DEVLIN · GUO · KUNIN · XIANG

SEEING THEORY

www.dbooks.org

Copyright © 2018 Seeing Theory

Contents

Basic Probability 5

Compound Probability 19

Probability Distributions 31

Frequentist Inference 41

Bayesian Inference 49

Regression Analysis 55

www.dbooks.org

Basic Probability

Introduction

The world is an uncertain place. Making predictions about something as seemingly mundane as tomorrow's weather, for example, is actually quite a difficult task. Even with the most advanced computers and models of the modern era, weather forecasters still cannot say with absolute certainty whether it will rain tomorrow. The best they can do is to report their best estimate of the *chance* that it will rain tomorrow. For example, if the forecasters are fairly confident that it will rain tomorrow, they might say that there is a 90% chance of rain. You have probably heard statements like this your entire life, but have you ever asked yourself what exactly it means to say that there is a 90% chance of rain?

Let us consider an even more basic example: tossing a coin. If the coin is fair, then it is just as likely to come up heads as it is to come up tails. In other words, if we were to repeatedly toss the coin many times, we would expect about about half of the tosses to be heads and and half to be tails. In this case, we say that the **probability** of getting a head is 1/2 or 0.5.

Note that when we say the probability of a head is 1/2, we are *not* claiming that any sequence of coin tosses will consist of exactly 50% heads. If we toss a fair coin ten times, it would not be surprising to observe 6 heads and 4 tails, or even 3 heads and 7 tails. But as we continue to toss the coin over and over again, we expect the long-run frequency of heads to get ever closer to 50%. In general, it is important in statistics to understand the distinction between *theoretical* and *empirical* quantities. Here, the true (theoretical) probability of a head was 1/2, but any realized (empirical) sequence of coin tosses may have more or less than exactly 50% heads. (See Figures 1 – 3.)

Now suppose instead that we were to toss an unusual coin with heads on both of its faces. Then every time we flip this coin we will observe a head — we say that the probability of a head is 1. The probability of a tail, on the other hand, is 0. Note that there is no way

Figure 1: The true probability of a head is 1/2 for a fair coin.

Figure 2: A sequence of 10 flips happened to contain 3 head. The empirical frequency of heads is thus 3/10, which is quite different from 1/2.

Figure 3: A sequence of 100 flips happened to contain 45 heads. The empirical frequency of heads is 45/100, which is much closer to 1/2.

www.dbooks.org

we can further modify the coin to make flipping a head even more likely. Thus, *a probability is always a number between 0 and 1 inclusive.*

First Concepts

Terminology

When we later discuss examples that are more complicated than flipping a coin, it will be useful to have an established vocabulary for working with probabilities. A probabilistic **experiment** (such as tossing a coin or rolling a die) has several components. The **sample space** is the set of all possible **outcomes** in the experiment. We usually denote the sample space by Ω, the Greek capital letter "Omega." So in a coin toss experiment, the sample space is

$$\Omega = \{\mathsf{H}, \mathsf{T}\},$$

since there are only two possible outcomes: heads (H) or tails (T). Different experiments have different sample spaces. So if we instead consider an experiment in which we roll a standard six-sided die, the sample space is

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Collections of outcomes in the sample space Ω are called **events**, and we often use capital Roman letters to denote these collections. We might be interested in the event that we roll an even number, for example. If we call this event E, then

$$E = \{2, 4, 6\}.$$

Any subset of Ω is a valid event. In particular, one-element subsets are allowed, so we can speak of the event F of rolling a 4, $F = \{4\}$.

Assigning probabilities to dice rolls and coin flips

In a random experiment, every event gets assigned a probability. Notationally, if A is some event of interest, then $P(A)$ is the probability that A occurs. The probabilities in an experiment are not arbitrary; they must satisfy a set of rules or **axioms**. We first require that *all probabilities be nonnegative.* In other words, in an experiment with sample space Ω, it must be the case that

$$P(A) \geq 0 \tag{1}$$

for any event $A \subseteq \Omega$. This should make sense given that we've already said that a probability of 0 is assigned to an impossible event,

and there is no way for something to be less likely than something that is impossible!

The next axiom is that *the sum of the probabilities of all the outcomes in Ω must be 1.* We can restate this requirement by the equation

$$\sum_{\omega \in \Omega} P(\omega) = 1. \tag{2}$$

This rule can sometimes be used to deduce the probability of an outcome in certain experiments. Consider an experiment in which we roll a fair die, for example. Then each outcome (i.e. each face of the die) is equally likely. That is,

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = a,$$

for some number a. Equation (2) now allows us to conclude

$$1 = \sum_{k=1}^{6} P(k) = \sum_{k=1}^{6} a = 6a,$$

so $a = 1/6$. In this example, we were able to use the symmetry of the experiment along with one of the probability axioms to determine the probability of rolling any number.

Once we know the probabilties of the outcomes in an experiment, we can compute the probability of any event. This is because *the probability of an event is the sum of the probabilities of the outcomes it comprises.* In other words, for an event $A \subseteq \Omega$, the probability of A is

$$P(A) = \sum_{\omega \in A} P(\omega). \tag{3}$$

To illustrate this equation, let us find the probability of rolling an even number, an event which we will denote by E. Since $E = \{2, 4, 6\}$, we simply add the probabilities of these three outcomes to obtain

$$\begin{aligned} P(E) &= \sum_{\omega \in E} P(\omega) \\ &= P(2) + P(4) + P(6) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\ &= \frac{1}{2}. \end{aligned}$$

What is the probability that we get at least one H?

Solution. One way to solve this problem is to add up the probabilities

www.dbooks.org

of all outcomes that have at least one H. We would get

$$\begin{aligned} P(\text{flip at least one H}) &= P(\text{HH}) + P(\text{HT}) + P(\text{TH}) \\ &= p^2 + p \cdot (1-p) + (1-p) \cdot p \\ &= p^2 + 2 \cdot (p - p^2) \\ &= 2p - p^2 \\ &= p \cdot (2-p). \end{aligned}$$

Another way to do this is to find the probability that we **don't** flip at least one H, and subtract that probability from 1. This would give us the probability that we **do** flip at least one H.

The only outcome in which we don't flip at least one H is if we flip T both times. We would then compute

$$P(\text{don't flip at least one H}) = P(\text{TT}) = (1-p)^2$$

Then to get the **complement** of this event, i.e. the event where we **do** flip at least one H, we subtract the above probability from 1. This gives us

$$\begin{aligned} P(\text{flip at least one H}) &= 1 - P(\text{don't flip at least one H}) \\ &= 1 - (1-p)^2 \\ &= 1 - (1 - 2p + p^2) \\ &= 2p - p^2 \\ &= p \cdot (2-p). \end{aligned}$$

Wowee! Both methods for solving this problem gave the same answer. Notice that in the second calculation, we had to sum up fewer probabilities to get the answer. It can often be the case that computing the probability of the complement of an event and subtracting that from 1 to find the probability of the original event requires less work. □

Independence

If two events A and B don't influence or give any information about the other, we say A and B are independent. Remember that this is not the same as saying A and B are disjoint. If A and B were disjoint, then given information that A happened, we would know with certainty that B did *not* happen. Hence if A and B are disjoint they could never be independent. The mathematical statement of independent events is given below.

Definition 0.0.1. *Let A and B both be subsets of our sample space Ω. Then we say A and B are independent if*

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

In other words, if the probability of the intersection factors into the product of the probabilities of the individual events, they are independent.

We haven't defined set intersection in this section, but it is defined in the set theory chapter. The $\cap$ symbol represents *A and B* happening, i.e. the intersection of the events.

Example 0.0.1. *Returning to our double coin flip example, our sample space was*

$$\Omega = \{HH, HT, TH, TT\}$$

Define the events

$$A \doteq \{\textit{first flip heads}\} = \{HH, HT\}$$
$$B \doteq \{\textit{second flip heads}\} = \{HT, TT\}$$

Notation: *We write the sign $\doteq$ to represent that we are defining something. In the above expression, we are defining the arbitrary symbols A and B to represent events.*

Intuitively, we suspect that A and B are independent events, since the first flip has no effect on the outcome of the second flip. This intuition aligns with the definition given above, as

$$P(A \cap B) = P(\{HT\}) = \frac{1}{4}$$

and

$$P(A) = P(B) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

We can verify that

$$P(A \cap B) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A)P(B)$$

Hence A and B are independent. This may have seemed like a silly exercise, but in later chapters, we will encounter pairs of sets where it is not intuitively clear whether or not they are independent. In these cases, we can simply verify this mathematical definition to conclude independence.

www.dbooks.org

Expectation

Consider the outcome of a single die roll, and call it X. A reasonable question one might ask is "What is the average value of X?". We define this notion of "average" as a weighted sum of outcomes.

Since X can take on 6 values, each with probability $\frac{1}{6}$, the weighted average of these outcomes should be

$$\begin{aligned}\text{Weighted Average} &= \frac{1}{6}\cdot 1 + \frac{1}{6}\cdot 2 + \frac{1}{6}\cdot 3 + \frac{1}{6}\cdot 4 + \frac{1}{6}\cdot 5 + \frac{1}{6}\cdot 6 \\ &= \frac{1}{6}\cdot(1+2+3+4+5+6) \\ &= \frac{21}{6} \\ &= 3.5\end{aligned}$$

This may seem dubious to some. How can the average roll be a non-integer value? The confusion lies in the interpretation of the phrase *average roll*. A more correct interpretation would be the long term average of the die rolls. Suppose we rolled the die many times, and recorded each roll. Then we took the average of all those rolls. This average would be the fraction of 1's, times 1, plus the fraction of 2's, times 2, plus the fraction of 3's, times 3, and so on. But this is exactly the computation we have done above! In the long run, the fraction of each of these outcomes is nothing but their probability, in this case, $\frac{1}{6}$ for each of the 6 outcomes.

From this very specific die rolling example, we can abstract the notion of the *average value* of a random quantity. The concept of average value is an important one in statistics, so much so that it even gets a special bold faced name. Below is the mathematical definition for the **expectation**, or average value, of a random quantity X.

Definition 0.0.2. *The **expected value**, or **expectation** of X, denoted by $E(X)$, is defined to be*

$$E(X) = \sum_{x \in X(\Omega)} xP(X = x)$$

This expression may look intimidating, but it is actually conveying a very simple set of instructions, the same ones we followed to compute the average value of X.

The $\sum$ sign means to sum over, and the indices of the items we are summing are denoted below the $\sum$ sign. The $\in$ symbol is shorthand for "contained in", so the expression below the $\sum$ is telling us to sum over all items *contained in* our sample space Ω. We can think of the expression to the right of the $\sum$ sign as the actual items we are summing, in this case, the weighted contribution of each item in our sample space.

The notation $X(\Omega)$ is used to deal with the fact that Ω may not be a set of numbers, so a weighted sum of elements in Ω isn't even well defined. For instance, in the case of a coin flip, how can we compute $\mathtt{H} \cdot \frac{1}{2} + \mathtt{T} \cdot \frac{1}{2}$? We would first need to assign *numerical values* to $\mathtt{H}$ and $\mathtt{T}$ in order to compute a meaningful expected value. For a coin flip we typically make the following assignments,

$$\mathtt{T} \mapsto 0$$
$$\mathtt{H} \mapsto 1$$

So when computing an expectation, the indices that we would sum over are contained in the set

$$X(\Omega) = \{0, 1\}$$

Let's use this set of instructions to compute the expected value for a coin flip.

Expectation of a Coin Flip

Now let X denote the value of a coin flip with bias p. That is, with probability p we flip $\mathtt{H}$, and in this case we say $X = 1$. Similarly, with probability $1 - p$ we flip $\mathtt{T}$, and in this case we say $X = 0$. The expected value of the random quantity X is then

$$\begin{aligned} E(X) &= \sum_{x \in X(\Omega)} xP(X = x) \\ &= \sum_{x \in \{0,1\}} xP(X = x) \\ &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \\ &= 0 \cdot P(\mathtt{T}) + 1 \cdot P(\mathtt{H}) \\ &= 0 \cdot (1 - p) + 1 \cdot p \\ &= p \end{aligned}$$

So the expected value of this experiment is p. If we were flipping a fair coin, then $p = \frac{1}{2}$, so the average value of X would be $\frac{1}{2}$.

Again, we can never get an outcome that would yield $X = \frac{1}{2}$, but this is not the interpretation of the expectation of X. Remember, the correct interpretation is to consider what would happen if we flipped the coin many times, obtained a sequence of 0's and 1's, and took the average of those values. We would expect around half of the flips to give 0 and the other half to give 1, giving an average value of $\frac{1}{2}$.

Exercise 0.0.1. *Show the following properties of expectation.*

(a) If X and Y are two random variables, then

$$E(X + Y) = E(X) + E(Y)$$

www.dbooks.org

(b) If X is a random variable and c is a constant, then

$$E(cX) = cE(X)$$

(c) If X and Y are independent random variables, then

$$E[XY] = E[X]E[Y]$$

Proof. For now, we will take (a) and (c) as a fact, since we don't know enough to prove them yet (and we haven't even defined independence of random variables!). (b) follows directly from the definition of expectation given above. □

Variance

The variance of a random variable X is a nonnegative number that summarizes on average how much X differs from its mean, or expectation. The first expression that comes to mind is

$$X - E(X)$$

i.e. the difference between X and its mean. This itself is a random variable, since even though EX is just a number, X is still random. Hence we would need to take an expectation to turn this expression into the average amount by which X differs from its expected value. This leads us to

$$E(X - EX)$$

This is almost the definition for variance. We require that the variance always be nonnegative, so the expression inside the expectation should always be ≥ 0. Instead of taking the expectation of the difference, we take the expectation of the squared difference.

