > 0$ and notice

$$\begin{aligned} F_{X_n}(x) &= \mathbb{P}(X_n \le x) \\ &= \mathbb{P}(X_n \le x, X > x+\epsilon) + \mathbb{P}(X_n \le x, X \le x+\epsilon) \\ &\underbrace{\le}_{\text{think carefully}} \mathbb{P}(|X_n - X| \ge \epsilon) + \mathbb{P}(X \le x+\epsilon) \\ &= \mathbb{P}(|X_n - X| \ge \epsilon) + F_X(x+\epsilon) \end{aligned}$$

Also, $F_{X_n}(x) \ge F_X(x-\epsilon) - \mathbb{P}(|X_n - X| \ge \epsilon)$. Thus

$$F_X(x-\epsilon) \le \liminf_{n \nearrow \infty} F_{X_n}(x) \le \limsup_{n \nearrow \infty} F_{X_n}(x) \le F_X(x+\epsilon)$$

Taking ϵ small enough we get as tight an inequality as we want proving the claim.
Lastly, we consider the more difficult claim: almost sure convergence implies convergence in probability. We proceed as follows. First, for all integers $k > 0$ define the sets

$$S_{k,l} \doteq \cap_{n=l}^{\infty} \{\omega \mid |X_n - X| < \frac{1}{k}\}$$

Then

$$\{\omega \mid \lim_{n \nearrow \infty} X_n(\omega) = X(\omega)\} = \cap_{k=1}^{\infty} \cup_{l=1}^{\infty} S_{k,l}$$

Therefore we see that $\mathbb{P}(\cap_{k=1}^{\infty} \cup_{l=1}^{\infty} S_{k,l}) = 1 \iff X_n \xrightarrow{a.s.} X$. Notice that

$$S_{k,l} \overset{l \uparrow \infty}{\nearrow} \cup_{l \ge 1} S_{k,l} \quad \text{and} \quad \cup_{l \ge 1} S_{k,l} \overset{k \uparrow \infty}{\searrow} \cap_{k \ge 1} \cup_{l \ge 1} S_{k,l}$$

In other words, we have

$$\mathbb{P}(\cup_{l \ge 1} S_{k,l}) = \lim_{k \to \infty} \mathbb{P}(S_{k,l})$$

Since we have almost sure convergence we have that $\mathbb{P}(\cap_{k=1}^{\infty} \cup_{l=1}^{\infty} S_{k,l}) = 1$ but $\cup_{l \ge 1} S_{k,l} \overset{k \uparrow \infty}{\searrow} \cap_{k \ge 1} \cup_{l \ge 1} S_{k,l}$ shows that the sets $\cup_{l \ge 1} S_{k,l}$ get smaller as they approach the limit. Therefore we must have $\mathbb{P}(\cup_{l \ge 1} S_{k,l}) = 1$ as the limit of 1 is approached from above. Thus, for $\epsilon > 0$ $\{\omega \mid |X_n(\omega) - X(\omega)| \ge \epsilon\} \subset \{\omega \mid |X_n(\omega) - X(\omega| \ge \frac{1}{k}\} \subset \Omega \cap S_{k,l}^c$ whenever $\frac{1}{k} < \epsilon$ and $n \ge l$. This implies that

$$\mathbb{P}(\{\omega \mid |X_n(\omega) - X(\omega)| \ge \epsilon\}) \xrightarrow{n \to \infty} 0$$

as desired.

□

Now that we have introduced what it means for a sequence of random variables to converge to a random variable, we can now discuss some of the most common laws surrounding limiting behaviour of random variables. We begin with the laws of large numbers. In it we use the common convention that iid is shorthand for **independent, identically distributed**. This assumption is very often made on random variables up for consideration for limit laws.

Theorem 18: Laws of Large Numbers

Let $X_1, ..., X_n$ be iid random variables with means $\mathbb{E}[X_i] = \mu$ and finite variance $var(X_i) = \sigma^2 < \infty$. Define the **sample mean** random variable

$$S_n = \overline{X}_n \doteq \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then the following hold true.

