Understanding Uncertainty Course Notes

Jeffrey Woo

2025-06-11

Contents

Preface

The examples in this preface is based on OpenIntro Statistics (Diez, Ceytinka-Rundel, Barr), Chapter 9.4 and 9.5, which provide more background information. You can access the book for free at https://www.openintro.org/book/os/

The main goal using data science is to understand data. Broadly speaking, this will involve building a statistical model for predicting, or estimating a response variable based on one or more predictors. Such models are used in a wide variety of fields such as finance, medicine, public policy, sports, and so on. We will look a couple of examples.

0.1 Examples

0.1.1 Example 1: Mario Kart Auction Prices

In this first example, we will look at Ebay auctions of a video game called Mario Kart that is played on Nintendo Wii. We want to predict the price of an auction based on whether the game is new or not, whether the auction's main photo is a stock photo, the duration of the auction in days, and the number of Wii wheels included with the auction.

A model that we can use for this example is the linear regression model:

```
library(openintro)

Data<-mariokart
##fit model
result<-lm(total_pr~cond+stock_photo+duration+wheels, data=Data)
```

Generally speaking, a linear regression equation takes the following form:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

where $\hat{y}$ denotes the predicted value of the response variable, the price of the action in this example, $x_1, x_2, \cdots, x_k$ denote the values of the predictors. This is example, we have: x_1 for whether the game is new or not, x_2 for whether the

auction's main photo is a stock photo, x_3 for the duration of the auction in days, and x_4 for the number of Wii wheels included with the auction. $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ represent the estimated regression parameters. If we know what these values are, we can easily plug in the values of the predictors to obtain the predicted price of the auction.

Fitting the model in R, we obtain the estimated regression parameters:

```
##get estimated regression parameters
result
```

```
##
## Call:
## lm(formula = total_pr ~ cond + stock_photo + duration + wheels,
##     data = Data)
##
## Coefficients:
##    (Intercept)        condused  stock_photoyes        duration          wheels
##        43.5201         -2.5816         -6.7542          0.3788          9.9476
```

so we have:

$$\hat{y} = 43.5201 - 2.5816x_1 - 6.7542x_2 + 0.3788x_3 + 9.9476x_4$$

So for an auction for Mario Kart game that is used, that uses a stock photo, is listed for 2 days, and comes with 0 wheels, the predicted price will be $\hat{y} = 43.5201 - 2.5816 - 6.7542 + 0.3788 \times 2 = 34.94$ or about 35 dollars.

0.1.2 Example 2: Job Application Callback Rates

In this example, we look at data from an experiment that sought to evaluate the effect of race and gender on job application callback rates. For the experiment, researchers created fake resumes to job postings in Boston and Chicago to see which resumes resulted in a callback. The fake resumes included relevant information such as the applicant's educational attainment, how many year's of experience the applicant as well as a first and last name. The names on the fake resume were meant to imply the applicant's race and gender. Only two races were considered (Black or White) and only two genders were considered (Make or Female) for the experiment.

Prior to the experiment, the researchers conducted surveys to check for racial and gender associations for the names on the fake resumes; only names that passed a certain threshold from the surveys were included in the experiment.

A model that can be used in this example is the logistic regression model

```
Data2<-resume
##fit model
result2<-glm(received_callback~job_city + college_degree+years_experience+race+gender,
```

Generally speaking, a logistic regression equation takes the following form

$$\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

where $\hat{\pi}$ denotes the predicted probability that the applicant receives a call back. $x_1, x_2, \cdots, x_k$ denote the values of the predictors. This is example, we have: x_1 for which city is the job posting located in, x_2 for whether the applicant has a college degree or not, x_3 for the experience of the applicant, x_4 for associated race of the applicant, and x_5 for the associated gender of the applicant. Similar to linear regression, $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ represent the estimated regression parameters. If we know what these values are, we can easily plug in the values of the predictors to obtain the predicted probability for an applicant with those characteristics to receive a callback.

Fitting the model in R, we obtain the estimated regression parameters

```
##get estimated regression parameters
result2
```

```
##
## Call:  glm(formula = received_callback ~ job_city + college_degree +
##     years_experience + race + gender, family = "binomial", data = Data2)
##
## Coefficients:
##      (Intercept)   job_cityChicago    college_degree  years_experience
##          -2.63974          -0.39206          -0.06550           0.03152
##        racewhite           genderm
##          0.44299          -0.22814
##
## Degrees of Freedom: 4869 Total (i.e. Null);  4864 Residual
## Null Deviance:       2727
## Residual Deviance: 2680  AIC: 2692
```

so we have

$$\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -2.63974 - 0.39206 x_1 - 0.0655 x_2 + 0.03152 x_3 + 0.44299 x_4 - 0.22814 x_5$$

So for an applicant in Boston, who has a college degree, has 10 years of experience and has a name that is associated with being a Black male, the logistic regression equation becomes $\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -2.63974 - 0.0655 + 0.03152 \times 10 - 0.22814 = -2.61818$. Doing a little bit of algebra to solve, we get $\hat{\pi} = 0.06797751$. Such an applicant has about a 6.8 percent chance of receiving a callback.

0.2 How were Estimated Parameters Calculated?

In the two examples, notice how I used some R functions, supplied the names of the variables, and the R functions generated the values of the estimated parameters? One thing you will learn is how the functions actually calculate these numbers. It turns out that these calculations are based on foundational concepts associated with measures of uncertainty, probability, and expected values. We will be learning about these concepts in this class.

Why do we want to know how these calculations are performed? So that we understand the intuition and logic behind how these models are built. It becomes a lot easier to work with these models when we understand their logic (for example, we know when these models can be used or cannot be used, we know what steps to take when we notice our data have certain characteristics, etc), instead of memorizing a bunch of steps.

When presenting models and data to people, some people may occasionally questions our methods and models. Why should we trust the model? Should we trust these numbers that seem to come out from some black box?

Notice we used two different models, linear regression and logistic regression, for examples 1 and 2. Why did we use these models? Could we have swapped the type of model used in these examples? The answer is actually no. One of the main considerations when deciding what model to use is to identify if our response variable is quantitative or categorical. You will learn why the linear regression model works when the response variable is quantitative, and why the logistic regression model works when the response variable is categorical.

0.3 The Course: Understanding Uncertainty

As mentioned in the previous section, we will be learning about foundational concepts associated with measures of uncertainty, probability, and expected values. All of these concepts will then help explain the intuition and how statistical models are built.

At the end of the course, we will apply these concepts and revisit the linear regression and logistic regression models. These are two of the most widely used models used in data science, as they are relatively easier to understand and explain. More modern methods (that you will learn about in future classes) such as decision trees and neural networks can be viewed as extensions of the linear and logistic regression models.

Chapter 1

Descriptive Statistics

This module is based on OpenIntro Statistics (Diez, Ceytinka-Rundel, Barr), Chapter 2.1. You can access the book for free at https://www.openintro.org/book/os/ Please note that I cover additional topics, and skip certain topics from the book.

1.1 Uncertainty with Data

When we are analyzing data, there is always going to be some degree of uncertainty, as there is randomness in a lot of phenomena that we observe in our world. An event is **random** if individual outcomes of the event are unpredictable. For example, the weight of the next baby born in a local hospital. Without knowing any information about the biological parents, we have a high degree of uncertainty if we try to predict this baby's weight. Even if we know detailed information about the biological parents (for example they are both very tall), we may feel more confident in predicting that the baby is likely to be heavier than average, but we cannot be certain about this prediction.