Definition 0.0.3. *The **variance** of X, denoted by $Var(X)$ is defined*

$$Var(X) = E[(X - EX)^2]$$

Below we give and prove some useful properties of the variance.

Proposition 0.0.1. *If X is a random variable with mean EX and $c \in$ is a real number,*

(a) $Var(X) \geq 0$.

(b) $Var(cX) = c^2 Var(X)$.

(c) $Var(X) = E(X^2) - E(X)$.

(d) If X and Y are independent random variables, then

$$Var(X+Y) = Var(X) + Var(Y)$$

Proof.

(a) Since $(X - EX)^2 \geq 0$, its average is also ≥ 0. Hence $E[(X - EX)^2] \geq 0$.

(b) Going by the definition, we have

$$\begin{aligned}
\text{Var}(cX) &= E[(cX - E[cX])^2] \\
&= E[(cX - cEX)^2] \\
&= E[c^2(X - EX)^2] \\
&= c^2 E[(X - EX)^2] \\
&= c^2 \text{Var}(X)
\end{aligned}$$

(c) Expanding out the square in the definition of variance gives

$$\begin{aligned}
\text{Var}(X) &= E[(X - EX)^2] \\
&= E[X^2 - 2XEX + (EX)^2] \\
&= E[X^2] - E(2XEX) + E((EX)^2) \\
&= E[X^2] - 2EXEX + (EX)^2 \\
&= E[X^2] - (EX)^2
\end{aligned}$$

where the third equality comes from linearity of E (Exercise 2.3 (a)) and the fourth equality comes from Exercise 2.3 (b) and the fact that since EX and $(EX)^2$ are constants, their expectations are just EX and $(EX)^2$ respectively.

(d) By the definition of variance,

$$\begin{aligned}
\text{Var}(X+Y) &= E[(X+Y)^2] - (E[X+Y])^2 \\
&= E[X^2 + 2XY + Y^2] - \left((E[X])^2 + 2E[X]E[Y] + (E[Y])^2\right) \\
&= E[X^2] - (E[X])^2 + E[Y^2] - (E[Y]^2) + 2E[XY] - 2E[X]E[Y] \\
&= E[X^2] - (E[X])^2 + E[Y^2] - (E[Y]^2) \\
&= \text{Var}(X) + \text{Var}(Y)
\end{aligned}$$

where the fourth equality comes from the fact that if X and Y are independent, then $E[XY] = E[X]E[Y]$. Independence of random variables will be discussed in the "Random Variables" section, so don't worry if this proof doesn't make any sense to you yet.

□

www.dbooks.org

Exercise 0.0.2. *Compute the variance of a die roll, i.e. a uniform random variable over the sample space* $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Solution. Let X denote the outcome of the die roll. By definition, the variance is

$$\begin{aligned}
\text{Var}(X) &= E[(X - EX)]^2 \\
&= E(X^2) - (EX)^2 && \text{(Proposition 2.11 (c))} \\
&= \Big(\sum_{k=1}^{6} k^2 \cdot \frac{1}{6} \Big) - (3.5)^2 && \text{(Definition of Expectation)} \\
&= \frac{1}{6} \cdot (1 + 4 + 9 + 16 + 25 + 36) - 3.5^2 \\
&= \frac{1}{6} \cdot 91 - 3.5^2 \\
&\approx 2.92
\end{aligned}$$

□

Remark 0.0.1. *The square root of the variance is called the* ***standard deviation****.*

Markov's Inequality

Here we introduce an inequality that will be useful to us in the next section. Feel free to skip this section and return to it when you read "Chebyschev's inequality" and don't know what's going on.

Markov's inequality is a bound on the probability that a nonnegative random variable X exceeds some number a.

Theorem 0.0.1 (Markov's inequality)**.** *Suppose X is a nonnegative random variable and $a \in$ is a positive constant. Then*

$$P(X \geq a) \leq \frac{EX}{a}$$

Proof. By definition of expectation, we have

$$\begin{aligned}
EX &= \sum_{k \in X(\Omega)} kP(X = k) \\
&= \sum_{k \in X(\Omega) \text{ s.t. } k \geq a} kP(X = k) + \sum_{k \in X(\Omega) \text{ s.t. } k < a} kP(X = k) \\
&\geq \sum_{k \in X(\Omega) \text{ s.t. } k \geq a} kP(X = k) \\
&\geq \sum_{k \in X(\Omega) \text{ s.t. } k \geq a} aP(X = k) \\
&= a \sum_{k \in X(\Omega) \text{ s.t. } k \geq a} P(X = k) \\
&= aP(X \geq a)
\end{aligned}$$

where the first inequality follows from the fact that X is nonnegative and probabilities are nonnegative, and the second inequality follows from the fact that $k \geq a$ over the set $\{k \in X(\Omega) \text{ s.t. } k \geq a\}$.

Notation: "s.t." stands for "such that".

Dividing both sides by a, we recover

$$P(X \geq a) \leq \frac{EX}{a}$$

□

Corollary 0.0.1 (Chebyschev's inequality)**.** *Let X be a random variable. Then*

$$P(|X - EX| > \epsilon) \leq \frac{Var(X)}{\epsilon^2}$$

Proof. This is marked as a corollary because we simply apply Markov's inequality to the nonnegative random variable $(X - EX)^2$. We then have

$$\begin{aligned} P(|X - EX| > \epsilon) &= P((X - EX)^2 > \epsilon^2) && \text{(statements are equivalent)} \\ &\leq \frac{E[(X - EX)^2]}{\epsilon^2} && \text{(Markov's inequality)} \\ &= \frac{\text{Var}(X)}{\epsilon^2} && \text{(definition of variance)} \end{aligned}$$

□

www.dbooks.org

Estimation

One of the main reasons we do statistics is to make inferences about a population given data from a subset of that population. For example, suppose there are two candidates running for office. We could be interested in finding out the true proportion of the population that supports a particular political candidate. Instead of asking every single person in the country their preferred candidate, we could randomly select a couple thousand people from across the country and record their preference. We could then estimate the true proportion of the population that supports the candidate using this sample proportion. Since each person can only prefer one of two candidates, we can model this person's preference as a coin flip with bias $p =$ the true proportion that favors candidate 1.

Estimating the Bias of a Coin

Suppose now that we are again flipping a coin, this time with bias p. In other words, our coin can be thought of as a random quantity X defined

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

where 1 represents `H` and 0 represents `T`. If we were just handed this coin, and told that it has some bias $0 \leq p \leq 1$, how would we estimate p? One way would be to flip the coin n times, count the number of heads we flipped, and divide that number by n. Letting X_i be the outcome of the i^{th} flip, our estimate, denoted $\hat{p}$, would be

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

As the number of samples n gets bigger, we would expect $\hat{p}$ to get closer and closer to the true value of p.

Estimating π

In the website's visualization, we are throwing darts uniformly at a square, and inside that square is a circle. If the side length of the square that inscribes the circle is L, then the radius of the circle is $R = \frac{L}{2}$, and its area is $A = \pi(\frac{L}{2})^2$. At the i^{th} dart throw, we can define

$$X_i = \begin{cases} 1 & \text{if the dart lands in the circle} \\ 0 & \text{otherwise} \end{cases}$$

The event “dart lands in the circle" has probability

$$p = \frac{\text{Area of Circle}}{\text{Area of Square}} = \frac{\pi\left(\frac{L}{2}\right)^2}{L^2} = \frac{\pi}{4}$$

So with probability $p = \frac{\pi}{4}$, a dart lands in the circle, and with probability $1 - \frac{\pi}{4}$, it doesn't.

By the previous section, we can estimate p using $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ so that for large enough n, we have

$$\hat{p} \approx p = \frac{\pi}{4}$$

so that rearranging for π yields

$$\pi \approx 4\hat{p}$$

Hence our approximation gets closer and closer to π as the number of sample $n \to \infty$ causing $\hat{p} \to p$.

Consistency of Estimators

What exactly do we mean by “closer and closer"? In this section, we describe the concept of **consistency** in order to make precise this notion of convergence. Our estimator in the last section, $4\hat{p}$ is itself random, since it depends on the n sample points we used to compute it. If we were to take a different set of n sample points, we would likely get a different estimate. Despite this randomness, intuitively we believe that as the number of samples n tends to infinity, the estimator $4\hat{p}$ will converge in some probabilistic sense, to π.

Another way to formulate this is to say, no matter how small a number we pick, say 0.001, we should always be able to conclude that the probability that our estimate differs from π by more than 0.001, goes to 0 as the number of samples goes to infinity. We chose 0.001 in this example, but this notion of probabilistic convergence should hold for any positive number, no matter how small. This leads us to the following definition.

Definition 0.0.4. *We say an estimator $\hat{p}$ is a **consistent** estimator of p if for any $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|\hat{p} - p| > \epsilon) = 0.$$

Let's show that $4\hat{p}$ is a *consistent* estimator of π.

www.dbooks.org

Proof. Choose any $\epsilon > 0$. By Chebyshev's inequality (Corollary 2.13),

$$
\begin{aligned}
P(|4\hat{p} - \pi| > \epsilon) &\leq \frac{\text{Var}(4\hat{p})}{\epsilon^2} \\
&= \frac{\text{Var}\left(4 \cdot \frac{1}{n}\sum_{i=1}^{n} X_i\right)}{\epsilon^2} && \text{(Definition of } \hat{p}) \\
&= \frac{\frac{16}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right)}{\epsilon^2} && (\text{Var}(cY) = c^2\text{Var}(Y)) \\
&= \frac{\frac{16}{n^2}\sum_{i=1}^{n}\text{Var}(X_i)}{\epsilon^2} && (X_i\text{'s are independent}) \\
&= \frac{\frac{16}{n^2} \cdot n \cdot \text{Var}(X_1)}{\epsilon^2} && (X_i\text{'s are identically distributed}) \\
&= \frac{\frac{16}{n} \cdot p(1-p)}{\epsilon^2} && (\text{Var}(X_i) = p(1-p)) \\
&= \frac{16 \cdot \frac{\pi}{4}\left(1 - \frac{\pi}{4}\right)}{n\epsilon^2} && \left(p = \frac{\pi}{4}\right) \\
&\to 0
\end{aligned}
$$

as $n \to \infty$. Hence we have shown that $4\hat{p}$ is a consistent estimator of π. □

Compound Probability

Set Theory

A probability measure P is a function that maps subsets of the state space Ω to numbers in the interval $[0, 1]$. In order to study these functions, we need to know some basic set theory.

Basic Definitions

Definition 0.0.5. *A **set** is a collection of items, or elements, with no repeats. Usually we write a set A using curly brackets and commas to distinguish elements, shown below*

$$A = \{a_0, a_1, a_2\}$$

*In this case, A is a set with three distinct elements: a_0, a_1, and a_2. The size of the set A is denoted $|A|$ and is called the **cardinality** of A. In this case, $|A| = 3$. The **empty set** is denoted $\varnothing$ and means*

$$\varnothing = \{\quad\}$$

Some essential set operations in probability are the intersection, union, and complement operators, denoted $\cap, \cup$, and ^{c}. They are defined below

Definition 0.0.6. ***Intersection** and **Union** each take two sets in as input, and output a single set. **Complementation** takes a single set in as input and outputs a single set. If A and B are subsets of our sample space Ω, then we write*

(a) $A \cap B = \{x \in \Omega : x \in A \text{ and } x \in B\}$.

(b) $A \cup B = \{x \in \Omega : x \in A \text{ or } x \in B\}$.

(c) $A^c = \{x \in \Omega : x \notin A\}$.

Another concept that we need to be familiar with is that of disjointness. For two sets to be disjoint, they must share no common elements, i.e. their intersection is empty.

www.dbooks.org

Definition 0.0.7. *We say two sets A and B are **disjoint** if*

$$A \cap B = \varnothing$$

It turns out that if two sets A and B are disjoint, then we can write the probability of their union as

$$P(A \cup B) = P(A) + P(B)$$

Set Algebra

There is a neat analogy between set algebra and regular algebra. Roughly speaking, when manipulating expressions of sets and set operations, we can see that " $\cup$ " acts like " $+$ " and " $\cap$ " acts like " $\times$ ". Taking the complement of a set corresponds to taking the negative of a number. This analogy isn't perfect, however. If we considered the union of a set A and its complement A^c, the analogy would imply that $A \cup A^c = \varnothing$, since a number plus its negative is 0. However, it is easily verified that $A \cup A^c = \Omega$ (Every element of the sample space is either in A or not in A.)

Although the analogy isn't perfect, it can still be used as a rule of thumb for manipulating expressions like $A \cap (B \cup C)$. The number expression analogy to this set expression is $a \times (b + c)$. Hence we could write it

$$a \times (b + c) = a \times b + a \times c$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

The second set equality is true. Remember that what we just did was not a proof, but rather a non-rigorous rule of thumb to keep in mind. We still need to actually prove this expression.

Exercise 0.0.3. *Show that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.*

Proof. To show set equality, we can show that the sets are contained in each other. This is usually done in two steps.

Step 1: "$\subset$". First we will show that $A \cap (B \cup C) \subset (A \cap B) \cup (A \cap C)$.

