TYLER DEVLIN, JINGRU GUO, DANIEL KUNIN, DAN XIANG

SEEING THEORY

A VISUAL INTRODUCTION TO PROBABILITY AND STATISTICS

BROWN UNIVERSITY

PUBLISHED BY BROWN UNIVERSITY

STUDENTS.BROWN.EDU/SEEINGTHEORY

First printing, August 2017

Contents

Introduction

The prerequisites of this text are basic algebra for the earlier chapters and basic calculus for the sections on continuous distributions. [1]

[1] Some familiarity with poker is also expected.

Basic Probability

Basic probability is an introduction to the foundational ideas in probability theory.

Likelihood

A **probability** is a number between 0 and 1 that describes how likely an event is to occur. When setting up an experiment, we first specify a set of outcomes, which we call the **sample space**, typically denoted with the symbol Ω. Collections of items in Ω are called **events**, and to these events, we associate likelihoods, or probabilities.

Let's see how to set up sample spaces in some familiar settings.

Sample Spaces for Coin and Dice Experiments

Example 0.0.1. Imagine we are flipping a coin. The set of possible outcomes is

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

i.e. we can flip either a "Heads" or a "Tails".

Now suppose we flip two coins. The set of possible outcomes becomes

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

Example 0.0.2. Now suppose we are rolling a die. The sample space of a die roll is similar to that of a coin flip, except we have more outcomes. Since a typical die has 6 sides, we can write the sample space as

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

An example of a **subset** A of Ω (we write $A \subset \Omega$) could be the set of even rolls $\{2, 4, 6\}$. [2]

[2] Should we use the terminology of "event" and "collection" instead of "subset" which we will introduce later?

Assigning Probabilities to Dice Rolls and Coin Flips

To each element of our sample space Ω, we associate a probability. We require that the sum of the probabilities of all the outcomes in Ω must be 1. To find the probability of a subset of our sample space, we would add up the probabilities of the elements contained in that subset.

Example 0.0.3. Assume that the die we rolled above was a fair die. Then each of the outcomes is equally likely to occur, i.e.

$$P(\text{roll } 1) = P(\text{roll } 2) = P(\text{roll } 3) = P(\text{roll } 4) = P(\text{roll } 5) = P(\text{roll } 6) = \frac{1}{6}$$

To find the probability of rolling an even number, we add up all the probabilities of the even numbers in Ω. This gives us

$$\begin{aligned} P(\text{roll an even number}) &= P(\text{roll } 2) + P(\text{roll } 4) + P(\text{roll } 6) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\ &= \frac{1}{2}. \end{aligned}$$

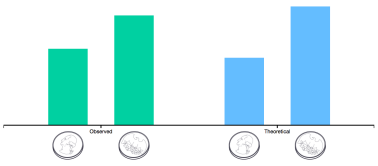

Figure 1: We should include screen shots directly from the visualizations

Example 0.0.4. Suppose that we flip a single fair coin. Then as stated above, the possible outcomes are `H` or `T`. Since it is equally likely to flip a `H` as a `T`, we have

$$P(\texttt{H}) = P(\texttt{T}) = \frac{1}{2}$$

Similarly, if we flip the coin twice, each of the possible outcomes should be equally likely, so

$$P(\texttt{HH}) = P(\texttt{HT}) = P(\texttt{TH}) = P(\texttt{TT}) = \frac{1}{4}$$

Now assume that our coin has a bias, that is, the probability of flipping `H` is some number between 0 and 1, denoted p. Then since the probabilities of all the outcomes must sum to 1, we have

$$\begin{aligned} P(\texttt{H}) &= p \\ P(\texttt{T}) &= 1 - p \end{aligned}$$

Similarly, the probabilities for the second sample space become

$$\begin{aligned} P(\texttt{HH}) &= p^2 \\ P(\texttt{HT}) &= p(1-p) \\ P(\texttt{TH}) &= (1-p)p \\ P(\texttt{TT}) &= (1-p)^2 \end{aligned}$$

Since we flip a `H` with probability p, the probability of then flipping another `H` in sequence would be p^2. The probability of flipping a `H` and then a `T` is the product of their probabilities, $p \cdot (1-p)$. The last two probabilities above are obtained similarly.

We check that these probabilities sum to 1 below.

$$\begin{aligned} P(\texttt{HH}) + P(\texttt{HT}) + P(\texttt{TH}) + P(\texttt{TT}) &= p^2 + p \cdot (1-p) + (1-p) \cdot p + (1-p)^2 \\ &= p^2 + (p - p^2) + (p - p^2) + (1 - 2p + p^2) \\ &= 2p - p^2 + (1 - 2p + p^2) \\ &= 1 \end{aligned}$$

Exercise 0.0.5. What is the probability that we get at least one H?

Solution. One way to solve this problem is to add up the probabilities 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 0.0.6. 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 0.0.7. Returning to our double coin flip example, our sample space was

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

Define the events

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

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

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

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.

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 0.0.8. 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 $\mathsf{H} \cdot \frac{1}{2} + \mathsf{T} \cdot \frac{1}{2}$? We would first need to assign *numerical values* to H and T in order to compute a meaningful expected value. For a coin flip we typically make the following assignments,

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

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 H, and in this case we say $X = 1$. Similarly, with probability $1 - p$ we flip T, and in this case we say $X = 0$. The expected value of the random quantity X is then

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

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 0.0.9. Show the following properties of expectation.