On the other end hand, an event is **deterministic** if we can predict individual outcomes of the event with certainty. For example, if we know the length of a cube is 2 inches, we know for sure that its volume is is $2^3 = 8$ cubic inches, based on rules of mathematics. The volume of a cube with length 2 inches is always going to be 8 cubic inches, so the volume is deterministic.

Thought question: think about data that you see in real life. Write these down. Are these data random or deterministic?

We will explore tools to help us quantify uncertainty in data. In this module, we will explore fairly standard tools that are used to describe data and give us an idea about the degree of uncertainty we have in the data. When describing

data that is quantitative, we usually describe the following: the shape of its distribution, its average or typical value, and its spread and uncertainty.

1.2 Visualizing Data

Data visualization is the representation of information in the form of pictures. Imagine have access to weights of all newborn babies at a local hospital. Examining each numerical value could be time consuming. So instead, we can use visualizations to give us an idea about the values of the weights. For example, what weights of newborns are common? What proportion of babies have dangerously low weights (which may indicate health risks)? Good data visualizations can give us such information fairly quickly. Next, we will explore some common visualizations that are used for quantitative (or numerical) variables.

1.2.1 Dot Plots

We will start with a **dot plot**, as it is the most basic visualization for a quantitative variable. We will use the `loan50` dataset from the `openintro` package. The data originally consist of thousands of loans made through the Lending Club platform, but we will randomly select 50 of these loans. Let us study the interest rate the loans the 50 applicants received.

```
library(tidyverse)
library(openintro)

##create object for data
Data<-loan50
```

For simplicity, we will round the numerical values of the interest rates to the nearest whole number:

```
##round interest rate to whole number
Data<- Data%>%
  mutate(r_int_rate = round(interest_rate))
```

We can create the corresponding dot plot, per Figure 1.1:

```
##dotplot
ggplot(Data,aes(x=r_int_rate))+
  geom_dotplot(binwidth=1)+
  theme(
    axis.text.y = element_blank(),  # Remove y-axis labels
    axis.title.y = element_blank(), # Remove y-axis title
    axis.ticks.y = element_blank()  # Remove y-axis ticks
  )+
  labs(x="Interest Rates (Rounded)")
```

Notice there is 1 black dot that corresponds to an interest rate of 20 (presumably

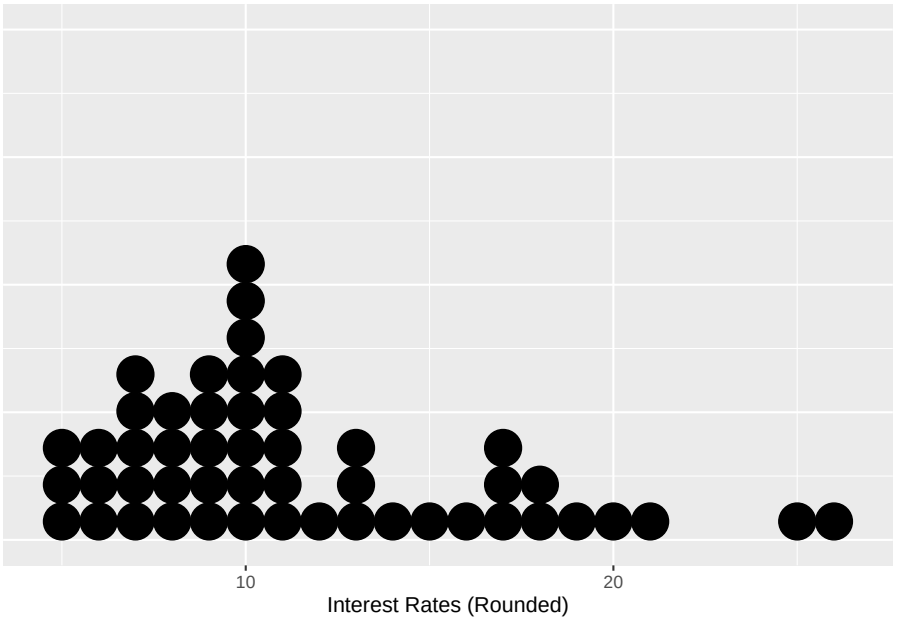

Figure 1.1: Dot Plot for 50 Interest Rates (rounded)

in percent), so there is one applicant who has a rounded interest rate of 20 percent. There are 8 black dots that correspond to an interest rate to 10 percent, so there are 8 applicants with a rounded interest rate of 10 percent. So interest rates of 10 percent are much more commonly occurring than interest rate of 20 percent. So we can use the height, or number of dots, to help us glean how often the value of a certain interest rate occurs. Based on this dotplot, interest rates between 5 and 11 percent are common, with higher values being less common.

Note: do not get too torn up about the details in the code to produce this dot plot. I have chosen the present the dot plot this way to highlight how we use it, without getting bogged down in the details of how it can be produced. We will not be using dot plots in this class.

1.2.2 Histograms

It turns out that dot plots are often not useful for large data sets, but they provide the general idea of how other visualizations for larger data sets work. The height of the dots inform us the frequency of those values occurring.

A visualization that is more commonly used for larger data sets is a histogram. Instead of displaying how common each value of the variable exists, we think of the values as belonging to a **bin** of values. For example, we can create a bin that contains interest rates between 5 and 7.5 percent, another bin containing

interest rates between 7.5 and 10 percent, and so on. A few things to note about histograms:

- By convention, values that lie exactly on the boundary of a bin will belong to the lower bin. For example, an interest rate that is exactly 12.5 percent will belong to the bin between 10 and 12.5 percent, and not the bin between 12.5 to 15 percent.
- Each bin should have the same width. In our example, the width is 2.5.

We create this histogram (using the original interest rates) below, per Figure 1.2:

```
##set up sequence to specify the bins
s25<-seq(5,27.5,2.5)

ggplot(Data,aes(x=interest_rate))+
  geom_histogram(breaks=s25,fill="blue",color="orange")+
  labs(x="Interest Rate", title="Histogram of Interest Rates")
```

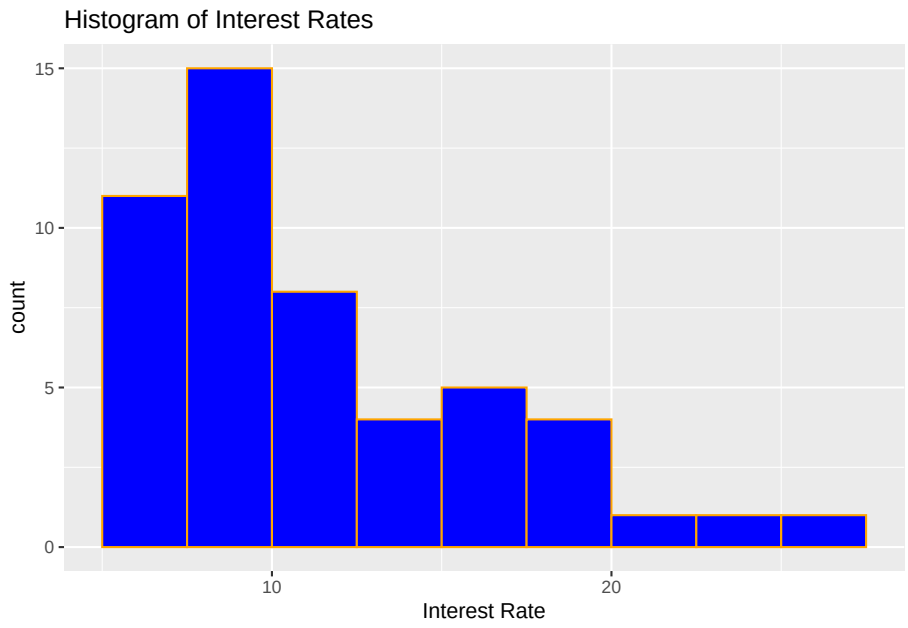

Figure 1.2: Historgram for 50 Interest Rates

Similar to the dot plot in Figure 1.1, the height of the histogram inform us what values are more commonly occurring. We can see from this histogram that interest rates between 5 and 10 percent are common, much more so than loans with interest rates greater than 20 percent. We could say that we have

more certainty that a randomly selected loan applicant will have an interest rate between 5 and 10 percent than an interest rate that is greater than 20 percent.