Select an arbitrary element in $A \cap (B \cup C)$, denoted ω. Then by definition of intersection, $\omega \in A$ and $\omega \in (B \cup C)$. By definition of union, $\omega \in (B \cup C)$ means that $\omega \in B$ or $\omega \in C$. If $\omega \in B$, then since ω is also in A, we must have $\omega \in A \cap B$. If $\omega \in C$, then since ω is also in A, we must have $\omega \in A \cap C$. Thus we must have either

$$\omega \in A \cap B \text{ or } \omega \in A \cap C$$

Hence, $\omega \in (A \cap B) \cup (A \cap C)$. Since ω was arbitrary, this shows that any element of $A \cap (B \cup C)$ is also an element of $(A \cap B) \cup (A \cap C)$.

Thus we have shown

$$A \cap (B \cup C) \subset (A \cap B) \cup (A \cap C)$$

Step 2: " $\supset$ ". Next we will show that $(A \cap B) \cup (A \cap C) \subset A \cap (B \cup C)$.

Select an arbitrary element in $(A \cap B) \cup (A \cap C)$, denoted ω. Then $\omega \in (A \cap B)$ or $\omega \in (A \cap C)$. If $\omega \in A \cap B$, then $\omega \in B$. If $\omega \in A \cap C$, then $\omega \in C$. Thus ω is in either B or C, so $\omega \in B \cup C$. In either case, ω is also in A. Hence $\omega \in A \cap (B \cup C)$. Thus we have shown

$$(A \cap B) \cup (A \cap C) \subset A \cap (B \cup C)$$

Since we have shown that these sets are included in each other, they must be equal. This completes the proof. □

On the website, plug in each of the sets $(A \cap B) \cup (A \cap C)$ and $A \cap (B \cup C)$. Observe that the highlighted region doesn't change, since the sets are the same!

DeMorgan's Laws

In this section, we will show two important set identities useful for manipulating expressions of sets. These rules known as DeMorgan's Laws.

Theorem 0.0.2 (DeMorgan's Laws)**.** *Let A and B be subsets of our sample space Ω. Then*

(a) $(A \cup B)^c = A^c \cap B^c$

(b) $(A \cap B)^c = A^c \cup B^c$.

Proof.

(a) We will show that $(A \cup B)^c$ and $A^c \cap B^c$ are contained within each other.

Step 1: "$\subset$". Suppose $\omega \in (A \cup B)^c$. Then ω is not in the set $A \cup B$, i.e. in neither A nor B. Then $\omega \in A^c$ and $\omega \in B^c$, so $\omega \in A^c \cap B^c$. Hence $(A \cup B)^c \subset A^c \cap B^c$.

Step 2: "$\supset$". Suppose $\omega \in A^c \cap B^c$. Then ω is not in A and ω is not in B. So ω is in neither A nor B. This means ω is not in the set $(A \cup B)$, so $\omega \in (A \cup B)^c$. Hence $A^c \cap B^c \subset (A \cup B)^c$.

Since $A^c \cap B^c$ and $(A \cup B)^c$ are subsets of each other, they must be equal.

(b) Left as an exercise.

□

www.dbooks.org

If you're looking for more exercises, there is a link on the Set Theory page on the website that links to a page with many set identities. Try to prove some of these by showing that the sets are subsets of each other, or just plug them into the website to visualize them and see that their highlighted regions are the same.

Combinatorics

In many problems, to find the probability of an event, we will have to count the number of outcomes in Ω which satisfy the event, and divide by $|\Omega|$, i.e. the total number of outcomes in Ω. For example, to find the probability that a single die roll is even, we count the total number of even rolls, which is 3, and divide by the total number of rolls, 6. This gives a probability of $\frac{1}{2}$. But what if the event isn't as simple as "roll an even number"? For example if we flipped 10 coins, our event could be "flipped 3 heads total". How could we count the number of outcomes that have 3 heads in them without listing them all out? In this section, we will discover how to count the outcomes of such an event, and generalize the solution to be able to conquer even more complex problems.

Permutations

Suppose there are 3 students waiting in line to buy a spicy chicken sandwich. A question we could ask is, "How many ways can we order the students in this line?" Since there are so few students, let's just list out all possible orderings. We could have any of

$$\text{6 of these} \begin{cases} (1,2,3) \\ (1,3,2) \\ (2,1,3) \\ (2,3,1) \\ (3,1,2) \\ (3,2,1) \end{cases}$$

So there are 6 total possible orderings. If you look closely at the list above, you can see that there was a systematic way of listing them. We first wrote out all orderings starting with 1. Then came the orderings starting with 2, and then the ones that started with 3. In each of these groups of orderings starting with some particular student, there were two orderings. This is because once we fixed the first person in line, there were two ways to order the remaining two students. Denote N_i to be the number of ways to order i students. Now we

observe that the number of orderings can be written

$$N_3 = 3 \cdot N_2$$

since there are 3 ways to pick the first student, and N_2 ways to order the remaining two students. By similar reasoning,

$$N_2 = 2 \cdot N_1$$

Since the number of ways to order 1 person is just 1, we have $N_1 = 1$. Hence,

$$N_3 = 3 \cdot N_2 = 3 \cdot (2 \cdot N_1) = 3 \cdot 2 \cdot 1 = 6$$

which is the same as what we got when we just listed out all the orderings and counted them.

Now suppose we want to count the number of orderings for 10 students. 10 is big enough that we can no longer just list out all possible orderings and count them. Instead, we will make use of our method above. The number of ways to order 10 students is

$$N_{10} = 10 \cdot N_9 = 10 \cdot (9 \cdot N_8) = \cdots = 10 \cdot 9 \cdot 8 \cdot 7 \cdot \ldots \cdot 2 \cdot 1 = 3,628,800$$

It would have nearly impossible for us to list out over 3 million orderings of 10 students, but we were still able to count these orderings using our neat trick. We have a special name for this operation.

Definition 0.0.8. *The number of **permutations**, or orderings, of n distinct objects is given by the **factorial** expression,*

$$n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1$$

The factorial symbol is an exclamation point, which is used to indicate the excitement of counting.

Combinations

Now that we've established a quick method of counting the number of ways to order n distinct objects, let's figure out how to do our original problem. At the start of this section we asked how to count the number of ways we could flip 10 coins and have 3 of them be heads. The valid outcomes include

$$(\mathsf{H},\mathsf{H},\mathsf{H},\mathsf{T},\mathsf{T},\mathsf{T},\mathsf{T},\mathsf{T},\mathsf{T},\mathsf{T})$$
$$(\mathsf{H},\mathsf{H},\mathsf{T},\mathsf{H},\mathsf{T},\mathsf{T},\mathsf{T},\mathsf{T},\mathsf{T},\mathsf{T})$$
$$(\mathsf{H},\mathsf{H},\mathsf{T},\mathsf{T},\mathsf{H},\mathsf{T},\mathsf{T},\mathsf{T},\mathsf{T},\mathsf{T})$$
$$\vdots$$

www.dbooks.org

But its not immediately clear how to count all of these, and it definitely isn't worth listing them all out. Instead let's apply the permutations trick we learned in Section 3.2.2.

Suppose we have 10 coins, 3 of which are heads up, the remaining 7 of which are tails up. Label the 3 heads as coins 1, 2, and 3. Label the 7 tails as coins 4,5,6,7,8,9, and 10. There are 10! ways to order, or permute, these 10 (now distinct) coins. However, many of these permutations correspond to the same string of H's and T's. For example, coins 7 and 8 are both tails, so we would be counting the two permutations

$$(1, 2, 3, 4, 5, 6, \mathbf{7}, \mathbf{8}, 9, 10)$$
$$(1, 2, 3, 4, 5, 6, \mathbf{8}, \mathbf{7}, 9, 10)$$

as different, when they both correspond to the outcome

$$(\mathtt{H}, \mathtt{H}, \mathtt{H}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T})$$

hence we are *over counting* by just taking the factorial of 10. In fact, for the string above, we could permute the last 7 coins in the string (all tails) in 7! ways, and we would still get the same string, since they are all tails. To any particular permutation of these last 7 coins, we could permute the first 3 coins in the string (all heads) in 3! ways and still end up with the string

$$(\mathtt{H}, \mathtt{H}, \mathtt{H}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T})$$

This means that to each string of H's and T's, we can rearrange the coins in $3! \cdot 7!$ ways without changing the actual grouping of H's and T's in the string. So if there are 10! total ways of ordering the labeled coins, we are counting each unique grouping of heads and tails $3! \cdot 7!$ times, when we should only be counting it once. Dividing the total number of permutations by the factor by which we over count each unique grouping of heads and tails, we find that the number of unique groupings of H's and T's is

$$\#\text{ of outcomes with 3 heads and 7 tails} = \frac{10!}{3!7!}$$

This leads us to the definition of the binomial coefficient.

Definition 0.0.9. *The **binomial coefficient** is defined*

$$\binom{n}{k} \doteq \frac{n!}{k!(n-k)!}$$

The binomial coefficient, denoted $\binom{n}{k}$, represents the number of ways to pick k objects from n objects where the ordering within the chosen k objects doesn't matter. In the previous example, $n = 10$ and

$k = 3$. We could rephrase the question as, "How many ways can we pick 3 of our 10 coins to be heads?" The answer is then

$$\binom{n}{k} = \binom{10}{3} = \frac{10!}{3!(10-3)!} = \frac{10!}{3!7!} = 120$$

We read the expression $\binom{n}{k}$ as "n choose k". Let's now apply this counting trick to make some money.

Poker

One application of counting includes computing probabilities of poker hands. A poker hand consists of 5 cards drawn from the deck. The order in which we receive these 5 cards is irrelevant. The number of possible hands is thus

$$\binom{52}{5} = \frac{52!}{5!(52-5)!} = 2,598,960$$

since there are 52 cards to choose 5 cards from.

In poker, there are types of hands that are regarded as valuable in the following order form most to least valuable.

1. Royal Flush: A, K, Q, J, 10 all in the same suit.
2. Straight Flush: Five cards in a sequence, all in the same suit.
3. Four of a Kind: Exactly what it sounds like.
4. Full House: 3 of a kind with a pair.
5. Flush: Any 5 cards of the same suit, but not in sequence.
6. Straight: Any 5 cards in sequence, but not all in the same suit.
7. Three of a Kind: Exactly what it sounds like.
8. Two Pair: Two pairs of cards.
9. One Pair: One pair of cards.
10. High Card: Anything else.

Let's compute the probability of drawing some of these hands.

Exercise 0.0.4. *Compute the probabilities of the above hands.*

Solution.

1. There are only 4 ways to get this hand. Either we get the royal cards in diamonds, clubs, hearts, or spades. We can think of this has choosing 1 suit from 4 possible suits. Hence the probability of this hand is

$$P(\text{Royal Flush}) = \frac{\binom{4}{1}}{\binom{52}{5}} \approx 1.5 \cdot 10^{-6}$$

www.dbooks.org

2. Assuming hands like K, A, 2, 3, 4 don't count as consecutive, there are in total 10 valid consecutive sequences of 5 cards (each starts with any of A,2,...,10). We need to pick 1 of 10 starting values, and for each choice of a starting value, we can pick 1 of 4 suits to have them all in. This gives a total of $\binom{10}{1} \cdot \binom{4}{1} = 40$ straight flushes. However, we need to subtract out the probability of a royal flush, since one of the ten starting values we counted was 10 (10, J, Q, K, A is a royal flush). Hence the probability of this hand is

$$P(\text{Straight Flush}) = \frac{\binom{10}{1}\binom{4}{1} - \binom{4}{1}}{\binom{52}{5}} \approx 1.5 \cdot 10^{-5}$$

3. There are 13 values and only one way to get 4 of a kind for any particular value. However, for each of these ways to get 4 of a kind, the fifth card in the hand can be any of the remaining 48 cards. Formulating this in terms of our choose function, there are $\binom{13}{1}$ ways to choose the value, $\binom{12}{1}$ ways to choose the fifth card's value, and $\binom{4}{1}$ ways to choose the suit of the fifth card. Hence the probability of such a hand is

$$P(\text{Four of a Kind}) = \frac{\binom{13}{1}\binom{12}{1}\binom{4}{1}}{\binom{52}{5}} \approx 0.00024$$

4. For the full house, there are $\binom{13}{1}$ ways to pick the value of the triple, $\binom{4}{3}$ ways to choose which 3 of the 4 suits to include in the triple, $\binom{12}{1}$ ways to pick the value of the double, and $\binom{4}{2}$ ways to choose which 2 of the 4 suits to include in the double. Hence the probability of this hand is

$$P(\text{Full House}) = \frac{\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}}{\binom{52}{5}} \approx 0.0014$$

5. through 10. are left as exercises. The answers can be checked on the Wikipedia page titled "Poker probability".

□

Conditional Probability

Suppose we had a bag that contained two coins. One coin is a fair coin, and the other has a bias of 0.95, that is, if you flip this biased coin, it will come up heads with probability 0.95 and tails with probability 0.05. Holding the bag in one hand, you blindly reach in with your other, and pick out a coin. You flip this coin 3 times and see that all three times, the coin came up heads. You suspect that this coin is "likely" the biased coin, but how "likely" is it?