- **Weak Law of Large Numbers (WLLN):** $S_n \xrightarrow{\mathbb{P}} \mu$
- **Strong Law of Large Numbers (SLLN):** $S_n \xrightarrow{a.s} \mu$

In the above, the qualifiers “strong" and “weak" are obviously used because of the prior theorem showing that almost sure convergence implies convergence in probability. The above laws of large numbers are important because they say that not only is the sample mean an *unbiased* (a term we’ll explore later) estimator of the true mean of the random variables in question, but it can be used to properly *determine* the true mean of the distribution from which $X_1, ..., X_n$ are sampled from. It is one of the foundational results in mathematical statistics. The other landmark result in mathematical statistics and probability is the following version of the Central Limit Theorem (CLT).

Definition 19: Central Limit Theorem

Let $X_1, ..., X_n$ be iid random variables with means $\mathbb{E}[X_i] = \mu$ and finite variance $var(X_i) = \sigma^2 < \infty$. Define the **sample mean** random variable, as before, as $S_n = \overline{X}_n \doteq \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\frac{\sqrt{n}(S_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

Notice that the above shows that standard normal random variables are somehow *universal* distributions. Under very weak assumptions on the distribution underlying the samples X_i we can get an approximant of a distribution of a standard normal.

Proof

Define $M_n = \frac{\sqrt{n}(S_n-\mu)}{\sigma} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i$ where $Z_i = \frac{X_i-\mu}{\sigma}$. The Z_i are also iid. Then

$$\begin{aligned}
\phi_{M_n}(t) &= \mathbb{E}[e^{itS_n}] \\
&= \mathbb{E}[e^{it\frac{1}{\sqrt{n}}\sum_{i=j}^n Z_j}] \\
&= \mathbb{E}[\Pi_{j=1}^n e^{i\frac{t}{n}Z_j}] \\
&\overbrace{=}^{\text{independence of } Z_j' s} \quad \Pi_{j=1}^n \mathbb{E}[e^{i\frac{t}{n}Z_j}] \\
&= \Pi_{j=1}^n \phi_{Z_j}(\frac{t}{\sqrt{n}}) \\
&= (\phi_{Z_j}(\frac{t}{\sqrt{n}}))^n
\end{aligned}$$

Now, Taylor's theorem gives us that $\phi_{Z_j}(\frac{t}{\sqrt{n}}) = 1 - \frac{t^2}{2n} + o(\frac{t}{\sqrt{n}}^3)$. Therefore

$$\begin{aligned}
\phi_{S_n}(t) &= (1 - \frac{t^2}{2n} + o(\frac{t}{\sqrt{n}}^3))^n \\
&\xrightarrow{n\to\infty} e^{-\frac{t^2}{2}} \\
&= \phi_{\mathcal{N}(0,1)}(t)
\end{aligned}$$

□

We mention without proof a version of the multivariate case of CLT.

Theorem 20: Multivariate WLLN and CLT

Suppose that $X_1, .., X_n$ are iid random vectors in $\mathbb{R}^k$ with mean vector $\mathbb{E}X_i = \mu$ and covariance matrix Σ. Define the centroid random variable $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then

- $\overline{X}_n \xrightarrow{\mathbb{P}} \mu$
- $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$

The central limit theorems tell us that when looking at averages over independent samples, one should expect to see a normal distribution. One very useful application of this is to *noise* in data. Noise is often the result of several corrupting influences on the data acquisition or generation process averaged out. So, *a priori* we ought to expect that noise in our data should follow a Gaussian distribution. We'll have more to say on this when we discuss Bayesian models of inference. Note that the central limit theorem allows us to treat the sample mean as a modification of a standard normally distributed variable viz

$$\overline{X}_n \approx \frac{\sigma}{\sqrt{n}}Z + \frac{\mu}{\sqrt{n}}, \qquad Z \sim \mathcal{N}(0,1)$$

i.e. $\overline{X}_n \approx \mathcal{N}(\frac{\mu}{\sqrt{n}}, \frac{\sigma^2}{n})$.

Exercises

1. Show that the probability of finding at least one success in a time interval of length t of a Poisson-distributed random variable is $1 - e^{-\lambda t}$.

2. Prove that a pdf p_X must satisfy the normalization condition $\int_{\mathbb{R}} p_X(x)dx = 1$.

3. Prove equation (0.0.1)

4. Prove that the components of $X \sim \mathcal{N}(\mu, \Sigma)$, an n-dimensional Gaussian distributed random vector are independent if and only if Σ is diagonal.

5. Prove that the correlation coefficient ρ_{XY} satisfies $|\rho_{XY}| \leq 1$ for all random variables X and Y.

6. Give an example, with proof, of two non-independent, uncorrelated random variables.

7. Show **directly** that $var(aX + bY) = a^2 var(X) + b^2 var(Y) + (2ab)cov(X, Y)$ for all a, b constant and random variables X and Y.