1.2.2.1 Shapes of Distribution

Histograms can also give us an idea about the **shape** of the distribution of interest rates. For the histogram in Figure 1.2, most of the loans are less than 15 percent, with only a small number of loans greater than 20 percent. We can say that we have greater certainty that a loan will have an interest rate less than 15 percent. When the data tail off to the right as in our histogram, the shape is said to be **right-skewed**. When a variable is said to be right-skewed, large values of the variable are much less common than small values of the variable; smaller values are more likely occur.

- If the histogram has the reverse characteristic, i.e. the data tail off to the left instead, the shape is said to be **left-skewed**. This implies that small values of the variable are much less common than large values of the variable; larger values are more likely to occur.
- Histograms that trail off similarly in both directions are called **symmetric**. Large and small values are of the variable are equally likely.
- Histograms that have a peak in the middle, and then trail off on both sides are not only symmetic, but also **bell-shaped**, or have a **normal** distribution. Note: it turns out one of the assumptions in linear regression is that the response variable follow a normal distribution. This may seem restrictive, however, we will see in later modules that this assumption is not particular crucial under some circumstances.

Thought question: Can you think of real life variables that have symmetric, right-skewed, left-skewed distributions? Feel free to search the internet for examples.

1.2.2.2 Considerations with Histograms

With our interest rate example, you may have noticed that I made a specific choice to the width of the bins when I created the histograms. It turns out that the width of the bins can impact the shape of the histogram, and potentially, how we interpret the histogram.

Consider creating a histogram with bin width of 0.5, instead of 2.5, per Figure 1.3:

```
##set up sequence to specify the bins. width now 0.5
s05<-seq(5,27.5,0.5)

ggplot(Data,aes(x=interest_rate))+
  geom_histogram(breaks=s05,fill="blue",color="orange")+
  labs(x="Interest Rate", title="Histogram of Interest Rates")
```

Comparing Figure 1.3 with Figure 1.2, note the following:

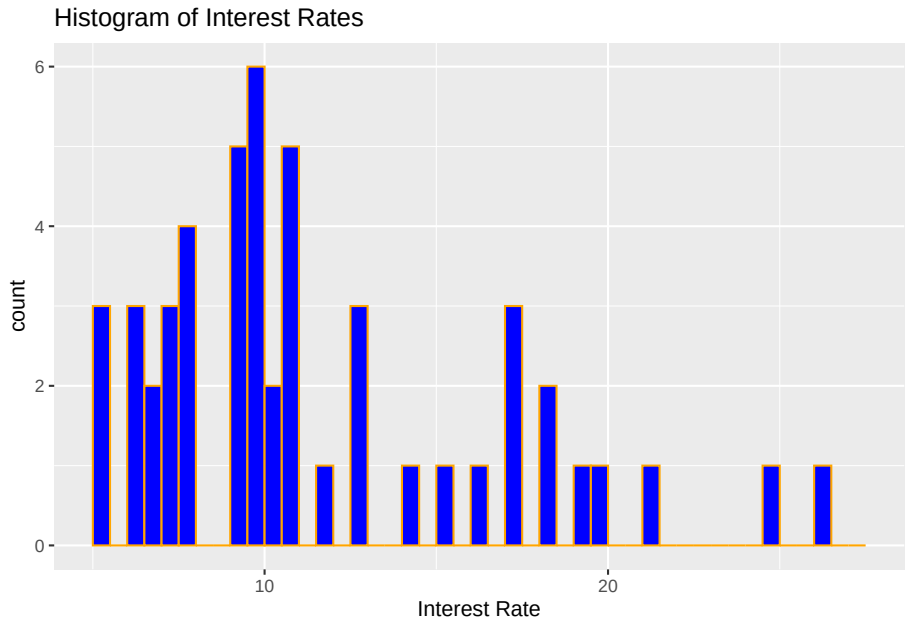

Figure 1.3: Historgram for 50 Interest Rates, with Bin Width 0.5

- Visually, the histogram looks more jagged with smaller bin width, whereas the histogram looks smoother with a larger bin width.
- Smaller bin widths may be preferred if we need information about smaller ranges of interest rates. However, it can be difficult to write about general trends.
- Larger bin widths may be more useful if we are trying to look for more general trends in the interest rates.

Thought question: What happens if we create a histogram with a bin width that is too large?

1.2.3 Density Plots

Another visualization for a quantitative variable is a density plot. A density plot can be viewed as a smoothed version of the histogram. We can use the heights to inform us about what values are more common. We create a density plot for the interest rates in Figure 1.4:

```
##density plot
plot(density(Data$interest_rate), main="Density Plot of Interest Rates")
```

Based on Figure 1.4, we see that low interest rates (between 5 and 12.5 percent)

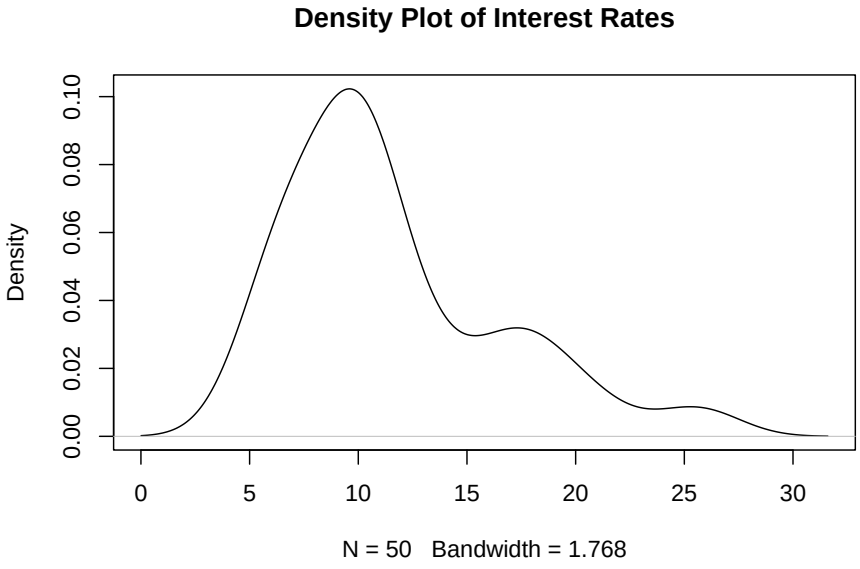

Figure 1.4: Density Plot for 50 Interest Rates

are much more common and high interest rates (higher than 20 percent). A few things to note about interpreting density plots:

- The area under the density plot is always equals to 1.
- To find the proportion of interest rates that are between two values, for example between 10 and 15 percent, we would integrate this density plot over this range, i.e. $\int_{10}^{15} f(x)dx$, where $f(x)$ is a mathematical equation that describes the density plot.
- The values on the vertical axis do not equal to probabilities (a common misconception).