This problem highlights a typical situation in which new information changes the likelihood of an event. The original event was "we pick the biased coin". Before reaching in to grab a coin and then flipping it, we would reason that the probability of this event occurring (picking the biased coin) is $\frac{1}{2}$. After flipping the coin a couple of times and seeing that it landed heads all three times, we gain new information, and our probability should no longer be $\frac{1}{2}$. In fact, it should be much higher. In this case, we "condition" on the event of flipping 3 heads out of 3 total flips. We would write this new probability as

$$P(\text{picking the biased coin} \mid \text{flipping 3 heads out of 3 total flips})$$

The "bar" between the two events in the probability expression above represents "conditioned on", and is defined below.

Definition 0.0.10. *The probability of an event A conditioned on an event B is denoted and defined*

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

The intuition of this definition can be gained by playing with the visualization on the website. Suppose we drop a ball uniformly at random in the visualization. If we ask "What is the probability that a ball hits the orange shelf?", we can compute this probability by simply dividing the length of the orange shelf by the length of the entire space. Now suppose we are given the information that our ball landed on the green shelf. What is the probability of landing on the orange shelf now? Our green shelf has become our "new" sample space, and the proportion of the green shelf that overlaps with the orange shelf is now the only region in which we could have possibly landed on the orange shelf. To compute this new conditional probability, we would divide the length of the overlapping, or "intersecting", regions of the orange and green shelves by the total length of the green shelf.

www.dbooks.org

Bayes Rule

Now that we've understood where the definition of conditional probability comes from, we can use it to prove a useful identity.

Theorem 0.0.3 (Bayes Rule)**.** *Let A and B be two subsets of our sample space Ω. Then*

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Proof. By the definition of conditional probability,

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

Similarly,

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

Multiplying both sides by $P(A)$ gives

$$P(B \mid A)P(A) = P(A \cap B)$$

Plugging this into our first equation, we conclude

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

□

Coins in a Bag

Let's return to our first example in this section and try to use our new theorem to find a solution. Define the events

$$A \doteq \{\text{Picking the biased coin}\}$$
$$B \doteq \{\text{Flipping 3 heads out of 3 total flips}\}$$

We were interested in computing the probability $P(A \mid B)$. By Bayes Rule,

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

$P(B \mid A)$, i.e. the probability of flipping 3 heads out of 3 total flips given that we picked the biased coin, is simply $(0.95)^3 \approx 0.857$. The probability $P(A)$, i.e. the probability that we picked the biased coin is $\frac{1}{2}$ since we blindly picked a coin from the bag. Now all we need to do is compute $P(B)$, the overall probability of flipping 3 heads in this experiment. Remember from the set theory section, we can write

$$B = B \cap \Omega = B \cap (A \cup A^c) = (B \cap A) \cup (B \cap A^c)$$

So

$$P(B) = P((B \cap A) \cup (B \cap A^c)) = P(B \cap A) + P(B \cap A^c)$$

since the two sets $B \cap A$ and $B \cap A^c$ are disjoint. By the definition of conditional probability, we can write the above expression as

$$= P(B \mid A)P(A) + P(B \mid A^c)P(A^c)$$

We just computed $P(B \mid A)$ and $P(A)$. Similarly, the probability that we flip 3 heads given that we *didn't* pick the biased coin, denoted $P(B \mid A^c)$, is the probability that we flip 3 heads given we picked the fair coin, which is simply $(\frac{1}{2})^3 = 0.125$. The event A^c represents the event in which A does not happen, i.e. the event that we pick the fair coin. We have $P(A^c) = 1 - P(A) = 1 - \frac{1}{2} = \frac{1}{2}$. Hence

$$\begin{aligned} P(B) &= P(B \mid A)P(A) + P(B \mid A^c)P(A^c) \\ &= 0.857 \cdot 0.5 + 0.125 \cdot 0.5 \\ &= 0.491 \end{aligned}$$

Plugging this back into the formula given by Bayes Rule,

$$P(A \mid B) = \frac{0.857 \cdot 0.5}{0.491} = 0.873$$

Thus, given that we flipped 3 heads out of a total 3 flips, the probability that we picked the biased coin is roughly 87.3%.

Conditional Poker Probabilities

Within a game of poker, there are many opportunities to flex our knowledge of conditional probability. For instance, the probability of drawing a full house is 0.0014, which is less than 2%. But suppose we draw three cards and find that we have already achieved a pair. Now the probability of drawing a full house is higher than 0.0014. How much higher you ask? With our new knowledge of conditional probability, this question is easy to answer. We define the events

$$\begin{aligned} A &\doteq \{\text{Drawing a Full House}\} \\ B &\doteq \{\text{Drawing a Pair within the first three cards}\} \end{aligned}$$

By Bayes Rule,

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

$P(B \mid A)$, i.e. the probability that we draw a pair within the first three cards given that we drew a full house eventually, is 1. This is because every grouping of three cards within a full house must contain a

www.dbooks.org

pair. From Section 3.2.3, the probability of drawing a full house is $P(A) = 0.0014$.

It remains to compute $P(B)$, the probability that we draw a pair within the first three cards. The total number of ways to choose 3 cards from 52 is $\binom{52}{3}$. The number of ways to choose 3 cards containing a pair is $\binom{13}{1}\binom{4}{2}\binom{50}{1}$. There are $\binom{13}{1}$ to choose the value of the pair, $\binom{4}{2}$ ways to pick which two suits of the chosen value make the pair, and $\binom{50}{1}$ ways to pick the last card from the remaining 50 cards. Hence the probability of the event B is

$$P(B) = \frac{\binom{13}{1}\binom{4}{2}\binom{50}{1}}{\binom{52}{3}} \approx 0.176$$

Plugging this into our formula from Bayes Rule,

$$P(A \mid B) = \frac{1 \cdot 0.0014}{0.176} \approx 0.00795$$

It follows that our chance of drawing a full house has more than quadrupled, increasing from less than 2% to almost 8%.

Probability Distributions

Throughout the past chapters, we've actually already encountered many of the topics in this section. In order to define things like expectation and variance, we introduced random variables denoted X or Y as mappings from the sample space to the real numbers. All of the distributions we've so far looked at have been what are called *discrete* distributions. We will soon look at the distinction between discrete and continuous distributions. Additionally we will introduce perhaps the most influential theorem in statistics, the *Central Limit Theorem*, and give some applications.

Random Variables

In Section 2.2 (Expectation), we wanted to find the expectation of a coin flip. Since the expectation is defined as a weighted sum of outcomes, we needed to turn the outcomes into numbers before taking the weighted average. We provided the mapping

$$\mathsf{T} \mapsto 0$$
$$\mathsf{H} \mapsto 1$$

Here was our first encounter of a random variable.

Definition 0.0.11. *A function X that maps outcomes in our sample space to real numbers, written $X : \Omega \to$, is called a* ***random variable****.*

In the above example, our sample space was

$$\Omega = \{\mathsf{H}, \mathsf{T}\}$$

and our random variable $X : \Omega \to$, i.e. our function from the sample space Ω to the real numbers , was defined by

$$X(\mathsf{T}) = 0$$
$$X(\mathsf{H}) = 1$$

Now would be a great time to go onto the website and play with the "Random Variable" visualization. The sample space is represented

www.dbooks.org

by a hexagonal grid. Highlight some hexagons and specify the value your random variable X assigns to those hexagons. Start sampling on the grid to see the empirical frequencies on the left.

Independence of Random Variables

In previous sections we've mentioned independence of random variables, but we've always swept it under the rug during proofs since we hadn't yet formally defined the concept of a random variable. Now that we've done so, we can finally define a second form of independence (different from independence of *events*).

Definition 0.0.12. *Suppose X and Y are two random variables defined on some sample space Ω. We say X and Y are* ***independent random variables*** *if*

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for any two subsets A and B of .

Let's go back and prove Exercise 2.9 (c), i.e. that if X and Y are independent random variables, then

$$E[XY] = E[X]E[Y]$$

Proof. Define the random variable $Z(\omega) = X(\omega)Y(\omega)$. By the definition of expectation, the left hand side can be written

$$\begin{aligned}
E[XY] &= \sum_{z \in Z(\Omega)} z \cdot P(Z = z) \\
&= \sum_{x \in X(\Omega), y \in Y(\Omega)} xyP(X = x, Y = y) \\
&= \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} xyP(X \in \{x\}, Y \in \{y\}) \\
&= \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} xyP(X \in \{x\})P(Y \in \{y\}) \\
&= \sum_{x \in X(\Omega)} xP(X \in \{x\}) \sum_{y \in Y(\Omega)} yP(Y \in \{y\}) \\
&= E[X]E[Y]
\end{aligned}$$

This completes the proof. □

Discrete vs. Continuous

Thus far we have only studied discrete random variables, i.e. random variables that take on only up to *countably* many values. The word "countably" refers to a property of a set. We say a set is *countable* if we can describe a method to list out all the elements in the set such that for any particular element in the set, if we wait long enough in our listing process, we will eventually get to that element. In contrast, a set is called *uncountable* if we cannot provide such a method.

Countable vs. Uncountable

Let's first look at some examples.

Example 0.0.2. *The set of all natural numbers*

$$\mathbb{N} \doteq \{1, 2, 3, \dots\}$$

is countable. Our method of enumeration could simply be to start at 1 and add 1 every iteration. Then for any fixed element $n \in \mathbb{N}$, this process would eventually reach and list out n.

Example 0.0.3. *The integers,*

$$\mathbb{Z} \doteq \{0, 1, -1, 2, -2, 3, -3, \dots\}$$

is countable. Our method of enumeration as displayed above is to start with 0 for the first element, add 1 to get the next element, multiply by -1 to get the third element, and so on. Any integer $k \in \mathbb{Z}$, if we continue this process long enough, will be reached.

Example 0.0.4. *The set of real numbers in the interval $[0, 1]$ is uncountable. To see this, suppose for the sake of contradiction that this set were countable. Then there would exist some enumeration of the numbers in decimal form. It might look like*

$$\begin{aligned}
&0 . 1\ 3\ 5\ 4\ 2\ 9\ 5 \dots \\
&0 . 4\ 2\ 9\ 4\ 7\ 2\ 6 \dots \\
&0 . 3\ 9\ 1\ 6\ 8\ 3\ 1 \dots \\
&0 . 9\ 8\ 7\ 3\ 4\ 3\ 5 \dots \\
&0 . 2\ 9\ 1\ 8\ 1\ 3\ 6 \dots \\
&0 . 3\ 7\ 1\ 6\ 1\ 8\ 2 \dots \\
&\vdots
\end{aligned}$$

Consider the element along the diagonal of such an enumeration. In this case the number is

$$a \doteq 0.121318 \dots$$

www.dbooks.org

Now consider the number obtained by adding 1 to each of the decimal places, i.e.

$$a' \doteq 0.232429\ldots$$

This number is still contained in the interval $[0,1]$*, but does not show up in the enumeration. To see this, observe that* a' *is not equal to the first element, since it differs in the first decimal place by 1. Similarly, it is not equal to the second element, as* a' *differs from this number by 1 in the second decimal place. Continuing this reasoning, we conclude that* a' *differs from the* n^{th} *element in this enumeration in the* n^{th} *decimal place by 1. It follows that if we continue listing out numbers this way, we will never reach the number* a'*. This is a contradiction since we initially assumed that our enumeration would eventually get to every number in* $[0,1]$*. Hence the set of numbers in* $[0,1]$ *is uncountable.*

If you're left feeling confused after these examples, the important take away is that an uncountable set is *much* bigger than a countable set. Although both are infinite sets of elements, uncountable infinity refers to a "bigger" notion of infinity, one which has no gaps and can be visualized as a continuum.

Discrete Distributions

Definition 0.0.13. *A random variable X is called* ***discrete*** *if X can only take on finitely many or countably many values.*

For example, our coin flip example yielded a random variable X which could only take values in the set $\{0,1\}$. Hence, X was a discrete random variable. However, discrete random variables can still take on infinitely many values, as we see below.

Example 0.0.5 (Poisson Distribution)**.** *A useful distribution for modeling many real world problems is the Poisson Distribution. Suppose* $\lambda > 0$ *is a positive real number. Let X be distributed according to a Poisson distribution with parameter* λ*, i.e.*

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

where $k \in \mathbb{N}$*. The shorthand for stating such a distribution is* $X \sim Poi(\lambda)$*. Since k can be any number in* $\mathbb{N}$*, our random variable X has a positive probability on infinitely many numbers. However, since* $\mathbb{N}$ *is countable, X is still considered a discrete random variable.*

On the website there is an option to select the "Poisson" distribution in order to visualize its probability mass function. Changing the value of λ *changes the probability mass function, since* λ *shows up in the probability*

expression above. Drag the value of λ from 0.01 up to 10 to see how varying λ changes the probabilities.

Example 0.0.6 (Binomial Distribution)**.** *Another useful distribution is called the Binomial Distribution. Consider n coin flips, i.e. n random variables $X_1, \ldots, X_n$ each of the form*

$$X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

Now consider the random variable defined by summing all of these coin flips, i.e.

$$S \doteq \sum_{i=1}^{n} X_i$$

We might then ask, "What is the probability distribution of S?" Based on the definition of S, it can take on values from 0 to n, however it can only take on the value 0 if all the coins end up tails. Similarly, it can only take on the value n if all the coins end up heads. But to take on the value 1, we only need one of the coins to end up heads and the rest to end up tails. This can be achieved in many ways. In fact, there are $\binom{n}{1}$ ways to pick which coin gets to be heads up. Similarly, for $S = 2$, there are $\binom{n}{2}$ ways to pick which two coins get to be heads up. It follows that for $S = k$, there are $\binom{n}{k}$ ways to pick which k coins get to be heads up. This leads to the following form,

$$P(S = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The p^k comes from the k coins having to end up heads, and the $(1-p)^{n-k}$ comes from the remaining $n-k$ coins having to end up tails. Here it is clear that k ranges from 0 to n, since the smallest value is achieved when no coins land heads up, and the largest number is achieved when all coins land heads up. Any value between 0 and n can be achieved by picking a subset of the n coins to be heads up.