8. Verify that for $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mathbb{E}X = \mu$ and $var(X) = \sigma^2$.

9. Prove the first inequality in Theorem 14.

10. Given a random variable Y, verify that $\sigma(Y)$ as defined by $\{Y^{-1}(B) \mid B \in \mathcal{B}\}$ is in fact a σ-algebra.

11. Prove that for $p \in (0, 1)$ we have $\sum_{k \geq 1} kp(1-p)^{k-1} = \frac{1}{p}$

12. Verify that for $X \sim p_X$ taking outcomes $\{X_1, ..., X_N\}$ with $p_X(X = X_i) = \frac{1}{N}$ we have $H(X) = \log_2 N$.

13. In the section of joint entropy we considered X and Y such that X takes one of 8 possible outcomes with equal likelihood and Y takes one of 4 probabilities with equal probability and upon any observation of the pair we saw that either both X and Y were odd or both were even. Prove that this implies $H(X, Y) = 1$.

14. Prove that $I(X, Y) = 0 \iff X \perp\!\!\!\perp Y$

15. Prove that $I(X, Y) = D_{KL}(p_{XY} \parallel p_X p_Y)$

16. Prove that if X and Y are independent with respect to both p and q then $D_{KL}(p_{XY} \parallel q_{XY}) = D_{KL}(p_X \parallel q_X) + D_{KL}(p_Y \parallel q_Y)$

17. Prove that $\phi(0) = 1$, $\phi(-t) = \bar{\phi}(t)$ and $|\phi(t)| \leq 1$ holds for the characteristic function of any random variable.

18. Calculate the characteristic function of a uniformly distributed random variable $X \sim \mathcal{U}(-a, a)$, $a > 0$.

19. In the proof of the central limit theorem we used the fact that if $X_1, ..., X_n$ are iid then so is $Z_1, ..., Z_n$ for $Z_i = \frac{X_i - \mu}{\sigma}$. Prove this.

Estimation and Inference

Contents

Introduction 44

Elementary Decision Theory 45

Predictors 47

Estimators 48

Bias, Variance, and the Bias-Variance Trade-off 48

Consistency 50

Minimum Variance Estimators 52

A Priori Estimates on Unbiased Estimators 52

Exercises **53**

In this chapter we cover the tools for making inference about distributions given access to values of random variables in a data set. We cover quantitative bounds which control accuracy and confidence in our estimators. These tools are fundamental to making reasonable assertions regarding distributions which we can only access via limited samples of random variables.

Introduction

The purpose of this chapter is to address what is one of the fundamental questions a data scientist can be asked to consider: given the values of numbers in a data set (which we will interpret as *realizations of a random variable*) what can we deduce about the data-generating distribution? Put formally, a **data set** is a collection of numerical values $D = \{x_1, ..., x_n\}$ which we assume is of the form $\{X(\omega_1), X(\omega_2), ..., X(\omega_n)\}$. Of course, things could be even more complicated, and the data could be instead samples from *differently distributed* random variables, namely $\{X(\omega_1), Y(\omega_2), ..., Z(\omega_n)\}$. Based solely on the numbers, we'd like to understand the *nature of the randomness* which they are instantiations of. Broadly speaking, we can go about this along one of two lines.

1. **Parametric:** This is where we will assume a function form of the unknown distribution. The functional form will depend on some parameters which we then want to determine. In general this means we have a density of the form $p(x \mid \theta)$ where theta is (possibly a vector of) parameters of the distribution in question.

2. **Non-Parametric:** This is the case in which no assumptions are made about the functional form of the data-generating distribution. In this case the "data speaks for itself" and guides the inferential process.

Of course the non-parametric view is still parametric. It's simply that in this case the parameterization is in an infinite, rather than a finite, dimensional space.