The density plot is found using a method called kernel density estimation (KDE). We will over details about KDE in a later module as we need to cover quite a bit of material before doing so.

1.2.3.1 Considerations with Density Plots

Similar to bins and histograms, density plots are affected by the **bandwidth**. Larger bandwidths lead to smoother density plots, while smaller bandwidths lead to more jagged density plots. We create a density plot that uses a bandwidth that is twice the default in Figure 1.5 below:

```
plot(density(Data$interest_rate, adjust=2), main="Density Plot of Interest Rates, with
```

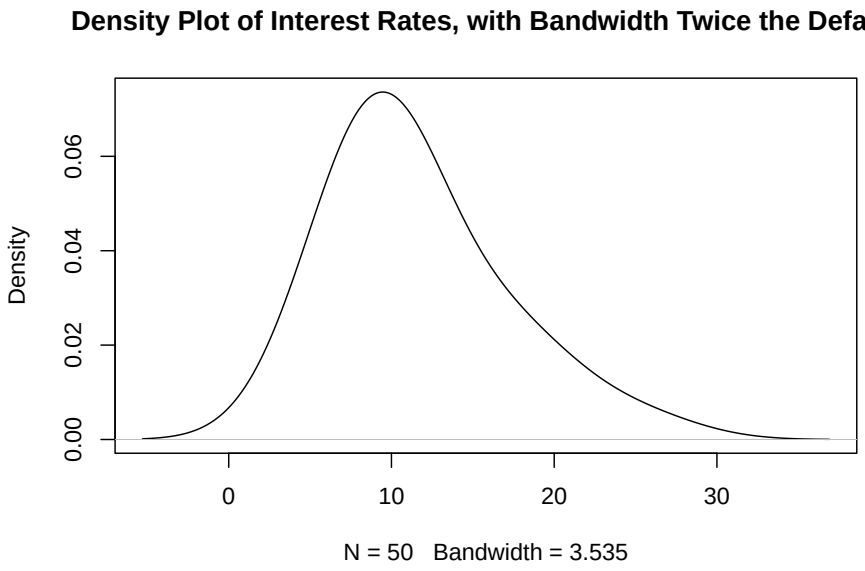

Figure 1.5: Density Plot for 50 Interest Rates with Larger Bandwidth

Notice in Figure 1.5 that the little peak for interest rates between 15 and 20 (which existed in Figures 1.4 and also 1.2) no longer exists. Using bandwidths that are too large can smooth out some of these peaks.

Thought question: What if we create a density plot with a bandwidth that is too small?

Thought question: How are bin widths for histograms and bandwidths for density plots related?

1.3 Ordered Statistics

The idea behind ordered statistics is pretty self-explanatory: take your numerical variable, and order the values from smallest to largest. Going back to our example of the interest rates from 50 loan applicants, let X denote the interest rate. Then $x_{(1)}$ will denote the interest rate that is the smallest, $x_{(2)}$ denotes the second smallest interest rate, and $x_{(50)}$ denotes the largest interest rate in our sample of 50.

1.3.1 Quantiles

Quantiles partition the range of numerical data into continuous intervals (groups) with (nearly) equal proportions. Common quartiles have their own names:

- Quartiles: 4 groups
- Percentiles: 100 groups

There is one less quantiles than the number of groups. We will go over quartiles in more detail.

1.3.1.1 Quartiles

Quartiles divide the data into 4 groups, and each group has (nearly) equal number of observations. So there will be three quartiles, denoted by Q_1, Q_2, Q_3.

- The first group will have values between negative infinity and Q_1.
- The second group will have values between negative Q_1 and Q_2.
- The third group will have values between negative Q_2 and Q_3.
- The fourth group will have values between negative Q_3 and infinity.

Q_2, sometimes called the second quartile, is the easiest value to find. It is also called the **median** of the data. Going back to our interest rates from the 50 loan applicants. Using our ordered statistics, the median is the middle observation. Since we have an even number of observations, we have two middle observations, $x_{(25)}$ and $x_{(26)}$. In this situation, the median will be the average of these two middle observations. Using R, we find the median to be:

```
median(Data$interest_rate)
```

```
## [1] 9.93
```

So half the interest rates are less than 9.93 percent, and half the interest rates are greater than 9.93 percent. You might also recognize another term for the median: the 50th percentile, as 50 percent of the interest rates are less than 9.93.

To find the middle observation(s) based on a sample of size n:

- If n is even, the 2 middle observations will be position $\frac{n}{2}$ and $\frac{n}{2}+1$ in the ordered statistics.
- If n is odd, the middle observation will be position $\frac{n}{2}+0.5$ in the ordered statistics.

Q_1 and Q_3 (also called the first and third quartiles) are found together, after finding Q_2. Note that Q_2 divides the data into two groups. Using our interest rates example, one group contains $x_{(1)}, \cdots, x_{(25)}$, and another group contains $x_{(26)}, \cdots, x_{(50)}$. Q_1 is the median of the first group, and Q_3 is the median of the second group. So for our 50 loan applicants:

- Q_1 is $x_{(13)}$, and

- Q_3 is $x_{(38)}$.

To find these values in R, we could type:

```
quantile(Data$interest_rate, prob=c(0.25,0.75), type = 1)
```

```
##   25%   75%
##  7.96 14.08
```

So Q_1 is 7.96 percent, and Q_3 is 14.08 percent. It turns out that Q_1 is also the 25th percentile, and Q_3 is also the 75th percentile, by definition.

Remember we wrote the following earlier:

- The first group will have values between negative infinity and Q_1. So about a quarter of observations are have interest rates less than 7.96 percent.
- The second group will have values between negative Q_1 and Q_2. So about a quarter of observations have interest rates between 7.96 and 9.93 percent.
- The third group will have values between negative Q_2 and Q_3. So about a quarter of observations have interest rates between 9.93 and 14.08 percent.
- The fourth group will have values between negative Q_3 and infinity. So about a quarter of observations have interest rates above 14.08 percent.

Note: you may notice that we used `type = 1` inside the `quantile()` function. Using `type = 1` gives the values of the first and third quartiles that are based on the method that was just described. There are actually several ways to find quantiles, which may result in slightly differing values, although they all generally meet the definition that Q_1 is the 25th percentile, and Q_3 is the 75th percentile.

1.3.1.2 Percentiles

Another common quantile is the percentile. In general the **k-th percentile** is the value of the data point below which k percent of observations are found. So in our earlier example, we said that Q_3 of the interest rates is 14.08 percent, and this is also the 75th percentile. So 75 percent of interest rates are less than 14.08 percent.

We will not go over the details of finding percentiles by hand.