Selecting the "Binomial" distribution on the website will allow you to visualize the probability mass function of S. Play around with n and p to see how this affects the probability distribution.

Continuous Distributions

Definition 0.0.14. *We say that X is a **continuous** random variable if X can take on uncountably many values.*

If X is a continuous random variable, then the probability that X takes on any particular value is 0.

www.dbooks.org

Example 0.0.7. *An example of a continuous random variable is a Uniform[0,1] random variable. If $X \sim$ Uniform[0,1], then X can take on any value in the interval [0,1], where each value is equally likely. The probability that X takes on any particular value in $[0, 1]$, say $\frac{1}{2}$ for example, is 0. However, we can still take probabilities of subsets in a way that is intuitive. The probability that x falls in some interval (a, b) where $0 \leq a < b \leq 1$ is written*

$$P(X \in (a, b)) = b - a$$

The probability of this event is simply the length of the interval (a, b).

A continuous random variable is distributed according to a *probability density function*, usually denoted f, defined on the domain of X. The probability that X lies in some set A is defined as

$$P(X \in A) = \int_A f$$

This is informal notation but the right hand side of the above just means to integrate the density function f over the region A.

Definition 0.0.15. *A **probability density function** f (abbreviated **pdf**) is valid if it satisfies the following two properties.*

1. $f(x) \geq 0$ *for all* $x \in$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Example 0.0.8 (Exponential Distribution)**.** *Let $\lambda > 0$ be a positive real number. Suppose X is a continuous random variable distributed according to the density*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Let's check that f defines a valid probability density function. Since $\lambda > 0$ and e^y is positive for any $y \in$, we have $f(x) \geq 0$ for all $x \in$. Additionally, we have

$$\begin{aligned} \int_0^{\infty} f(x)dx &= \int_0^{\infty} \lambda e^{-\lambda x} \\ &= \left[\lambda \frac{-1}{\lambda} e^{-\lambda x}\right]_0^{\infty} \\ &= 0 - (-1) \\ &= 1 \end{aligned}$$

Since f is nonnegative and integrates to 1, it is a valid pdf.

Example 0.0.9 (Normal Distribution). *We arrive at perhaps the most known and used continuous distributions in all of statistics. The Normal distribution is specified by two parameters, the mean μ and variance σ^2. To say X is a random variable distributed according to a Normal distribution with mean μ and variance σ^2, we would write $X \sim N(\mu, \sigma^2)$. The corresponding pdf is*

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

Some useful properties of normally distributed random variables are given below.

Proposition 0.0.2. *If $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ are independent random variables, then*

(a) The sum is normally distributed, i.e.

$$X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

(b) Scaling by a factor $a \in$ results in another normal distribution, i.e. we have

$$aX \sim N(a\mu_x, a^2\sigma_x^2)$$

(c) Adding a constant $a \in$ results in another normal distribution, i.e.

$$X + a \sim N(\mu_x + a, \sigma_x^2)$$

Heuristic. In order to rigorously prove this proposition, we need to use moment generating functions, which aren't covered in these notes.

However, if we believe that $X + Y$, aX, and $X + a$ are all still normally distributed, it follows that the specifying parameters (μ and σ^2) for the random variables in (a), (b), and (c) respectively are

$$\begin{aligned} E(X + Y) &= EX + EY = \mu_x + \mu_y \\ \text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y) = \sigma_x^2 + \sigma_y^2 \end{aligned}$$

and

$$\begin{aligned} E(aX) &= aEX = a\mu_x \\ \text{Var}(aX) &= a^2\text{Var}(X) = a^2\sigma_x^2 \end{aligned}$$

and

$$\begin{aligned} E(X + a) &= EX + a = \mu_x + a \\ \text{Var}(X + a) &= \text{Var}(X) + \text{Var}(a) = \text{Var}(X) = \sigma_x^2 \end{aligned}$$

□

www.dbooks.org

The Central Limit Theorem

We return to dice rolling for the moment to motivate the next result. Suppose you rolled a die 50 times and recorded the average roll as $\bar{X}_1 = \frac{1}{50}\sum_{k=1}^{50} X_k$. Now you repeat this experiment and record the average roll as $\bar{X}_2$. You continue doing this and obtain a sequence of sample means $\{\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots\}$. If you plotted a histogram of the results, you would begin to notice that the $\bar{X}_i$'s begin to look normally distributed. What are the mean and variance of this approximate normal distribution? They should agree with the mean and variance of $\bar{X}_i$, which we compute below. Note that these calculations don't depend on the index i, since each $\bar{X}_i$ is a sample mean computed from 50 independent fair die rolls. Hence we omit the index i and just denote the sample mean as $\bar{X} = \frac{1}{50}\sum_{k=1}^{50} X_k$.

$$\begin{aligned} E(\bar{X}) &= E\Big(\frac{1}{50}\sum_{k=1}^{50} X_k\Big) \\ &= \frac{1}{50}\sum_{k=1}^{50} E(X_k) \\ &= \frac{1}{50}\sum_{k=1}^{50} 3.5 \\ &= \frac{1}{50}\cdot 50 \cdot 3.5 \\ &= 3.5 \end{aligned}$$

where the second equality follows from linearity of expectations, and the third equality follows from the fact that the expected value of a die roll is 3.5 (See Section 2.2). The variance of $\bar{X}_i$ is

$$\begin{aligned} \mathrm{Var}(\bar{X}) &= \mathrm{Var}\Big(\frac{1}{50}\sum_{k=1}^{50} X_k\Big) && \text{(Definition of } \bar{X}_i\text{)} \\ &= \frac{1}{50^2}\mathrm{Var}\Big(\sum_{k=1}^{50} X_k\Big) && (\mathrm{Var}(cY) = c^2\mathrm{Var}(Y)) \\ &= \frac{1}{50^2}\sum_{k=1}^{50} \mathrm{Var}(X_k) && (X_k\text{'s are independent.}) \\ &= \frac{1}{50^2}\cdot 50 \cdot \mathrm{Var}(X_k) && (X_k\text{'s are identically distributed.}) \\ &\approx \frac{1}{50}\cdot 2.92 \\ &\approx 0.0583 \end{aligned}$$

where we computed $\mathrm{Var}(X_k) \approx 2.92$ in Exercise 2.12. So we would begin to observe that the sequence of sample means begins to resemble a normal distribution with mean $\mu = 3.5$ and variance

$\sigma^2 = 0.0582$. This amazing result follows from the Central Limit Theorem, which is stated below.

Theorem 0.0.4 (Central Limit Theorem)**.** *Let $X_1, X_2, X_3, \ldots$ be iid (independent and identically distributed) with mean μ and variance σ^2. Then*

$$\bar{X} \to N\left(\mu, \frac{\sigma^2}{n}\right)$$

in distribution as $n \to \infty$.

All this theorem is saying is that as the number of samples n grows large, independent observations of the sample mean $\bar{X}$ look as though they were drawn from a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. The beauty of this result is that this type of convergence to the normal distribution holds for any underlying distribution of the X_i's. In the previous discussion, we assumed that each X_i was a die roll, so that the underlying distribution was discrete uniform over the set $\Omega = \{1, 2, 3, 4, 5, 6\}$. However, this result is true for any underlying distribution of the X_i's.

A continuous distribution we have not yet discussed is the Beta distribution. It is characterized by two parameters α and β (much like the normal distribution is characterized by the parameters μ and σ^2.) On the Central Limit Theorem page of the website, choose values for α and β and observe that the sample means look as though they are normally distributed. This may take a while but continue pressing the "Submit" button until the histogram begins to fit the normal curve (click the check box next to "Theoretical" to show the plot of the normal curve).

Corollary 0.0.2. *Another way to write the convergence result of the Central Limit Theorem is*

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0, 1)$$

Proof. By the CLT, $\bar{X}$ becomes distributed $N(\mu, \frac{\sigma^2}{n})$. By Proposition 4.14 (c), $\bar{X} - \mu$ is then distributed

$$\bar{X} - \mu \sim N\left(\mu - \mu, \frac{\sigma^2}{n}\right) = N\left(0, \frac{\sigma^2}{n}\right)$$

Combining the above with Proposition 4.14 (a), we have that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is distributed

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N\left(0, \frac{\sigma^2}{n} \cdot \left(\frac{1}{\sigma/\sqrt{n}}\right)^2\right) = N(0, 1)$$

□

www.dbooks.org

Frequentist Inference

The topics of the next three sections are useful applications of the Central Limit Theorem. Without knowing anything about the underlying distribution of a sequence of random variables $\{X_i\}$, for large sample sizes, the CLT gives a statement about the sample means. For example, if Y is a $N(0,1)$ random variable, and $\{X_i\}$ are distributed iid with mean μ and variance σ^2, then

$$P\Big(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \in A\Big) \approx P(Y \in A)$$

In particular, if we want an interval in which Y lands with probability 0.95, we look online or in a book for a z table, which will tell us that for a $N(0,1)$ random variable Y,

$$P(Y \in (-1.96, 1.96)) = P(-1.96 \leq Y \leq 1.96) = 0.95$$

Since $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is nearly $N(0,1)$ distributed, this means

$$P\Big(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\Big) = 0.95$$

From the above statement we can make statements about experiments in order to quantify confidence and accept or reject hypotheses.

Confidence Intervals

Suppose that during the presidential election, we were interested in the proportion p of the population that preferred Hillary Clinton to Donald Trump. It wouldn't be feasible to call every single person in the country and write down who they prefer. Instead, we can take a bunch of samples, $X_1, \ldots, X_n$ where

$$X_i = \begin{cases} 1 & \text{if person } i \text{ prefers Hillary} \\ 0 & \text{otherwise} \end{cases}$$

www.dbooks.org

Then the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the proportion of our sample that prefers Hillary. Let p be the true proportion that prefer Hillary (p is not known). Note that $E\bar{X} = p$, since each X_i is 1 with probability p and 0 with probability $1 - p$. Then by the CLT,

$$\frac{\bar{X} - p}{\sigma/\sqrt{n}} \sim N(0,1)$$

Since we don't know the true value of σ, we estimate it using the sample variance, defined

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

This is a consistent estimator for σ^2, so for large n, the probability that it differs greatly from the true variance σ^2 is small. Hence we can replace σ in our expression with $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$. Since $\frac{\bar{X}-p}{S/\sqrt{n}}$ is approximately $N(0,1)$ distributed, we have

$$P\Big(-1.96 \leq \frac{\bar{X} - p}{S/\sqrt{n}} \leq 1.96\Big) = 0.95$$

Rearranging the expression for p, we have

$$\begin{aligned} P\Big(-1.96 \cdot \frac{S}{\sqrt{n}} \leq \bar{X} - p \leq 1.96 \cdot \frac{S}{\sqrt{n}}\Big) &= 0.95 \\ \Rightarrow P\Big(-1.96 \cdot \frac{S}{\sqrt{n}} - \bar{X} \leq -p \leq 1.96 \cdot \frac{S}{\sqrt{n}} - \bar{X}\Big) &= 0.95 \\ \Rightarrow P\Big(1.96 \cdot \frac{S}{\sqrt{n}} + \bar{X} \geq p \geq \bar{X} - 1.96 \cdot \frac{S}{\sqrt{n}}\Big) &= 0.95 \end{aligned}$$

Even though we do not know the true value for p, we can conclude from the above expression that with probability 0.95, p is contained in the interval

$$\Big(\bar{X} - 1.96 \cdot \frac{S}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{S}{\sqrt{n}}\Big)$$

This is called a 95% confidence interval for the parameter p. This approximation works well for large values of n, but a rule of thumb is to make sure $n > 30$ before using the approximation.

On the website, there is a confidence interval visualization. Try selecting the Uniform distribution to sample from. Choosing a sample size of $n = 30$ will cause batches of 30 samples to be picked, their sample means computed, and their resulting confidence intervals displayed on the right. Depending on the confidence level picked (the above example uses $\alpha = 0.05$, so $1 - \alpha = 0.95$), the generated confidence intervals will contain the true mean μ with probability $1 - \alpha$.

Hypothesis Testing

Let's return to the example of determining voter preference in the 2016 presidential election. Suppose we suspect that the proportion of voters who prefer Hillary Clinton is greater than $\frac{1}{2}$, and that we take n samples, denoted $\{X_i\}_{i=1}^n$ from the U.S. population. Based on these samples, can we support or reject our hypothesis that Hillary Clinton is more popular? And how confident are we in our conclusion? Hypothesis testing is the perfect tool to help answer these questions.

Constructing a Test

A hypothesis in this context is a statement about a parameter of interest. In the presidential election example, the parameter of interest was p, the proportion of the population who supported Hillary Clinton. A hypothesis could then be that $p > 0.5$, i.e. that more than half of the population supports Hillary.

There are four major components to a hypothesis test.

1. The *alternative hypothesis*, denoted H_a, is a claim we would like to support. In our previous example, the alternative hypothesis was $p > 0.5$.

2. The *null hypothesis*, denoted H_0 is the opposite of the alternative hypothesis. In this case, the null hypothesis is $p \leq 0.5$, i.e. that less than half of the population supports Hillary.