Our approach will often be based on variants of a **decision function**, $d = d(x_1, ..., x_n)$. This decision function is a tool for dividing up the space of variates and even will help for testing hypotheses. For instance, if we were looking at the final exam of only five students in a class of 100, we might want to know if we could determine whether the class average will be above 60%. In this case $d(x_1, ..., x_5) \in \{0, 1\}$ where 0 represents an answer of "no" and 1 represents the answer "yes". Notice that d divides $\mathbb{R}^5$ into two distinct regions, each of which corresponds to one of the two answers. Those regions, the level sets of d in this instance, are known as **decision regions** for the **decision problem** of determining whether or not the class average is above a 60%. When $d : \mathbb{R}^n \to M$ with $\#M > 2$ we have what is called a **multiple decision problem** involving discrimination among various answers. Perhaps, for example, rather than asking just about the class average I wanted to know about whether the class will have students whose grades fall into each of the regions $[50, 59), [60, 69), [70, 79), ...$ It may also be the case that we are trying to decide on the value of a continuous rather than discrete variable. Namely, we could have $d(x_1, ..., x_n) = \theta$ where θ can take values in a continuous region. Namely, the decision function is simply our *estimate of the parameter θ.* This will be the situation in most of our analysis.

Elementary Decision Theory

We are going to want to penalize incorrect decisions. We do this via a **loss function**

$$\mathcal{L}(\theta, a)$$

where θ represents the true (unknown) parameter we are wanting to estimate and a represents a decision made or action taken. For example, we could have $\mathcal{L}(\theta, a) = |a - \theta|$ or $\mathcal{L}(\theta, a) = ||a - \theta||^2$. The loss depends upon the decision function since we make our decisions a based on the output of decision function. So, in this way, we have

$$\mathcal{L} = \mathcal{L}(a, d(x_1, ..., x_n))$$

But we don't want our penalization, or our decision, to be based just on a single observation set $\{x_1, ..., x_n\}$. Rather, decisions should be based on overall performance. Namely, we have a random variable $d(X_1, ..., X_n)$ leading to the **loss random variable**

$$\mathcal{L}(\theta, d(X_1, ..., X_n))$$

The overall effectiveness of our decision function should be based on how well we *expect* to do, namely we should base our valuation of our model decision on a **risk function**

$$\mathcal{R}(\theta, d) = \mathbb{E}[\mathcal{L}(\theta, d(X_1, ..., X_n))]$$

where, in the above, the expected value is with respect to the joint distribution over $X_1, ..., X_n$, not over θ. Of course, if our sample $\{x_1, ..., x_n\}$ respresents realizations of iid draws of a variable $X \sim p(x \mid \theta)$ then we have

$$\mathcal{R}(\theta, d) = \int_{\mathbb{R}^n} \mathcal{L}(\theta, d(x_1, ..., x_n))\Pi_{i=1}^n p_{X_i}(x_i \mid \theta) dx_1 \cdots dx_n$$

To figure out which of competing decision functions should be used in a given problem we turn to the **minimax decision function**, d^* defined by

$$d^* = \arg\min_{d \in \mathcal{D}}(\max_{\theta} \mathcal{R}(\theta, d))$$

where $\mathcal{D}$ is a space of competing decision functions.

Example 22

We try to estimate parameter $\theta = \lambda$ from a single observation of $X \sim Pois(\theta)$. Namely $p(x \mid \theta) = \frac{e^{-\theta}\theta^x}{x!}$. Given only a single observation we try a linear function $d(x) = cx$ with unknown c. In addition, we will use the quadratic loss function

$$\mathcal{L}(\theta, d) = \frac{(d - \theta)^2}{\theta}$$

Then

$$\begin{aligned}\mathcal{R}(\theta, d) &= \mathbb{E}_X[\mathcal{L}(\theta, d)]\\ &= \sum_x \mathcal{L}(\theta, d(x))p(x \mid \theta)\\ &= \frac{1}{\theta}\sum_x (cx-\theta)^2 \frac{e^{-\theta}\theta^x}{x!}\end{aligned}$$

But $(cx-\theta)^2 = \frac{c^2}{\theta}\{(x-\theta)^2 + 2\theta(1-\frac{1}{c})(x-\theta) + \theta^2(1-\frac{1}{c})^2\}$. Therefore

$$\begin{aligned}\mathcal{R}(\theta, d) &= \frac{c^2}{\theta}(var(X) + 2\theta(1-\frac{1}{c})\mathbb{E}(X-\theta) + \theta^2(1-\frac{1}{c})^2)\\ &= c^2 + \theta(c-1)^2\end{aligned}$$

From this we see that if $c = 1$ then $\mathcal{R}(\theta, d) \equiv 1$ whereas if $c \neq 1$ then $\mathcal{R}$ is unbounded in θ. Therefore $d(x) = x$ is the best minimax decision function estimator of the mean of a Poisson distributed random variable.