1.3.2 Box Plots

Another visualization used to summarize quantitative data is the box plot. A **box plot** summarizes the 5-number summary. The 5 numbers are the minimum, Q_1, Q_2, Q_3, and the maximum. Using our interest rate data, the box plot is shown in Figure 1.6:

```
##box plot
ggplot(Data,aes(y=interest_rate))+
```

```
  geom_boxplot()+
  labs(y="Interest Rate", title="Box Plot of Interest Rates")
```

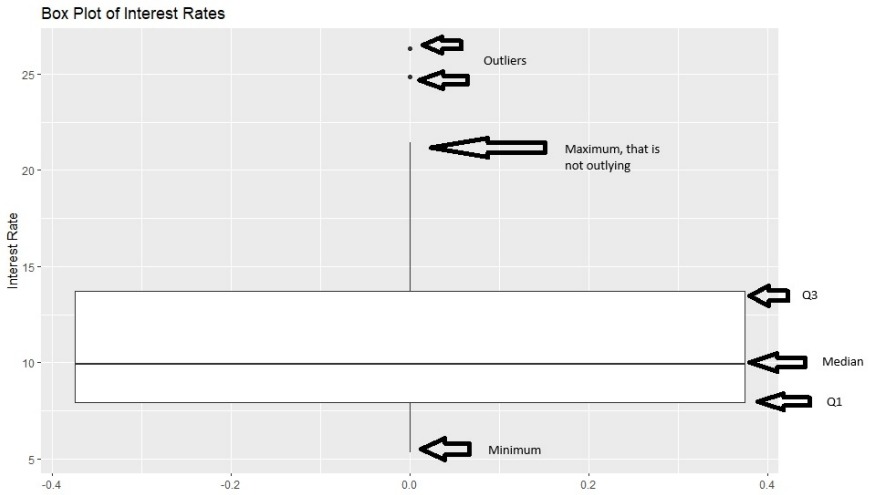

Figure 1.6: Box Plot of Interest Rates

Some people call a box plot a box and whisker plot.

- The boundaries of the box represent Q_1 and Q_3.
- The thick line in the box represents the median.
- The two whiskers on either side of the box extend to the minimum and maximum, if outliers do not exist. If outliers exist, the whiskers extend to the minimum and maximum values that are not outliers.

Generally, when we have one quantitative variable, an outlier is an observation whose numerical value is far away from the rest of the data. In other words, it is a lot smaller or larger relative to the rest of the data.

So for our 50 loans, there are two loan applicants with interest rates around 25 percent that are flagged as being a lot larger than the rest of the loans, which is reasonable since most of the loans are a lot smaller than 20 percent.

We will not go over the details of how outliers are determined in box plots. If you are interested, you can read Chapter 2.1.5 from OpenIntro Statistics (Diez, Ceytinka-Rundel, Barr).

Notice how much further large values (Q_3 and maximum) are from the median, compared to the distance of the small values (Q_1 and minimum) from the median. This indicates that the distribution of interest rates are right-skewed. Compare the boxplot of the interest rates in Figure 1.6 with its corresponding histogram (Figure 1.2) and density plot (Figure 1.4).

Thought question: can you sketch a box plot that represents a variable that is left-skewed? How about a variable that is symmetric?

1.3.3 Empirical Cumulative Distribution Function

From the previous sections, we can see how we could use histograms, density plots, and box plots to inform us about what proportion of observations take certain values, and the values of the data that correspond to certain percentiles. However, we are limited to quartiles and not any percentile when using box plots, and we need to find areas under the density plot (using integration, not a trivial task), or add up frequencies on a histogram (can be time consuming).

A plot that can easily give us values of the variable that correspond to percentiles is the **empirical cumulative distribution function (ECDF)** plot.

Let X denote a random variable, and we have observed n observations of X denoted by $x_1, \cdots, x_n$. Let $x_{(1)}, \cdots x_{(n)}$ denote the ordered statstics of the n observations. The ECDF, denoted by $\hat{F}_n(x)$ is the proportion of sample observations less than or equal to the value x of the random variable. Mathematically, the ECDF is:

$$\hat{F}_n(x) = \begin{cases} 0, & \text{for } x < x_{(1)} \\ \frac{k}{n}, & \text{for } x_{(k)} \leq x < x_{(k+1)}, k = 1, \cdots, n-1 \\ 1, & \text{for } x \geq x_{(n)}. \end{cases}$$

We shall use a simple toy example to illustrate how an ECDF is constructed. Suppose we ask 5 people how many times to go to the gym (at least 20 minutes) in a typical work week. The answers are: 3, 0, 1, 5, 3. The random variable X is how many times a person goes to the gym for at least 20 minutes, and the ordered statistics are $x_{(1)} = 0, x_{(2)} = 1, x_{(3)} = 3, x_{(4)} = 3, and x_{(5)} = 5$. Using the mathematical definition for the ECDF, we have:

- $\hat{F}_n(x) = 0$ for $x < x_{(1)} = 0$.
- $\hat{F}_n(x) = \frac{1}{5}$ for $0 \leq x < x_{(2)} = 1$.
- $\hat{F}_n(x) = \frac{2}{5}$ for $1 \leq x < x_{(3)} = 3$.
- $\hat{F}_n(x) = \frac{4}{5}$ for $3 \leq x < x_{(5)} = 5$. This value is special for this example since we have two observations where $x = 3$.
- $\hat{F}_n(x) = 1$ for $x \geq 5$.

The corresponding ECDF plot is shown in Figure 1.7:

```
##toy data
y<-c(3, 0, 1, 5, 3)
##ECDF plot
plot(ecdf(y), main = "ECDF for Toy Example")
```

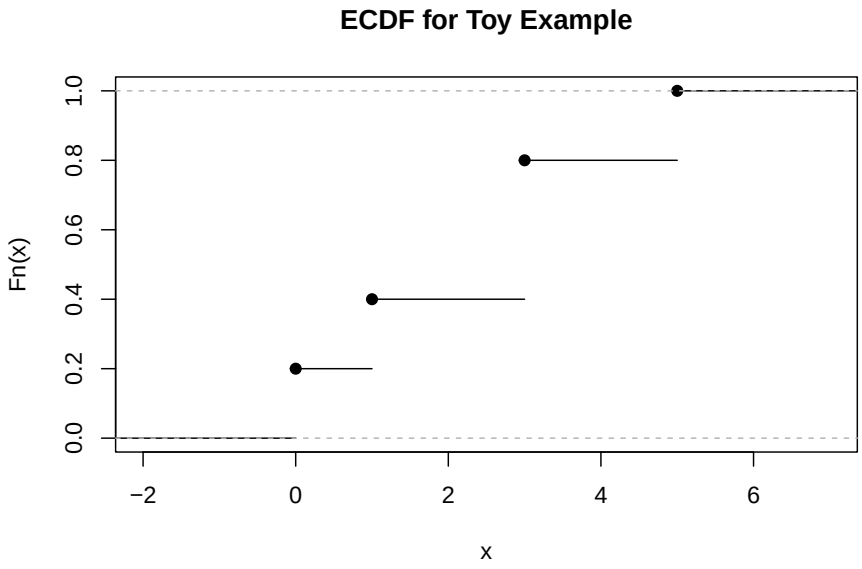

Figure 1.7: ECDF Plot for Toy Example

We can easily find percentiles from this plot, for example, the 40th percentile is equal to 1, going to the gym once a week. About 20 percent of observations go to the gym less than 1 time a week.

Next, we create the ECDF plot for the interest rates from the 50 loan applicants.

```
plot(ecdf(Data$interest_rate), main = "ECDF Plot of Interest Rates")
abline(h=0.8)
```

I overlaid a horizontal line for the 80th percentile, so we can read on the horizontal axis that this corresponds to an interest rate of about 17 percent. So about 80 percent of loan applicants have an interest rate less than 17 percent.

Thought question: try using the histogram and density plot for the interest rates (Figures 1.2 and 1.4) to find the interest rate that corresponds to the 80th percentile. Was this easy to perform?