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

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

(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. □

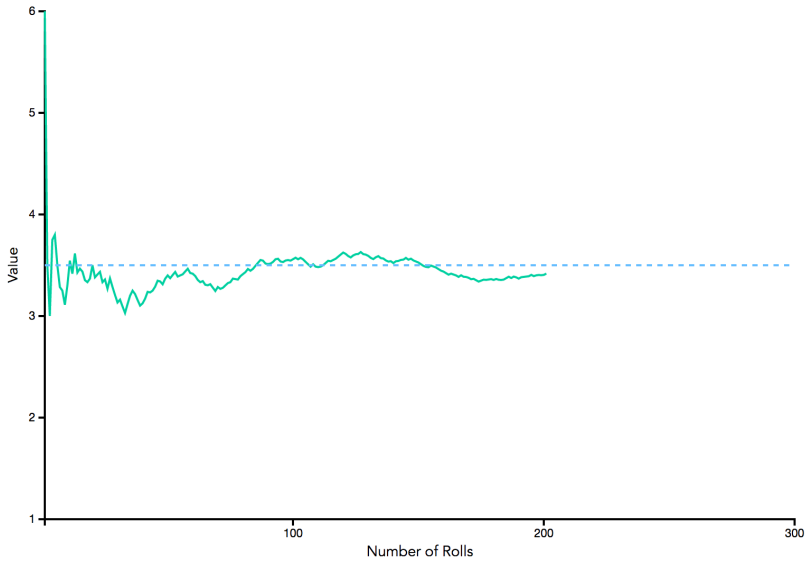

Figure 2: We can also include images in the main body of the text

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 0.0.10. The **variance** of X, denoted by $\text{Var}(X)$ is defined

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

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

Proposition 0.0.11. If X is a random variable with mean EX and $c \in R$ is a real number,

(a) $\text{Var}(X) \geq 0$.

(b) $\text{Var}(cX) = c^2\text{Var}(X)$.

(c) $\text{Var}(X) = E(X^2) - E(X)$.

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

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

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,

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

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.

□

Exercise 0.0.12. 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

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

□

Remark 0.0.13. 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 0.0.14 (Markov's inequality)**.** Suppose X is a nonnegative random variable and $a \in R$ 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 $\{k \in X(\Omega) \text{ s.t. } k \geq a\}$.[3]

[3] s.t. stands for "such that".

Dividing both sides by a, we recover

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

□

Corollary 0.0.15 (Chebyschev's inequality)**.** Let X be a random variable. Then

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

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

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

□

Compound Probability

Compound probability is the probability of joint occurrence of two or more simple events.

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 0.0.16. 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 0.0.17. **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.

Definition 0.0.18. 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 0.0.19. 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 0.0.20 (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.

□

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

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

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

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

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

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

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

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

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

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

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 0.0.22. 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 0.0.23. 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

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

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 0.0.24. 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.

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 0.0.25 (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 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%.

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 Variable

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

In the above example, our sample space was

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

and our random variable $X : \Omega \to R$, i.e. our function from the sample space Ω to the real numbers R, 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 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 0.0.27. 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 R.

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 and 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 0.0.28. The set of all natural numbers

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

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

Example 0.0.29. 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 0.0.30. 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

$$
\begin{aligned}
&0.1\,3\,5\,4\,2\,9\,5\ldots\\
&0.4\,2\,9\,4\,7\,2\,6\ldots\\
&0.3\,9\,1\,6\,8\,3\,1\ldots\\
&0.9\,8\,7\,3\,4\,3\,5\ldots\\
&0.2\,9\,1\,8\,1\,3\,6\ldots\\
&0.3\,7\,1\,6\,1\,8\,2\ldots\\
&\vdots
\end{aligned}
$$

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

$$a \doteq 0.121318\ldots$$

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 0.0.31. 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 0.0.32 (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 $k \in N$. The shorthand for stating such a distribution is $X \sim \text{Poi}(\lambda)$. Since k can be any number in N, our random variable X has a positive probability on infinitely many numbers. However, since 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 0.0.33 (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 0.0.34. 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.

Example 0.0.35. 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 0.0.36. A **probability density function** f (abbreviated **pdf**) is valid if it satisfies the following two properties.

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

Example 0.0.37 (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 $y \in R$, we have $f(x) \geq 0$ for all $x \in R$. 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 0.0.38 (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 0.0.39. 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 $a \in R$ results in another normal distribution, i.e. we have

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

(c) Adding a constant $a \in R$ 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

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

and

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

and

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

□

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

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

where we computed $\text{Var}(X_k) \approx 2.92$ 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 0.0.40 (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

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

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 0.0.41. 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

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

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

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

□

Statistical 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

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

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

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

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.

Types of Error

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

Definition 0.0.42. 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 0.0.43. 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.

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 0.0.44. The p-value, denoted p, is defined

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

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 0.0.45. 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,

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

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

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

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

$$P(\bar{X} > 0.7 \mid H_0) = P\Big(\frac{\bar{X} - 0.5}{S/\sqrt{n}} > \frac{0.7 - 0.5}{S/\sqrt{n}}\Big) \approx P\Big(Y > \frac{0.7 - 0.5}{S/\sqrt{n}}\Big) \doteq p \tag{1}$$

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

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

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 $\alpha \geq p$.

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

Linear Regression

Linear regression is an approach for modeling the linear relationship between two variables.

Ordinary Least Squares

Correlation

Analysis of Variance