The Bayesian view we discussed in the previous chapter essentially comes down to modelling all unknowns as random variables. Therefore we imagine that an unknown parameter θ follows a distribution $\theta \sim p(\theta)$. This Bayesian interpretation of dealing with unknowns leads to the **Bayes Risk function**

$$r(p, d) \doteq \mathbb{E}[\mathcal{R}(\theta, d)] = \int \mathcal{R}(\theta, d)p(\theta)d\theta$$

Notice the difference with the risk function discussed previously. In this instance the Bayes risk averages over all values of the unknown parameter which one expects to encounter provided the parameter is a random variable drawn from the distribution in question. The Bayes risk helps us pick our decision function accordingly; we choose the **Bayes decision function**

$$b = \arg\min_{d \in \mathcal{D}} r(p, d)$$

Example 23

As before we will try to estimate the mean of a Poisson-distributed random variable given a single observation. We assume $\theta \sim e^{-\theta}$ for $\theta > 0$. Then

$$\begin{aligned}r(p, d) &= \int_0^\infty [c^2 + \theta(c-1)^2]e^{-\theta}d\theta\\ &= c^2 + (c-1)^2\end{aligned}$$

Minimization with respect to c gives $c = \frac{1}{2}$. Namely, $d(x) = \frac{x}{2}$ is the Bayes decision function. Notice this is decidedly different than the decision function we obtained earlier using minimax.

Of course, the estimator we obtain implicitly depends on the choice of loss function we use. We've used in the past two examples the very common quadratic loss. This is a popular choice because it allows for the use of convex optimization routines to find the minima. We refer to the risk when using a quadratic loss function as the **MSE** or **mean squared error** of our decision.

Predictors

In doing work on machine learning we'll be trying to *predict* outcomes Y based on inputs X, where both are assumed to be random variables. In this setting we view $Y = f(X)$ and we try to *learn* (in a way made precise later) the function Y which behaves well. Of course, when doing prediction we're still doing decision problems so we have a loss here. In the context of machine learning the loss function is also called a **cost function**. We may use the quadratic cost, say, which takes the form

$$\mathcal{L}(Y, f(X)) = (Y - f(X))^2$$

When evaluating how well our prediction predicts we care about a **expected prediction error** associated to the decision function f, $EPE(f)$, defined by

$$EPE(f) = \mathbb{E}[\mathcal{L}(Y, f(X))] = \int_{\mathbb{R}^{2n}} |y - f(x)|^2 p(x,y) dxdy$$

It's easy to verify that

$$EPE(f) = \mathbb{E}[\mathbb{E}[(Y - f(X))^2 \mid X]]$$

Thus, we can be sure a minima of $EPE(f)$ by selecting the decision funcion

$$f(x) = \arg\min_{c} \mathbb{E}[(Y - c)^2 \mid X = x]$$

which is a pointwise minimizer. A calculation shows this give the **regression function**

$$f(x) = \mathbb{E}[Y \mid X = x]$$

Thus, conditional expectation gives an optimal prediction (the regression function) which is a type of decision function used in making predictions which should generalize well. One example of this is that if we were given a new input on which we wished to predict, we might take the average over nearby y's, via

$$\hat{f}(x) = Ave(y_i \mid x_i \in N_k(x))$$

where $N_k(x)$ is the neighbourhood around x containing the k nearest points to x. This regression function gives rise to a method called k-nearest neighbours, kNN, regression. We note that in kNN described here that

- The expected value operator is replaced with a average on sample data
- Conditioning on x is replaced with conditioning on a region *close* to x

Of course, as the dimension of the x_i's gets large the size of $N_k(x_i)$ must increase by an enormous factor leading to computational issues which we don't address.

Estimators

From now on we're going to be adopting the following conventions.

- We use the word "density" and "pdf" even in cases of discrete random variables
- We wish for the random variable $d = d(X_1, ..., X_n)$ to be estimate a parameter θ (or parameters whenever θ is vector valued) in a distribution $f(x \mid \theta)$ (notice we are now generally using f to denote the density rather than p). We say that d is an **estimator** of θ and write $d(x_1, ..., x_n) = \hat{\theta}$.
- We often use the quadratic loss funciton $\mathcal{L}(\theta, d) = (d - \theta)^2$ and the risk function $\mathcal{R}(\theta, d) = \mathbb{E}\mathcal{L}(\theta, d)$, i.e $MSE = MSE(\hat{\theta})$.