1.4 Measures of Centrality

So far, we have used visualizations to summarize the shape of the distribution of a quantitative variable. Next, we look at common measures of centrality. Loosely speaking, measures of centrality are measures that describe the average or typical value of a quantitative variable. The common measures of centrality

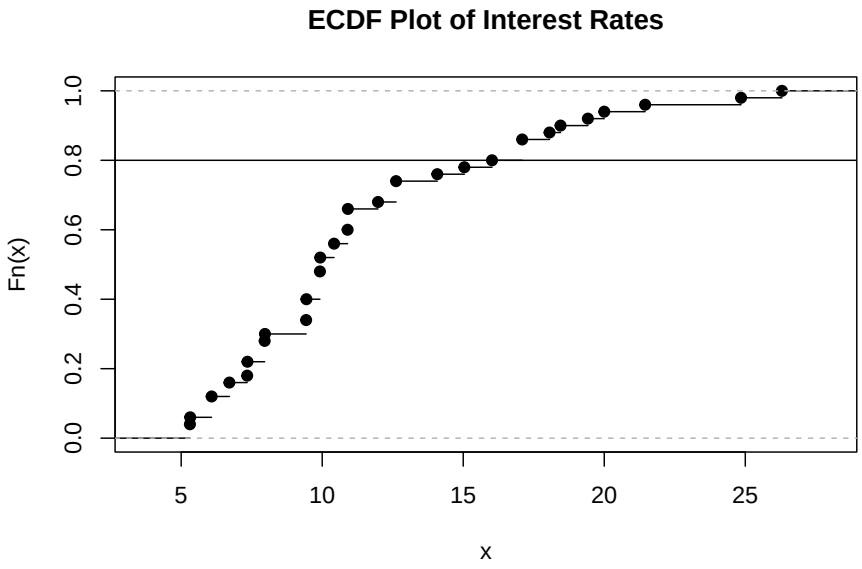

Figure 1.8: ECDF Plot of Interest Rates

are the mean, median, and mode.

1.4.1 Mean

The sample **mean** is simply the average value of the variable in our sample. The sample mean for a random variable X is denoted by $\bar{x}$, and is found by:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{1.1}$$

So, for our toy example of the 5 people and how often they go to the gym in a week, their sample mean is $\bar{x} = \frac{3+0+1+5+3}{5} = 2.4$.

1.4.2 Median

We went over how to find the median in section 1.3.1.1. The **median** is the value of the middle observation in ordered statistics. It is also called Q_2, the second quartile, and the 50th percentile, so approximately 50 percent of observations have values smaller than the median.

So, for our toy example of the 5 people and how often they go to the gym in a week, their sample median is $x_{(3)} = 3$. So about 50 percent of people went to gym less than 3 times in a week.

1.4.3 Mode

Another measure is the mode. Mathematically speaking, the **mode** is the most commonly occurring value in the data. So for our toy example, the mode is 3, since 3 occurs twice and occurs the most often in our data.

1.4.4 Considerations

A few things to consider when using these measures of centrality:

- The mean is a measure that most people are comfortable with, however, caution needs to be used if the variable is skewed, as extreme outliers and drastically alter the value of the mean. Using our toy example with the gym, suppose the person who visits the gym the most visits 50 times, instead of 5. The numerical value of the sample mean explodes, and does not give a good representation of the central value of how many visits to the gym a person makes in a week. The mean is fine if the variable is symmetric.
- The median is a measure that is recommended for skewed distributions, since the order associated with ordered statistics is not influenced by extreme outliers. Using the gym example, in the previous bullet point, the median is unaffected.
- The mean being larger than the median is an indication that the distribution is right-skewed. Using our interest rate example, we have:

```
mean(Data$interest_rate)
```

```
## [1] 11.5672
```

```
median(Data$interest_rate)
```

```
## [1] 9.93
```

which is consistent with the right skew we saw in the histogram and density plot in Figures 1.2 and 1.4. Conversely, a left-skewed distribution usually has a mean that is smaller than the median. A symmetric distribution typically has similar values for the mean and median.

- The mean is considered a **sensitive** measure, since its numerical value can be drastically affected by outliers. The median is considered a **robust** measure, since its numerical value is more resistant and is less affected by outliers.
- The mathematical definition of mode can be difficult to use for variables that are continuous, since it is likely that there are no observations that have the same value when the variable is continuous. In this instance, the mode typically refers to the bin in the histogram that has is the tallest. So, using the histogram in Figure 1.2 for the interest rates, the mode is between 7.5 to 10 percent.

1.5 Measures of Spread

In the previous sections, we learned about summarizing features a quantitative variable, by using visualizations to summarize its shape, and by using some measures of centrality that describe the average or typical values of the variable. One more feature we can summarize is the spread, associated with the values of a quantitative variable. Measures of spread are considered a way to measure uncertainty. Data that have larger spread have more uncertainty.

1.5.1 Variance and Standard Deviation

One measure of spread is the variance. The sample **variance** for a random variable X is denoted by s^2, or sometimes s_x^2, and is found by:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \tag{1.2}$$

The variance can be interpreted as the approximate average squared distance of the observations from the mean. The formula in equation (1.2) may look a bit complicated, but let us use the toy example where we asked 5 people how often they go to the gym in a workweek. The answers are: 3, 0, 1, 5, 3, and we had earlier found the sample mean to be $\bar{x} = 2.4$. To calculate the sample variance:

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \\ &= \frac{(3-2.4)^2 + (0-2.4)^2 + (1-2.4)^2 + (5-2.4)^2 + (3-2.4)^2}{5-1} \\ &= 3.8 \end{aligned}$$

Notice what we did in the numerator of equation (1.2): we take the difference between each observed value from the sample mean, square these differences, then add up the squared differences. We then divide by $n-1$, rather than n, hence the sample variance being the approximate averaged squared distance of the observations from the mean. There is some nuance in the mathematics as to why we divide by $n-1$ instead of n, and may not be intuitive as to why we do so. It turns out dividing by $n-1$ makes the sample variance an unbiased estimator of the true variance in the population (denoted by σ^2) and is more reliable than if we had divided by n. We will go over this in more detail in a later module after covering a few additional concepts.

Larger values of the sample variance indicate that the observations are generally further away from the sample mean, indicating larger spread, and a higher degree of uncertainty about future values.

Thought question: What does it mean if the sample variance of a set of observations is 0? Why does this indicate their there is little (or no) uncertainty about the set of observations?

Another related measure is the sample **standard deviation**, which is the square root of the sample variance. Similar to the variance, larger values indicated more spread in the data.

1.5.2 Interquartile Range

Another measure of spread is the **interquartile range (IQR)**, and it is the difference between the third and first queartiles,

$$IQR = Q_3 - Q_1. \tag{1.3}$$

The IQR is considered a robust measure of spread, while the sample variance and standard deviations are considered to be sensitive.

Chapter 2

Probability

This module is based on Introduction to Probability (Blitzstein, Hwang), Chapters 1 and 2. You can access the book for free at https://stat110.hsites.harvard.edu/ (and then click on Book). Please note that I cover additional topics, and skip certain topics from the book. You may skip Sections 1.4 and 1.5, and examples 2.4.5 and 2.5.12 from the book.

2.1 Probability

A way of quantifying uncertainty is through probability. Think about these statements: "I am 100% certain that it will rain in the next hour" and "I am 50% certain that it will rain in the next hour". The percentages are used to reflect the degree of certainty about the event happening. The first statement reflects certainty; the second reflects uncertainty as the statement implies the belief that it is equally likely that it will rain or not. In this module, we will learn about the basic concepts about probability.

2.1.1 Why Study Probability?

The book (Section 1.1) lists 10 different applications of probability, and there are many more applications. I will go as far as to say that anything that deals with data will also deal with probability.