3. The *test statistic* is a function of the sample observations. Based on the test statistic, we will either accept or reject the null hypothesis. In the previous example, the test statistic was the sample mean $\bar{X}$. The sample mean is often the test statistic for many hypothesis tests.

4. The *rejection region* is a subset of our sample space Ω that determines whether or not to reject the null hypothesis. If the test statistic falls in the rejection region, then we reject the null hypothesis. Otherwise, we accept it. In the presidential election example, the rejection region would be

$$\text{RR: } \{(x_1, \ldots, x_n) : \bar{X} > k\}$$

This notation means we reject if $\bar{X}$ falls in the interval (k, ∞), where k is some number which we must determine. k is determined by the Type I error, which is defined in the next section. Once k is computed, we reject or accept the null hypothesis depending on the value of our test statistic, and our test is complete.

www.dbooks.org

Types of Error

There are two fundamental types of errors in hypothesis testing. They are denoted Type I and II error.

Definition 0.0.16. *A **Type I error** is made when we reject H_0 when it is in fact true. The probability of Type I error is typically denoted as α.*

In other words, α is the probability of a false positive.

Definition 0.0.17. *A **Type II error** is made when we accept H_0 when it is in fact false. The probability of Type II error is typically denoted as β.*

In other words, β is the probability of a false negative.

In the context of hypothesis testing, α will determine the rejection region. If we restrict the probability of a false positive to be less than 0.05, then we have

$$P(\bar{X} \in \text{RR} \mid H_0) \leq 0.05$$

i.e. our test statistic falls in the rejection region (meaning we reject H_0), given that H_0 is true, with probability 0.05. Continuing along our example of the presidential election, the rejection region was of the form $\bar{X} > k$, and the null hypothesis was that $p \leq 0.5$. Our above expression then becomes

$$P(\bar{X} > k \mid p \leq 0.5) \leq 0.05$$

If $n > 30$, we can apply the CLT to say,

$$P(\frac{\bar{X} - p}{S/\sqrt{n}} > \frac{k - p}{S/\sqrt{n}} \mid p \leq 0.5) = P(Y > \frac{k - p}{S/\sqrt{n}} \mid p \leq 0.5)$$

where Y is a $N(0, 1)$ random variable. Since $p \leq 0.5$ implies $\frac{k-p}{S/\sqrt{n}} \geq \frac{k-0.5}{S/\sqrt{n}}$, we must also have

$$Y > \frac{k - p}{S/\sqrt{n}} \Rightarrow Y > \frac{k - 0.5}{S/\sqrt{n}}$$

Hence,

$$P(Y > \frac{k - p}{S/\sqrt{n}} \mid p \leq 0.5) \leq P(Y > \frac{k - 0.5}{S/\sqrt{n}})$$

So if we bound the probability on the right side of the inequality by 0.05, then we also bound the probability on the left (the Type I error, α) by 0.05. Since Y is distributed $N(0, 1)$, we can look up a z table to find that $z_{0.05} = -1.64$, so

$$P(Y > 1.64) = P(Y < -1.64) = 0.05$$

Letting $\frac{k-0.5}{S/\sqrt{n}} = 1.64$, we can solve for k to determine our rejection region.

$$k = 0.5 + 1.64 \cdot \frac{S}{\sqrt{n}}$$

Since our rejection region was of the form $\bar{X} > k$, we simply check whether $\bar{X} > 0.5 + 1.64 \cdot \frac{S}{\sqrt{n}}$. If this is true, then we reject the null, and conclude that more than half the population favors Hillary Clinton. Since we set $\alpha = 0.05$, we are $1 - \alpha = 0.95$ confident that our conclusion was correct.

In the above example, we determined the rejection region by plugging in 0.5 for p, even though the null hypothesis was $p \leq 0.5$. It is almost as though our null hypothesis was $H_0 : p = 0.5$ instead of $H_0 : p \leq 0.5$. In general, we can simplify H_0 and assume the border case ($p = 0.5$ in this case) when we are determining the rejection region.

www.dbooks.org

p-Values

As we saw in the previous section, a selected α determined the rejection region so that the probability of a false positive was less than α. Now suppose we observe some test statistic, say, the sample proportion of voters $\bar{X}$ who prefer Hillary Clinton. We then ask the following question. Given $\bar{X}$, what is the smallest value of α such that we still reject the null hypothesis? This leads us to the following definition.

Definition 0.0.18. *The p-value, denoted p, is defined*

$$p = \min\{\alpha \in (0,1) : \textit{Reject } H_0 \textit{ using an } \alpha \textit{ level test}\}$$

i.e. the smallest value of α for which we still reject the null hypothesis.

This definition isn't that useful for computing p-values. In fact, there is a more intuitive way of thinking about them. Suppose we observe some sample mean $\bar{X}_1$. Now suppose we draw a new sample mean, $\bar{X}_2$. The p-value is just the probability that our new sample mean is more *extreme* than the one we first observed, assuming the null hypothesis is true. By "extreme" we mean, more different from our null hypothesis.

Below we go through an example which verifies that the intuitive definition given above agrees with Definition 5.3.

Example 0.0.10. *Suppose that we sampled n people and asked which candidate they preferred. As we did before, we can represent each person as an indicator function,*

$$X_i = \begin{cases} 1 & \textit{if person } i \textit{ prefers Hillary} \\ 0 & \textit{otherwise} \end{cases}$$

Then $\bar{X}$ is the proportion of the sample that prefers Hillary. After taking the n samples, suppose we observe that $\bar{X} = 0.7$. If we were to set up a hypothesis test, our hypotheses, test statistic, and rejection region would be

$$\begin{aligned} &H_0 : q \leq 0.5 \\ &H_a : q > 0.5 \\ &\textit{Test statistic: } \bar{X} \\ &\textit{RR: } \{(x_1, \ldots, x_n) : \bar{X} > k\} \end{aligned}$$

where q is the true proportion of the entire U.S. population that favors Hillary. Using the intuitive definition, the p value is the probability that we observe something more extreme than 0.7. Since the null hypothesis is that $q \leq 0.5$, "more extreme" in this case means, "bigger than 0.7". So the p-value is the probability that, given a new sample, we observe the new $\bar{X}$ is

greater than 0.7, assuming the null, i.e. that $q \leq 0.5$. Normalizing $\bar{X}$, we have

$$P(\bar{X} > 0.7 \mid H_0) = P\Big(\frac{\bar{X} - 0.5}{S/\sqrt{n}} > \frac{0.7 - 0.5}{S/\sqrt{n}}\Big) \approx P\Big(Y > \frac{0.7 - 0.5}{S/\sqrt{n}}\Big) \doteq p \tag{4}$$

where $Y \sim N(0,1)$. We would then compute the value $z_p \doteq \frac{0.7-0.5}{S/\sqrt{n}}$ by plugging in the sample standard deviation, S, and the number of samples we took, n. We would then look up a z table and find the probability corresponding to z_p, denoted p (this is our p value).

We now claim that this p is equal to the smallest α for which we reject the null hypothesis, i.e. that our intuitive definition of a p-value agrees with Definition 5.3. To show that

$$p = \min\{\alpha \in (0,1) : \textit{Reject } H_0 \textit{ using an } \alpha \textit{ level test}\},$$

we need to show that for any $\alpha < p$, we accept the null hypothesis. We also need to show that for any $\alpha \geq p$, we reject the null hypothesis.

***Case 1:** Suppose $\alpha < p$. We need to show that the test statistic $\bar{X} = 0.7$ falls in the acceptance region determined by α. Using a z table, we could find z_α such that*

$$\alpha = P(Y > z_\alpha) \approx P(\frac{\bar{X} - 0.5}{S/\sqrt{n}} > z_\alpha \mid H_0) = P(\bar{X} > z_\alpha \cdot \frac{S}{\sqrt{n}} + 0.5 \mid H_0)$$

Since the RHS of the above expression is the probability of Type I error, the rejection region is determined by

$$\bar{X} > k_\alpha \doteq z_\alpha \cdot \frac{S}{\sqrt{n}} + 0.5$$

Since $\alpha < p$, the corresponding z_p such that $p = P(Y > z_p)$ satisfies $z_p < z_\alpha$. By the RHS of expression (1),

$$p = P\Big(Y > \frac{0.7 - 0.5}{S/\sqrt{n}}\Big)$$

which implies $z_p = \frac{0.7-0.5}{S/\sqrt{n}} \Rightarrow z_p \cdot \frac{S}{\sqrt{n}} + 0.5 = 0.7$. This implies that

$$0.7 = z_p \cdot \frac{S}{\sqrt{n}} + 0.5 < z_\alpha \cdot \frac{S}{\sqrt{n}} + 0.5 = k_\alpha$$

Therefore $\bar{X} = 0.7 < k_\alpha$ implies $\bar{X} = 0.7$ is in the acceptance region determined by α. Hence, we accept the null hypothesis for any $\alpha < p$.

***Case 2:** Suppose $\alpha \geq p$. We need to show that the test statistic $\bar{X} = 0.7$ falls in the rejection region determined by α. By reasoning similar to the kind in Case 1, we would have $z_\alpha \leq z_p$. This implies*

$$k_\alpha \doteq z_\alpha \cdot \frac{S}{\sqrt{n}} + 0.5 \leq z_p \cdot \frac{S}{\sqrt{n}} + 0.5 = 0.7$$

Hence $\bar{X} = 0.7 \geq k_\alpha$ implies that $\bar{X} = 0.7$ is in the rejection region determined by α. Hence, we reject the null hypothesis for any $\alpha \geq p$.

www.dbooks.org

Example 5.4 (above) justifies the definition of p-values which gives an easy way to compute them. Given some observation of our test statistic $\bar{X}$, we compute the p-value by calculating the probability of seeing something *more* different or "extreme" than our observed $\bar{X}$, assuming H_0 is true. By the argument in Example 5.4, this value is the same as the smallest α level for which we reject H_0.

Bayesian Inference

Introduction

The frequentist approach to inference holds that probabilities are intrinsicially tied (unsurprisingly) to frequencies. This interpretation is actually quite natural. What, according to a frequentist, does it mean to say that the probability a fair coin will come up heads is 1/2? Well, simply that in an infinite sequence of independent tosses of the same coin, half will come up heads (loosely speaking). Many random experiments are in fact repeatable, and the frequentist paradigm readily applies in such situations.

It is often desirable, however, to assign probabilities to events that are not repeatable. When the weather forecast tells you that there is a 90% chance of rain tomorrow, for example, it is assigning a probability to a one-off event, since tomorrow only happens once! What's more, there are many scenarios in which we would like to assign probabilities to non-random events that nevertheless involve uncertainty. A bank might be interested in designing an automated system that computes the probability that a signature on a check is genuine. Even though there is an underlying ground truth (the signature is either genuine or not), there is uncertainty from the bank's point of view, so the use of probability is justified. The pure frequentist interpretation of probabilities cannot be squared up with either of these use cases.

Bayesian inference takes a subjective approach and views probabilities as representing degrees of belief. It is thus perfectly valid to assign probabilities to non-repeating and non-random events, so long as there is uncertainty that we wish to quantify. The fact that Bayesian probabilities are subjective does not mean they are arbitrary. The rules for working with Bayesian probabilities are identical to those for working with the frequentist variety. Bayesians are simply happy to assign probabilities to a larger class of events than frequentists are.

The essential spirit of Bayesian inference is encapsulated by Bayes'

www.dbooks.org

theorem.

Bayes' theorem

Suppose that during a routine medical examination, your doctor informs you that you have tested positive for a rare disease. You are initially distressed, but as a good statistician, you are also aware that these tests can be finicky and there is some uncertainty in their results. Unfortunately for you, this test is quite accurate — it reports a positive result for 95% of the patients with the disease, and a negative result for 95% of the healthy patients. The outlook does not appear to be good.

As a good *Bayesian* statistician, however, you realize that these test accuracies are not quite the bottom line, as far as your health is concerned. If we let "+" and "−" denote a positive and negative test result, respectively, then the test accuracies are the conditional probabilities

$$P(+ \mid \text{disease}) = 0.95,$$
$$P(- \mid \text{healthy}) = 0.95.$$

But what you are really interested in is

$$P(\text{disease} \mid +).$$

In order to compute this last quantity, we need to "turn around" the conditional probabilities encoded in the test accuracies. This is achieved by Bayes' theorem.

Theorem 0.0.5 (Bayes' Theorem)**.** *Let $Y_1, \ldots, Y_k$ be a partition of the sample space Ω and let X be any event. Then*

$$P(Y_j|X) = \frac{P(X|Y_j)P(Y_j)}{\sum_{i=1}^{k} P(X|Y_i)P(Y_i)}.$$

Since "disease" and "healthy" partition the sample space of outcomes, we have

$$P(\text{disease}|+) = \frac{P(+|\text{disease})P(\text{disease})}{P(+|\text{disease})P(\text{disease}) + P(+|\text{healthy})P(\text{healthy})}.$$

Importantly, Bayes' theorem reveals that in order to compute the conditional probability that you have the disease given the test was positive, you need to know the "prior" probability you have the disease $P(\text{disease})$, given no information at all. That is, you need to know the overall incidence of the disease in the population to which you belong. We mentioned earlier that this is a rare disease. In fact, only 1 in 1,000 people are affected, so $P(\text{disease}) = 0.001$, which

in turn implies $P(\text{healthy}) = 0.999$. Plugging these values into the equation above gives

$$P(\text{disease} \mid +) \approx 0.019.$$

In other words, despite the apparent reliability of the test, the probability that you actually have the disease is still less than 2%. The fact that the disease is so rare means that most of the people who test positive will be healthy, simply because most people are healthy in general. Note that the test is certainly not useless; getting a positive result increases the probability you have the disease by about 20-fold. But it is incorrect to interpret the 95% test accuracy as the probability you have the disease.