As an easy warm-up, we consider the case where $X \sim \mathcal{N}(\theta, 1)$, i.e. we're trying to estimate the mean of a normally distributed random variable, and here $f(x \mid \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$. Of course, the intuition given to use the by Strong Law of Large Numbers is that we should use

$$d_1(X_1, ..., X_n) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

But we will compare this against $d_2(X_1, ..., X_2) \equiv 1$. If it turns out that $\theta = 1$ then $\mathcal{R}(\theta, d_2) = 0$ whereas

$$\begin{aligned}
\mathcal{R}(\theta, d_1) &= \mathbb{E}[(d_1 - \theta)^2] \\
&= \mathbb{E}[(\overline{X} - \theta)^2] \\
&= var(\overline{X}) \\
&= var(\frac{X_1}{n} + \cdots + \frac{X_n}{n}) \\
&= \frac{nvar(X_1)}{n^2} \\
&= \frac{1}{n}
\end{aligned}$$

In other words, $\mathcal{R}(\theta, d_1) > \mathcal{R}(\theta, d_2)$. In other words, d_2 is better in this instance (although, obviously, not very robust!) This illustrates the need for us to be careful in working out which estimators to use in any instance.

Bias, Variance, and the Bias-Variance Trade-off

We begin with a commonly encountered notion in parameter estimation.

Definition 21: Bias of an Estimator

An estimator $d(X_1, ..., X_n)$ has **bias** given by

$$B_{\hat{\theta}}(\theta) \doteq \mathbb{E}[d(X_1, ..., X_n) - \theta]$$

If $\mathbb{E}\hat{\theta} = \theta$ we say that $\hat{\theta} = d(X_1, ..., X_n)$ is an **unbiased estimator**.

Notice the convention of treating the bias *as a function of the true parameter* θ rather than of the estimate $\hat{\theta}$. The above indicates that an estimator is unbiased if it's "correct on average". Related to the concept of bias is the concept of variance of an estimator.

Definition 22: Variance of an Estimator

An estimator $\hat{\theta} = d(X_1, ..., X_n)$ has **variance** given by

$$V_{\hat{\theta}}(\theta) \doteq \mathbb{E}_\theta(\hat{\theta} - \mathbb{E}\hat{\theta})^2$$

Given these definitions of bias an variance we can look at the MSE (i.e. the risk under quadratic loss function),

$$\begin{aligned}\mathbb{E}_\theta[(\theta - \hat{\theta})^2] &= \mathbb{E}_\theta[((\theta - \mathbb{E}\hat{\theta}) + (\mathbb{E}\hat{\theta} - \hat{\theta}))^2] \\ &= \mathbb{E}_\theta[(\theta - \mathbb{E}\hat{\theta})^2] + 2\mathbb{E}_\theta[(\theta - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \hat{\theta})] + \mathbb{E}_\theta[(\mathbb{E}\hat{\theta} - \hat{\theta})^2]\end{aligned}$$

But notice that $(\theta - \mathbb{E}\hat{\theta})$ is a constant so $2\mathbb{E}_\theta[(\theta - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \hat{\theta})] = 2(\theta - \mathbb{E}\hat{\theta})\mathbb{E}_\theta(\mathbb{E}\hat{\theta} - \hat{\theta}) = 2(\theta - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \mathbb{E}\hat{\theta}) = 0$. Therefore we have

$$\begin{aligned}MSE(\hat{\theta}) &= \mathbb{E}_\theta[(\theta - \mathbb{E}\hat{\theta})^2] + \mathbb{E}_\theta[(\mathbb{E}\hat{\theta} - \hat{\theta})^2] \\ &= (\theta - \mathbb{E}\hat{\theta})^2 + \mathbb{E}_\theta[(\mathbb{E}\hat{\theta} - \hat{\theta})^2] \\ &= B^2_{\hat{\theta}}(\theta) + V_{\hat{\theta}}(\theta)\end{aligned}$$

In other words

$$MSE = (bias)^2 + variance$$

Which is known in the field as the **bias-variance trade-off**. This means that, when using the MSE to score a predictor, one can improve by trying to lower the bias or lower the variance. A corollary of the bias-variance trade-off is that when using unbiased estimators, one can only improve by reducing the variance. Various values of MSE may find

- Regions of low variance but high bias. This indicates a simple model, since the estimator cannot vary enough to the complexity in the samples.
- Regions of low bias but high variance. This indicates a high amount of model complexity, since the estimator may be able to adapt to fluctuations and noise in the samples.
- Regions with a bit of variance and bias. In these regions generally the MSE will be lowest and there is enough complexity to model the randomness but not enough to get stuck modelling the noise in the samples.