2.1.2 Frequentiest and Bayesian View of Probability

There are a couple of main viewpoints on how to interpret probability: **frequentist** and **Bayesian**. Consider the statement that "if we flip a fair coin, the coin has a 50% chance of landing heads".

- The frequentist viewpoint views probability as the relative frequency associated with an event that is repeated for an infinite number of times.

It will interpret the 50% probability as: if we were to flip the coin many many times, 50% of these times will result in the coin landing heads.

- The Bayesian viewpoint views probability as a measure of belief, or certainty, that an event will happen. It will interpret the 50% probability as: heads and tails are equally likely to occur with a coin flip.

In this coin flip example, both interpretations are reasonable. However, in some instances, the frequentist interpretation may not be as interpretable if we cannot repeat the event many times. For example, the earlier statement about rain: "I am 50% certain that it will rain in the next hour". Whether it will rain or not in the next hour is not a repeatable event, so the frequentist interpretation makes less sense here.

2.2 Key Concepts in Probability

In this section, we will cover the basic terminology and foundational ideas in probability.

2.2.1 Sample Space

The **sample space** of an experiment, denoted by S, is the set of all possible outcomes of an experiment.

For the rest of this module, we will use the following as an example: consider a standard deck of 52 cards, and we draw one card at random. What is the card drawn? The sample space for this experiment can be viewed as a list of all 52 cards, per Figure 2.1 below.

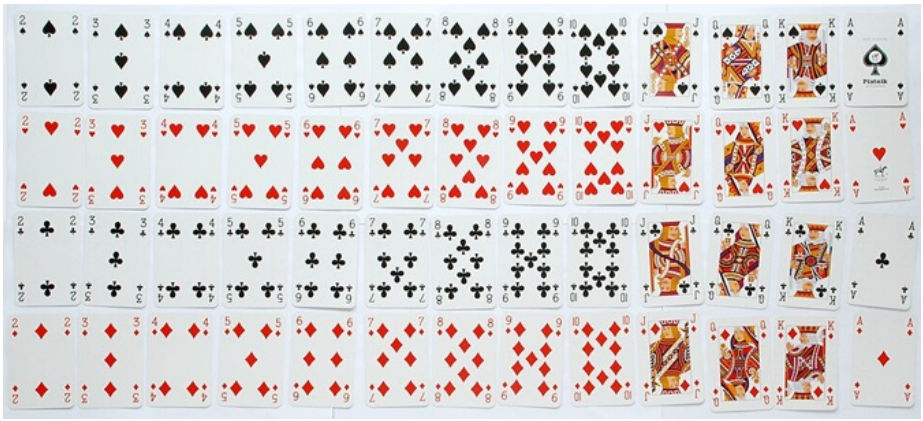

Figure 2.1: Sample Space of Drawing One Card from Standard Deck. Picture from https://en.wikipedia.org/wiki/Standard_52-card_deck

While the definition of sample space may appear elementary, writing out the

sample space is almost always the first step in performing any probability calculations.

2.2.2 Event

An **event** is a subset of the sample space, and is usually denoted by an upper case letter. For example, let A denote the event that I draw a card with a black suit (spades or clubs), and let B denote the event I draw a picture card (Jack, Queen, or King). Events A and B are each shown in Figures Figure 2.2 and Figure 2.3 below.

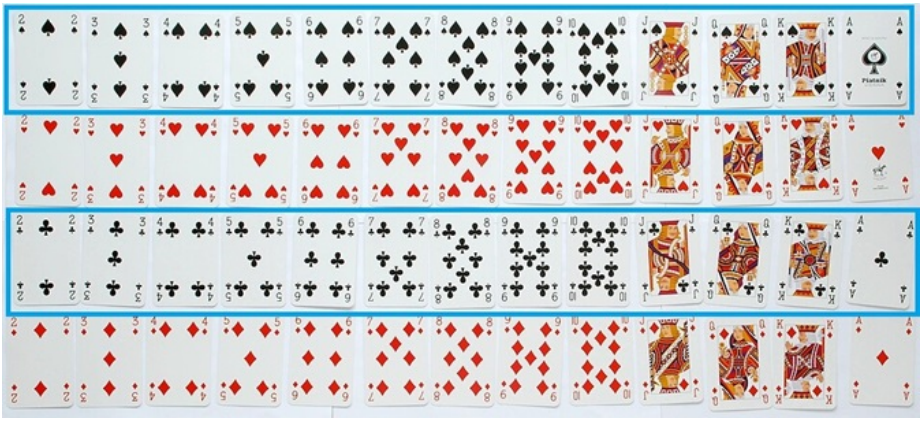

Figure 2.2: Event A (in Blue)

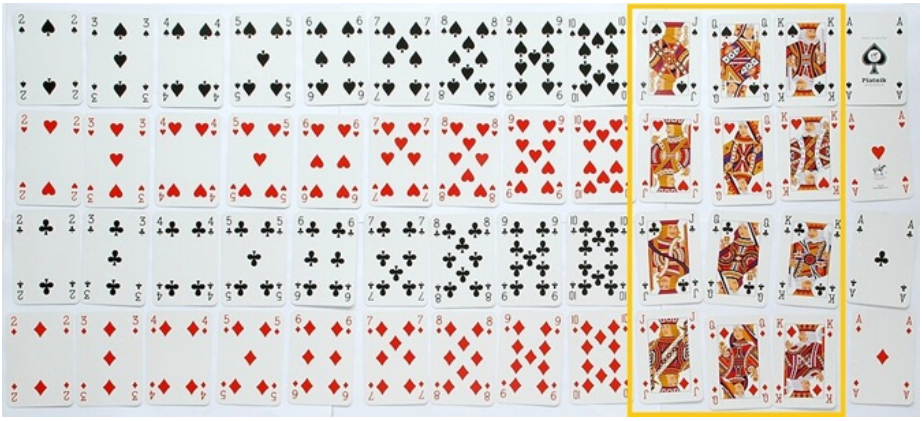

Figure 2.3: Event B (in gold)

The sample space of the experiment can be finite or infinite. In our card example, our sample space is finite since we can actually write out all possible outcomes.

If the number of possible outcomes is infinite (i.e. we cannot write out the entire list of all possible outcomes), the sample space is infinite.

We assign a probability to each event. The probability of event A happening is $P(A)$. **If each outcome of a sample space is equally likely and we have a finite sample space, the probability of the event is the number of outcomes belonging to the event divided by the number of outcomes in the sample space.** Using our card example, $P(A) = \frac{26}{52} = \frac{1}{2}$ and $P(B) = \frac{12}{52} = \frac{3}{13}$.

2.2.3 Complements

The **complement** of an event is the set of all outcomes that do not belong to the event. For example, the complement of A, denoted by A^c, will be drawing a card with a red suit (hearts or diamonds). One way to think about complements is that the complement of an event is the event not happening. Loking at Figure 2.2, this will be the cards that are not outlined in blue. In this example, $P(A^c) = \frac{26}{52} = \frac{1}{2}$.

Thought question: What is the probability of drawing a non picture card?

From these examples, you might realize the probability associated with the complement of an event can be found by subtracting the probability of the event from 1, i.e.

$$P(A^c) = 1 - P(A). \tag{2.1}$$

Sometimes, the calculation for the probability of the complement of an event is much less tedious than the probability of the event. In such an instance, equation (2.1) will be useful.

2.2.4 Unions

The **union** of events is when **at least one** of the events happen. For example, the union of events A and B, denoted by $A \cup B$, is the event that the card drawn is either a black suit, or a picture card, or both a black suit and a picture card. This is reflected in Figure 2.4.