The Bayesian procedure

The above example is illustrative of the general procedure for doing Bayesian inference. Suppose you are interested in some parameter θ.

1. Encode your initial beliefs about θ in the form of a *prior distribution* $P(\theta)$.

2. Collect data X via experimentation, observation, querying, etc.

3. Update your beliefs using Bayes' theorem to the *posterior distribution*

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}.$$

4. Repeat the entire process as more data become available.

Prior, likelihood, posterior

As it turns out, Bayes' theorem is so fundamental to Bayesian inference that special names are given to the terms in the equation.

Prior

The prior distribution is the unconditional distribution $P(\theta)$. The goal of the prior is to capture our pre-existing knowledge about θ, before we see any data. In the medical testing example, we used the incidence of the disease in the population as the prior probability that any particular individual has the disease.

www.dbooks.org

Likelihood

In Bayesian and frequentist statistics alike, the likelihood of a parameter θ given data X is $P(X|\theta)$. The likelihood function plays such an important role in classical statistics that it gets its own letter:

$$L(\theta|X) = P(X|\theta).$$

This notation emphasizes the fact that we view the likelihood as a function of θ for some fixed data X.

Figure 4 shows a random sample x of 8 points drawn from a standard normal distribution, along with the corresponding likelihood function of the mean parameter.

Figure 4: The orange curve shows the likelihood function for the mean parameter of a normal distribution with variance 1, given a sample of 8 points (middle) from a standard normal density (top).

In general, given a sample of n independent and identically distributed random variables $X_1, \ldots, X_n$ from some distribution $P(X|\theta)$, the likelihood is

$$\begin{aligned} L(\theta|X_1, \ldots, X_n) &= P(X_1, \ldots, X_n|\theta) \\ &= \prod_{i=1}^{n} P(X_i|\theta). \end{aligned}$$

In the case of the normal distribution with variance 1 and unknown mean θ, this equation suggests a way to visualize how the likelihood function is generated. Imagine sliding the probability density function of a Normal$(\theta, 1)$ distribution from left to right by gradually increasing θ. As we encounter each sample X_i, the density "lifts" the point off the x-axis. The dotted lines in the middle panel of Figure 5 represent the quantities $P(X_i|\theta)$. Their product is precisely the likelihood, which is plotted in orange at the bottom of Figure 5.

3slide.png

Figure 5: Each orange dotted line in the middle panel represents the quantity $P(x_i|\theta)$. The product of the lengths of these dotted lines is the likelihoood for the value of θ that produced the density in blue.

We can see that the likelihood is maximized by the value of θ for which the density of a Normal$(\theta, 1)$ distribution is able to lift the most points the furthest off the x-axis. It can be shown that this maximizing value is given by the sample mean

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

In this case we say that the sample mean is the *maximum likelihood estimator* of the parameter θ.

In Bayesian inference, the likelihood is used to measure quantify the degree to which a set of data X supports a particular parameter value θ. The essential idea is that if the data could be generated by a given parameter value θ with high probability, then such a value of θ is favorable in the eyes of the data.

Posterior

The goal of Bayesian inference is to update our prior beliefs $P(\theta)$ by taking into account data X that we observe. The end result of

this inference procedure is the posterior distribution $P(\theta|X)$. Bayes' theorem specifies the way in which the posterior is computed,

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}.$$

Since in any particular inference problem, the data is fixed, we are often interested in only the terms which are functions of θ. Thus, the essence of Bayes' theorem is

$$P(\theta|X) \propto P(X|\theta)P(\theta),$$

or in words,

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior},$$

where all the terms above are viewed as functions of θ. Our final beliefs about θ combine both the relevant information we had *a priori* and the knowledge we gained *a posteriori* by observing data.

Coin Tosses

To get an understanding of what the Bayesian machinery looks like in action, let us return to our coin toss example. Suppose you just found a quarter lying on the sidewalk. You are interested in determining the extent to which this quarter is biased. More precisely, you wish to determine the probability p that the coin will come up heads. The most natural way to determine the value of p is to start flipping the coin and see what happens. So you flip the coin once and observe that the coin comes up heads. What should you conclude?

It is tempting to say that we cannot conclude anything from a single coin toss. But this is not quite true. The result of this toss tells us at the very least that $p \neq 0$, whereas before the toss it was certainly possible that $p = 0$ (perhaps both sides were tails). Furthermore, we should now be slightly more inclined to believe that p takes on larger values than we were before the toss. Which values we believe are reasonable depends on what our prior beliefs were. Most of the coins I have encountered in my life have been fair, or at least very close to fair. So my prior distribution on the value of p for any particular coin might look something like this.

www.dbooks.org

Regression Analysis

Linear regression is one of the most widely used tools in statistics. Suppose we were jobless college students interested in finding out how big (or small) our salaries would be 20 years from now. There's no way to pin down this number for sure, but we know that there are many factors that contribute to how much money a college graduate will make. For example, a naive observation (but a good starting point) is that students with higher GPAs earn more money 20 years from now. In this case, we assume that there is some true distribution that governs the behavior of the random variables

$$X \doteq \text{GPA}$$
$$Y \doteq \text{Salary 20 years from now}$$

where X and Y are not independent. In this case, we call X *a predictor* of Y. Another way that people refer to X and Y are as independent and dependent variables (nothing to do with *probabilistic* independence), since Y *depends* on X. In the following sections, we set up a linear model to describe the relationship between Y and X, which we can then use to predict our own future salary, based on some sample data.

Ordinary Least Squares

The Linear Model

Since X and Y seem to have some relationship, it would be reasonable to assume that given some value of X, we have a better idea about what Y is. Intuitively, we would expect students with higher GPAs to have a larger future salary, so we could model the relationship between X and Y using a line. That is, for some real numbers w_0 and w_1,

$$Y = w_0 + w_1 X$$

This is our familiar $y = mx + b$ relationship from high school algebra, but with different names for m and b.

www.dbooks.org

Note that this is an extremely simple model that is likely to miss most of the nuances in predicting someone's salary 20 years from now. There are in fact many more predictors than someone's GPA that affect their future salary. Also notice that we can express the above relationship using the following vector form.

$$Y = \mathbf{X} \cdot \mathbf{w} \doteq (1, X) \cdot (w_0, w_1)$$

where "$\cdot$" represents the dot product. This form is why the method is called *linear* regression.

Exercise 0.0.5. *Verify the function $f : ^2 \to$ defined by*

$$f(\boldsymbol{w}) = \boldsymbol{X} \cdot \boldsymbol{w}$$

is linear in $\boldsymbol{w}$.

Solution. Remember that the term *linear* was used to describe the "Expectation" operator. The two conditions we need to check are

(a) For any vectors $\mathbf{w}, \mathbf{v} \in ^2$, we have

$$f(\mathbf{u} + \mathbf{v}) = f(\mathbf{w}) + f(\mathbf{v})$$

(b) For any vector $\mathbf{w} \in ^2$ and constant $c \in$,

$$f(c\mathbf{w}) = cf(\mathbf{w})$$

To show (a), we know that $\mathbf{w}$ and $\mathbf{v}$ are vectors of the form

$$\mathbf{w} \doteq (w_0, w_1)$$
$$\mathbf{v} \doteq (v_0, v_1)$$

so that

$$\begin{aligned}
f(\mathbf{w} + \mathbf{v}) &= f((w_0, w_1) + (v_0, v_1)) \\
&= f((w_0 + v_0, w_1 + v_1)) \\
&= \mathbf{X} \cdot (w_0 + v_0, w_1 + v_1) \\
&= (1, X) \cdot (w_0 + v_0, w_1 + v_1) && \text{(Definition of } \mathbf{X}\text{)} \\
&= (w_0 + v_0) + X(w_1 + v_1) && \text{(Definition of dot product)} \\
&= (w_0 + Xw_1) + (v_0 + Xv_1) && \text{(Rearranging)} \\
&= \mathbf{X} \cdot \mathbf{w} + \mathbf{X} \cdot \mathbf{v} \\
&= f(\mathbf{w}) + f(\mathbf{v})
\end{aligned}$$

For (b), observe that if $\mathbf{w} \in^2$ and $c \in$,

$$\begin{aligned} f(c\mathbf{w}) &= \mathbf{X} \cdot (cw_0, cw_1) \\ &= (1, X) \cdot (cw_0, cw_1) \\ &= cw_0 + cw_1 X \\ &= c(w_0 + w_1 X) \\ &= c\mathbf{X} \cdot \mathbf{w} \\ &= cf(\mathbf{w}) \end{aligned}$$

This completes the proof. □

The observation that f is linear in $\mathbf{w}$ as opposed to linear in $\mathbf{X}$ is an extremely important distinction. Take a moment and let it sink in. This means that we can transform $\mathbf{X}$ in crazy nonlinear ways while maintaining the linearity of this problem. For example, the proof above implies that we could replace $\mathbf{X}$ with $\log(\mathbf{X})$ or $\sin(\mathbf{X})$ and we still have a linear relationship between Y and $\mathbf{w}$.

The above example not realistic in the sense that its extremely unlikely that if we sampled n college graduates and their actual salaries 20 years after college, all their GPAs fall on a perfect line when plotted against their salaries. That is, if we took n sample points, written

$$\text{Sample} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$$

and plotted these points in the plane with "GPA" on the x-axis and "Salary" on the y-axis, the points would almost surely not fall on a perfect line. As a result, we introduce an error term ϵ, so that

$$Y = \mathbf{X} \cdot \mathbf{w} + \epsilon \tag{5}$$

All of this hasn't yet told us how to predict our salaries 20 years from now using only our GPA. The subject of the following section gives a method for determining the best choice for w_0 and w_1 given some sample data. Using these values, we could plug in the vector $(1, \text{our GPA})$ for $\mathbf{X}$ in equation (2) and find a corresponding predicted salary Y (within some error ϵ).

Method of Least Squares

Our current model for X and Y is the relationship

$$Y = \mathbf{X} \cdot \mathbf{w} + \epsilon$$

where ϵ is some error term. Suppose we go out and ask a bunch of 50 year olds for their college GPAs and their salaries 20 years out of college. We can pair these quantities and record this sample data as

$$\text{Data} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

www.dbooks.org

Remember that we assume these samples come from the relationship

$$y_i = (1, x_i) \cdot (w_0, w_1) + \epsilon_i$$

and we are trying to find w_0 and w_1 to best fit the data. What do we mean by "best fit"? The notion we use is to find w_0 and w_1 that minimize the sum of squared errors $\sum_{i=1}^{n} \epsilon_i^2$. Rearranging the above equation for ϵ_i, we can rewrite this sum of squared errors as

$$E(\mathbf{w}) \doteq \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \mathbf{w})^2$$

where the vector $\mathbf{x}_i$ is shorthand for $(1, x_i)$. As we can see above, the error E is a function of $\mathbf{w}$. In order to minimize the squared error, we minimize the function E with respect to $\mathbf{w}$. E is a function of both w_0 and w_1. In order to minimize E with respect to these values, we need to take partial derivatives with respect to w_0 and w_1. This derivation can be tricky in keeping track of all the indices so the details are omitted. If we differentiate E with respect to w_0 and w_1, we eventually find that minimizing $\mathbf{w}$ can be expressed in matrix form as

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \left(\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

This can be written in the following concise form,

$$[box = \boxed{\ }align* \mathbf{w}^T = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{y}$$

where $\mathbf{D}$ is the matrix made by stacking the sample vectors $\mathbf{x}_i$,

$$\mathbf{D} \doteq \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

and $\mathbf{y}$ is the column vector made by stacking the observations y_i,

$$\mathbf{y} \doteq \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

A sketch of the derivation using matrices is given in the following section for those who cringed at the sentence "This derivation can be tricky in keeping track of all the indices so the details are omitted." Some familiarity with linear algebra will also be helpful going through the following derivation.

Linear Algebra Derivation

We can write the error function E as the squared norm of the matrix difference $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{D}\mathbf{w}^T$.

$$E(\mathbf{w}) = \|\mathbf{y} - \mathbf{D}\mathbf{w}^T\|^2 = \|\mathbf{D}\mathbf{w}^T - \mathbf{y}\|^2$$

Differentiating with respect to $\mathbf{w}$, the two comes down from the exponent by the power rule, and we multiply by $\mathbf{D}^T$ to account for the chain rule. We get

$$\nabla E = 2\mathbf{D}^T(\mathbf{D}\mathbf{w}^T - \mathbf{y})$$

We set $\nabla E = \mathbf{0}$ (we use a bold "o" since it is actually a vector of zeros) so that

$$2\mathbf{D}^T(\mathbf{D}\mathbf{w}^T - \mathbf{y}) = \mathbf{0}$$

Dividing by 2 on both sides and distributing the $\mathbf{D}^T$ across the difference gives

$$\mathbf{D}^T\mathbf{D}\mathbf{w}^T - \mathbf{D}^T\mathbf{y} = \mathbf{0}$$

Adding $\mathbf{D}^T\mathbf{y}$ to both sides gives

$$\mathbf{D}^T\mathbf{D}\mathbf{w}^T = \mathbf{D}^T\mathbf{y}$$

Multiplying on the left by the inverse of the matrix $\mathbf{D}^T\mathbf{D}$ on both sides of the above equation finally yields the famous linear regression formula,

$$\mathbf{w}^T = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{y}$$

Now, assuming salaries are related to college GPAs according to the relation

$$Y = w_0 + w_1 X + \epsilon,$$

we can plug in our GPA for X, and our optimal w_0 and w_1 to find the corresponding predicted salary Y, give or take some error ϵ. Note that since we chose w_0 and w_1 to minimize the errors, it is likely that the corresponding error for our GPA and predicted salary is small (we assume that our (GPA, Salary) pair come from the same "true" distribution as our samples).