We remark that the bias and variance both depend on the unknown parameter θ.

Example 24

Consider $X_i \sim B(p)$, iid coin flips. Denote by n the number of heads obtained in N tosses. Then

$$\begin{aligned}\mathbb{E}[\frac{n}{N}] &= \frac{1}{N}\mathbb{E}[\text{binomial variable } B(N,p)] \\ &= \frac{1}{N}Np \\ &= p\end{aligned}$$

Thus, the random variable $\frac{n}{N}$ is an unbiased estimator of the parameter p. Similarly

$$\begin{aligned}V_{\hat{\theta}}(\theta) &= var_{B(N,p)}(\frac{n}{N}) \\ &= \frac{1}{N^2}var_{B(N,p)}(n) \\ &= \frac{Np(1-p)}{N^2} \\ &= \frac{p(1-p)}{N}\end{aligned}$$

Thus, the variance of the estimator decreases in N, meaning that the estimate becomes more and more reliable the larger the number of draws taken.

One common heuristic for selecting estimators, beyond the minimax or Bayes, is to select among unbiased estimators. Thus, we may be looking to select *optimal unbiased estimators*, where optimality will be made precise in given contexts.

Consistency

By **consistency** of an estimator $\hat{\theta}_n = d(x_1, ..., x_n)$ we mean that

$$\hat{\theta}_n \xrightarrow[n\to\infty]{a.s} \theta$$

So, for example, the Strong Law of Large Numbers implies that $\overline{X}_n$, the sample mean, is a consistent estimator of the true mean, $\mathbb{E}[X]$, i.e.

$$\overline{X}_n \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[X]$$

We say that the sample mean is *consistent and unbiased*. On the other hand, the **sample variance**

$$\hat{\sigma}^2 \doteq \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is *biased and consistent*. To see this, observe that

$$\begin{aligned} n\hat{\sigma}^2 &= \sum_i (X_i - \overline{X})^2 \\ &= \sum_i X_i^2 - \frac{2}{n}\sum_{i,j} X_i X_j + \frac{n}{n^2}\sum_{j,k} X_j X_k \\ &= \sum_i X_i^2 - \frac{1}{n}\sum_{i,j} X_i X_j \end{aligned}$$

Which means that $n\mathbb{E}\hat{\sigma}^2 = \sum_i \mathbb{E}X_i^2 - \frac{1}{n}\sum_{i,j}\mathbb{E}[X_i X_j]$. Now, in the second sum, when $i = j$ the summands become terms like $\mathbb{E}[X_i^2] = \mu_2$, the **second moment** of random variable X.[11] When $i \neq j$ we obtain terms like $\mathbb{E}[X_i X_j] = \mathbb{E}X_i \mathbb{E}X_j = \mu^2$. Then

$$\begin{aligned} n\mathbb{E}\hat{\sigma}^2 &= n\mu_2 - \frac{1}{n}\{n\mu_2 + n(n-1)\mu^2\} \\ &= (n-1)\{\mu_2 - \mu^2\} \\ &= (n-1)\sigma^2 \end{aligned}$$

From this we see that the bias of the sample variance is

$$B_{\hat{\sigma}^2}(\sigma^2) = -\frac{\sigma^2}{n}$$

Notice that this implies

$$\mathbb{E}[\frac{n}{n-1}\hat{\sigma}^2] = \sigma^2$$

so that $\frac{n}{n-1}\hat{\sigma}^2$ is an **unbiased sample variance** which can be written

$$\hat{\sigma}^2_{unbiased} = \hat{\sigma}^2_u = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$$

One should use caution in relying on software packages to calculate the sample variance since some calculate the unbiased version while others calculate the biased version. In other words, there is no universal agreement as to which should refer to “the sample variance". Despite it being biased, the sample variance is clearly consistent, since

$$\hat{\sigma}^2_n = \overline{X^2} - \overline{X}^2$$

so the Strong Law of Large Numbers dictates[12] that

$$\hat{\sigma}_n \xrightarrow[n\to\infty]{a.s.} \mathbb{E}X^2 - (\mathbb{E}X)^2 = \sigma^2$$

The sample variance is therefore **biased and consistent** or, put differently, wrong on average but right eventually.