Generalization

Our above example is a simplistic one, relying on the very naive assumption that salary is determined solely by college GPA. In fact

www.dbooks.org

there are many factors which influence someones salary. For example, earnings could also be related to the salaries of the person's parents, as students with more wealthy parents are likely to have more opportunities than those who come from a less wealthy background. In this case, there are more predictors than just GPA. We could extend the relationship to

$$Y = w_0 + w_1 X_1 + w_2 X_2 + w_3 X_3 + \epsilon$$

where X_1, X_2, and X_3 are the GPA, Parent 1 salary, and Parent 2 salary respectively.

By now it is clear that we can extend this approach to accomodate an arbitrary number of predictors $X_1, \ldots, X_d$ by modifying the relationship so that

$$Y = w_0 + w_1 X_1 + w_2 X_2 + \cdots + w_d X_d + \epsilon$$

or more concisely,

$$Y = \mathbf{X} \cdot \mathbf{w} + \epsilon$$

where the vectors $\mathbf{X}, \mathbf{w} \in^{d+1}$ are the extensions

$$\mathbf{X} \doteq (1, X_1, X_2, \ldots, X_d)$$
$$\mathbf{w} \doteq (w_0, w_1, w_2, \ldots, w_d)$$

the parameters w_i can be thought of as "weights" since the larger any particular weight is, the more influence its attached predictor has in the above equation. Recall that in Exercise 6.1, we verified the function $f(\mathbf{w}) = \mathbf{X} \cdot \mathbf{w}$ was linear in the vector $\mathbf{w} \in^2$. In fact, when we extend $\mathbf{w}$ to be a vector in $^{d+1}$, the function $f(\mathbf{w}) = \mathbf{X} \cdot \mathbf{w}$ is still linear in $\mathbf{w}$.

The linear regression formula still holds, i.e. that the optimal weights are given by

$$\mathbf{w}^T = (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T \mathbf{y}$$

where the matrix $\mathbf{D}$ is still constructed by stacking the observed samples,

$$\mathbf{D} \doteq \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(d)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(d)} \\ \vdots & & & & \\ 1 & x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(d)} \end{bmatrix}$$

where the i^{th} sample is written

$$\mathbf{x}_i \doteq (1, x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(d)})$$

Correlation

Throughout the past chapters, we often made the assumption that two random variables are independent in various exercises and methods. In reality, most random variables are not actually independent. In this section we give some measures to quantify how "related" a collection of random variables are.

The example in the Linear Regression chapter began with the observation that GPAs are positively correlated with future salaries. That is, we assumed that as college GPA increased, future salary also increased. Qualitatively, this was enough to motivate the problem of regression. However, there were other predictors that contributed to the future salary, some of which were also positively correlated to the projected salary. The fact that some variables contributed "more positively" than others was manifested in the size of the weights that were attached to the variables in the equation $Y = \mathbf{X} \cdot \mathbf{w} + \epsilon$. If one X_i were more predictive of Y than another, then its corresponding weight was larger. In the following section we examine the covariance of two random variables, which is another attempt to quantify the relationship between random variables.

Covariance

Suppose we have two random variables X and Y, not necessarily independent, and we want to quantify their relationship with a number. This number should satisfy two basic requirements.

(a) The number should be positive when X and Y increase/decrease together.

(b) It should be negative when one of X or Y decreases while the other increases.

Consider the following random variable.

$$(X - EX)(Y - EY)$$

Consider the possible realizations of the random variables $X = x$ and $Y = y$. The collection of these pairs is the sample space Ω. We can think of the outcomes of sampling an X and a Y as pairs $(x, y) \in \Omega$. Suppose the probability distribution governing X and Y on Ω assigns most of the probability mass on the pairs (x, y) such that $x > EX$ and $y > EY$. In this case, the random variable $(X - EX)(Y - EY)$ is likely to be positive most of the time. Similarly, if more mass were placed on pairs (x, y) such that $x < EX$ and $y < EY$, the product $(X - EX)(Y - EY)$ would be a negative number times a negative

www.dbooks.org

number, which means it would still be positive most of the time. Hence the product $(X - EX)(Y - EY)$ being positive is indicative of X and Y being mutually more positive or mutually more negative.

By similar reasoning, the product $(X - EX)(Y - EY)$ is more often negative if the distribution assigns more mass to pairs (x, y) that have $x < EX$ and $y > EY$, or that satisfy $x > EX$ and $y < EY$. In either case, the product $(X - EX)(Y - EY)$ will be a product of a positive and negative number, which is negative.

We are almost done. Remember at the beginning of this discussion we were searching for a number to summarize a relationship between X and Y that satisfied the requirements (a) and (b). But $(X - EX)(Y - EY)$ is a random variable, (that is, a function mapping Ω to) not a number. To get a number, we take the expectation. Finally we arrive at the definition of covariance.

Definition 0.0.19. *The **covariance** of two random variables X and Y, written $Cov(X, Y)$, is defined*

$$Cov(X, Y) = E[(X - EX)(Y - EY)]$$

This definition may look similar to the definition for variance of a random variable X, except we replace one of the terms in the product with the difference $Y - EY$. Similar to Proposition 2.11 (c), there is another useful form of the covariance.

Proposition 0.0.3. *Let X and Y be two random variables with means EX and EY respectively. Then*

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

Proof. By the definition of covariance, we can foil the product inside the expectation to get

$$\begin{aligned} \text{Cov}(X, Y) &= E[XY - XEY - YEX + EXEY] \\ &= E[XY] - E[XEY] - E[YEX] + E[EXEY] \quad \text{(linearity of } E\text{)} \\ &= E[XY] - EYEX - EXEY + EXEY \quad \text{(linearity of } E\text{)} \\ &= E[XY] - EXEY \end{aligned}$$

□

The Correlation Coefficient

The covariance quantity we just defined satisfies conditions (a) and (b), but can become arbitrarily large depending on the distribution of X and Y. Thus comparing covariances between different pairs of random variables can be tricky. To combat this, we normalize the quantity to be between -1 and 1. The normalized quantity is called the correlation, defined below.

Definition 0.0.20. *The **correlation coefficient** between two random variables X and Y with standard deviations σ_x and σ_y, is denoted ρ and is defined*

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

Exercise 0.0.6. *Verify that for given random variables X and Y, the correlation ρ_{xy} lies between -1 and 1.*

Heuristic. The rigorous proof for this fact requires us to view X and Y as elements in an infinite-dimensional normed vector space and apply the Cauchy Schwartz inequality to the quantity $E[(X - EX)(Y - EY)]$. Since we haven't mentioned any of these terms, we instead try to understand the result using a less fancy heuristic argument.

Given a random variable X, the first question we ask is,

What is the random variable *most positively* correlated with X?

The random variable that correlates most positively with X should increase *exactly* with X and decrease *exactly* with X. The only random variable that accomplishes this feat is X itself. This implies that the correlation coefficient between X and any random variable Y is less than that between X and itself. That is,

$$\rho_{xy} \leq \rho_{xx} = \frac{\text{Cov}(X,X)}{\sigma_x \sigma_x} = \frac{\text{Var}(X)}{\text{Var}(X)} = 1$$

By now you've probably guessed the second question we need to ask.

What is the random variable *least positively* correlated with X?

In other words, we are looking for a random variable with which the correlation between X and this random variable is the most negative it can be. This random variable should increase *exactly* as X decreases, and it should also decrease *exactly* as X increases. The candidate that comes to mind is $-X$. This would imply that the correlation coefficient between X and any random variable Y is greater than that between X and $-X$.

This implies that

$$\rho_{xy} \geq \rho_{x,-x} = \frac{\text{Cov}(X,-X)}{\sigma_x \sigma_{-x}}$$

By Proposition 6.3, the expression on the right becomes

$$= \frac{E[X(-X)] - E[X]E[-X]}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(-X)}} = \frac{-(E[X^2] - (EX)^2)}{\text{Var}(X)} = \frac{-\text{Var}(X)}{\text{Var}(X)} = -1$$

Hence, we conclude that $-1 \leq \rho_{xy} \leq 1$. □

www.dbooks.org

Interpretation of Correlation

The correlation coefficient between two random variables X and Y can be understood by plotting samples of X and Y in the plane. Suppose we sample from the distribution on X and Y and get

$$\text{Sample} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

There are three possibilities.

Case 1: $\rho_{xy} > 0$. We said that this corresponds to X and Y increasing mutually or decreasing mutually. If this is the case, then if we took n to be huge (taking many samples) and plotted the observations, the best fit line would have a positive slope. In the extreme case if $\rho_{xy} = 1$, the samples (X_i, Y_i) would all fall perfectly on a line with slope 1.

Case 2: $\rho_{xy} = 0$. This corresponds to X and Y having no *observable* relationship. However, this does not necessarily mean that X and Y have no relationship whatsoever. It just means that the measure we are using the quantify their relative spread (the correlation) doesn't capture the underlying relationship. We'll see an example of this later. In terms of the plot, the samples (X_i, Y_i) would look scattered on the ^{2} plane with no apparent pattern.

Case 3: $\rho_{xy} < 0$. We said that this case corresponds to one of X or Y decreasing while the other increases. If this were the case, then the best fit line is likely to have a negative slope. In the extreme case when $\rho_{xy} = -1$, all samples fall perfectly on a line with slope -1.

Independence vs Zero Correlation

There is a commonly misunderstood distinction between the following two statements.

1. "X and Y are independent random variables."

2. "The correlation coefficient between X and Y is 0."

The following statement is always true.

Proposition 0.0.4. *If X and Y are independent random variables, then $\rho_{xy} = 0$.*

The converse is not. That is, $\rho_{xy} = 0$ does not *necessarily* imply that X and Y are independent.

In "Case 2" of the previous section, we hinted that even though $\rho_{xy} = 0$ corresponded to X and Y having no observable relationship, there could still be some underlying relationship between the random variables, i.e. X and Y are still not independent. First let's prove Proposition 6.6

Proof. Suppose X and Y are independent. Then functions of X and Y are independent. In particular, the functions

$$f(X) \doteq X - EX$$
$$g(Y) \doteq Y - EY$$

are independent. By the definition of correlation,

$$\begin{aligned}
\rho_{xy} &= \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y} \\
&= \frac{E[(X-EX)(Y-EY)]}{\sigma_x \sigma_y} \\
&= \frac{E[f(X)g(Y)]}{\sigma_x \sigma_y} \\
&= \frac{E[f(X)]E[g(Y)]}{\sigma_x \sigma_y} && \text{(independence of } f(X) \text{ and } g(Y)) \\
&= \frac{0 \cdot 0}{\sigma_x \sigma_y} && (E[f(X)] = E(X - EX) = 0) \\
&= 0
\end{aligned}$$

Hence if X and Y are independent, $\rho_{xy} = 0$. □

Now let's see an example where the converse does not hold. That is, an example of two random variables X and Y such that $\rho_{xy} = 0$, but X and Y are *not* independent.

Example 0.0.11. *Suppose X is a discrete random variable taking on values in the set $\{-1, 0, 1\}$, each with probability $\frac{1}{3}$. Now consider the random variable $|X|$. These two random variables are clearly not independent, since once we know the value of X, we know the value of $|X|$. However, we can show that X and $|X|$ are uncorrelated. By the definition of correlation and Proposition 6.3,*

$$\rho_{x,|x|} = \frac{E(X \cdot |X|) - EX \cdot E|X|}{\sigma_x \sigma_{|x|}} \tag{6}$$

Let's compute the numerator. By looking at the distribution of X, we can see that the product $X \cdot |X|$ can only take on three possible values. If $X = 0$, then $|X| = 0$ so $X \cdot |X| = 0$. If $X = -1$, then $|X| = 1$ and $X \cdot |X| = -1$. Finally if $X = 1$, then $|X| = 1$ and $X \cdot |X| = 1$. Each of these cases occur with probability $\frac{1}{3}$. Hence,

$$X \cdot |X| \sim \mathit{Uniform}\{-1, 0, 1\}$$

It follows that the expectation of $X \cdot |X|$ is

$$E(X \cdot |X|) = \frac{1}{3} \cdot (-1) + \frac{1}{3} \cdot (0) + \frac{1}{3} \cdot (1) = 0.$$

www.dbooks.org

Also by the definition of expectation,

$$E[X] = \frac{1}{3} \cdot (-1) + \frac{1}{3} \cdot (0) + \frac{1}{3} \cdot (1) = 0.$$

Plugging these values into the numerator in expression (3), we get $\rho_{x,|x|} = 0$. *Thus, the two random variables* X *and* $|X|$ *are certainly not always equal, they are not independent, and yet they have correlation 0. It is important to keep in mind that zero correlation does not necessarily imply independence.*