[11] The **n’th moment of** X is $\mathbb{E}X^n$, so, e.g. the first moment is the mean. These are sometimes called the **moments about zero**

[12] Here we are using the fact that if $X_n \xrightarrow[n\to\infty]{a.s.} X$ the $g(X_n) \xrightarrow[n\to\infty]{a.s.} g(X)$ for continuous g.

Minimum Variance Estimators

An unbiased estimator possessing least variance among the class of unbiased estimators should be deemed “best" since it will then minimize the MSE. This ideal estimator is called the MVU or **minimum variance unbiased** estimator. Notice our **precision** is inversely proportional to the spread in our estimator and therefore is

$$\frac{1}{V_{\hat{\theta}}(\theta)}$$

since variance represents a spread in certainty. In this way, high variance means low precision and low variance means high precision. Unbiased estimators therefore give us a benchmark from which to compare among estimators. Suppose that $d^* = MVU$ for a parameter θ and let d be any other estimator for θ. The ratio

$$Eff(\theta, d) \doteq \frac{V_{d^*}(\theta)}{V_d(\theta)}$$

is called the **efficiency** of the estimator d. If $\frac{V_{d^*}(\theta)}{V_d(\theta)}$ were, say, .83 then we would say “d is 83% efficient". In a sample of size n, we suppose that an unbiased estimator was of the form

$$V(d) = \frac{c}{n}$$

as is common in practice. Then if d were 80% efficient we'd have $V(d^*) = \frac{.8c}{n}$. This would mean that *a sample of size 80 with d^* is as good as a sample of size 100 with d*. In other words the efficiency $Eff(\theta, d)$ helps to determine the optimal sample size for a given level of precision. In this way, efficiency is a measure of “bang for your buck" where bucks are measured in units of data. One issue with the MVU however, is that we don't know the true value θ, or even whether or not an MVU exists. This issue we address in the following section.

A Priori Estimates on Unbiased Estimators

As noted in the last section, we often don't know whether an MVU even exists, let alone how to construct one. To circumvent this, we'd like to get some a priori estimates on the variance of a given unbiased estimator. If we could come up with a universal a priori lower bound on unbiased estimators then any estimator which saturated the bound would clearly be worthy of the title “best". To attack this problem we introduce a few necessary terms.

- Given iid data $\{x_1, ..., x_n\}$ assumed draws from $X_i \sim f(x \mid \theta)$, we define the **likelihood function**

  $$L(\theta) \doteq \Pi_{i=1}^n f(x_i \mid \theta)$$

  This is the *probability of observing the sample for a given choice of θ.*

- Since the likelihood is a product it can often be easier to works with it's logarithm. We define the **log-likelihood function**

$$l(\theta) = \log L(\theta)$$

- The **Fisher score** is given by

$$s(\theta) = l'(\theta)$$

  or, in the case of vector parameters $s(\theta) = \nabla_\theta l(\theta)$. It can tell the sensitivity of the sample to changes in the parameters. In other words the Fisher score is a

- The **Fisher Information Measure** $\mathcal{F}_X$ is

$$\mathcal{F}_X(\theta) = \mathbb{E}_\theta[(\nabla_\theta f(x \mid \theta))^2]$$

  In the case of scalar parameters we have

$$\mathcal{F}_X(\theta) = \mathbb{E}[l'(\theta)]$$

  For a single sample, this is also often denoted by $I(\theta)$. For iid samples we have then that $\mathcal{F}_X(\theta) = nI(\theta)$.

Exercises

1. Prove that $\mathbb{E}[Y \mid X = x] = \arg\min_c \mathbb{E}[(Y - c)^2 \mid X = x]$.

Bibliography

[1] Patrick Billingsley. *Probability and Measure*. Wiley.

[2] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 4 edition, 2010.

[3] William Feller. *An Introduction to Probability Theory and its Applications*, volume 1. Wiley, 1968.

[4] Solomon Kullback. *Information Theory and Statistics*. Dover Publications Inc., 1978.

[5] David Williams. *Probability with Martingales*. Cambridge University Press.

[6] Elizier Yudkowsky. `https://wiki.lesswrong.com/wiki/A_Human's_Guide_to_Words`.