To find $P(A \cup B)$, we can refer to Figure 2.4 and just count the number of outcomes to belong to either event A (is black suit) or event B (is picture card), and find this is $\frac{32}{52}$.

The union of A and B can be viewed as the event where either event A or B (or both) happens.

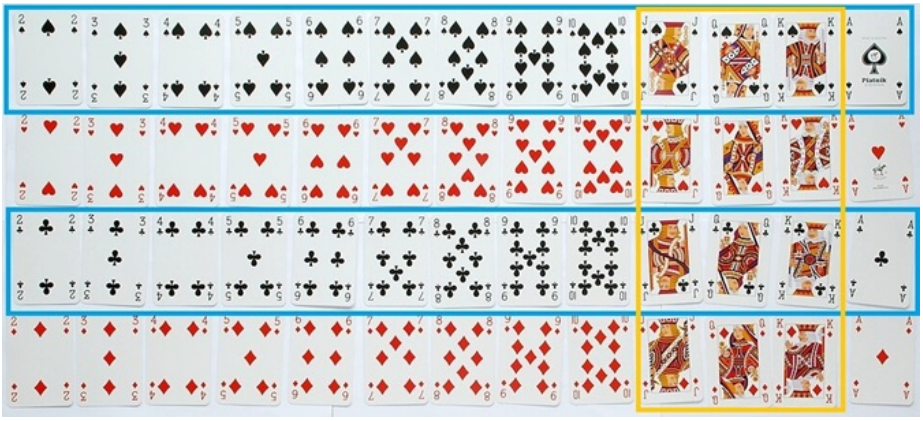

Figure 2.4: Union of A, B (in blue or gold, or both blue and gold)

2.2.5 Intersections

The **intersection** of events is when **all** of the events happen. Using our example, the intersection of events A and B is denoted by $A \cap B$, is the event that the card drawn is both a black suit and a picture card. Using Figure 2.4, the outcomes belonging to $A \cap B$ are the cards that are outlined in blue and gold. This probability is $P(A \cap B) = \frac{6}{32}$.

2.2.6 Addition rule

A common mistake that can be made in calculating $P(A \cup B)$ is to just add up the probabilities of each individual event, so the mistake will say this probability is $\frac{26}{52} + \frac{12}{52} = \frac{38}{52}$. The problem with this approach is that the outcomes that belong to both events (black picture cards) get counted twice, when we only want to count them once. This leads to the following formula for calculating probabilities involving unions of two events, and is sometimes called the **addition rule** in probability:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{2.2}$$

Using equation (2.2), $P(A \cup B) = \frac{26}{52} + \frac{12}{52} - \frac{6}{32} = \frac{32}{52}$.

2.2.7 Disjoint or Mutually Exclusive Events

The previous discussion leads to the idea of **disjoint**, or **mutually exclusive** events. Events are disjoint if they cannot happen simultaneously. In our card example, events A and B are not disjoint, since A and B can happen simultaneously, since a card that is drawn can be both black and a picture card, e.g. we draw a king of spades.

Using Figure 2.4 as a visual example, we can see that events A and B are not disjoint since the outcomes in blue overlap with the outcomes in gold.

Suppose we define another event, C, to denote that the card drawn is an Ace. The events B and C are disjoint since a card that is drawn cannot be both a picture card and an ace. This definition of disjoint events leads to the following: for events are disjoint, the probability of their intersection will be 0.

Using Figure 2.5 below as a visual example, we can see that events B and C are disjoint since the outcomes in gold and pink do not overlap.

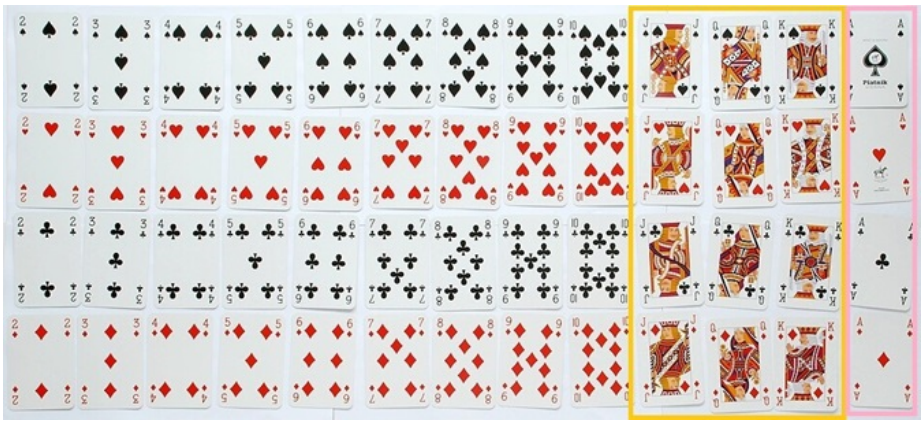

Figure 2.5: Events B, C (in gold and pink respectively)

Applying this idea to equation (2.2), we have the following for disjoint events: for disjoint events, the probability of at least one event happening is the sum of the probabilities for each event.

2.2.8 Axioms of Probability

The following are called the axioms of probability, which are considered foundation properties associated with probability:

1. The probability of any event, E, is non negative, i.e. $P(E) \geq 0$.
2. The probability that at least one outcome in the sample space occurs is 1, i.e.$P(S) = 1$.
3. If $A_1, A_2, \cdots$ are all disjoint events, then

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

In other words, for disjoint events, the probability that at least one event happens is the sum of their individual probabilities.

Note: most writers list these as three axioms. Our book combines the first two axioms into 1, and so write these as two axioms.

We can easily see how equations (2.1) and (2.2) can be derived from these axioms. Note that these equations and the axioms apply in all circumstances, regardless of whether the sample space is finite or not.

2.3 Conditional Probability

The concept of conditional probability appears in almost all statistical and data science models. In statistical models such as logistic regression, we are trying to use observable data (called predictors, input variables, etc) to model the probabilities associated with the different values of an outcome that is random (called response variable, output variable, etc). If the observable data are predictive of the outcome, then the probabilities associated with the outcome should indicate greater certainty, than if we do not have the observable data. Conditional probabilities allows us to incorporate observable data, or evidence, when evaluating uncertainty with random outcomes.

Consider that we are headed out for lunch, and we need to decide if we want to bring an umbrella (assuming we only bring an umbrella if we think it is going to rain). If we had been working in a windowless basement with no internet, we will have a high degree of uncertainty when evaluating if it will rain or not. However, if we were to look outside and observe the current weather conditions before heading out, we are likely to have a higher degree of certainty when evaluating if it will rain or not. Conditional probabilities allow us to incorporate what we see into our prediction of a random event.

If we were to use the language of probability to denote this example, let R denote the event that it will rain when we go for lunch. If we had been working in the windowless basement with no internet, we will be calculating $P(R)$, the probability it will rain when we go to lunch. If we are able to incorporate the current weather conditions, this probability will be denoted as $P(R|data)$, where data denotes the current observe weather conditions. $P(R|data)$ can be read as the probability that it will rain when we go to lunch, given what we have observed with the weather. With this example, we can see that $P(R)$ and $P(R|data)$ will be different, since we update our probability given useful information. Notice the | symbol inside the probability. This symbol implies that we are working with a conditional probability, with the given or observed information listed after the |.

2.3.1 Definition

If X and Y are events, with $P(X) > 0$, the conditional probability of Y given X, denoted by $P(Y|X)$, is

$$P(Y|X) = \frac{P(Y \cap X)}{P(X)}. \tag{2.3}$$

In this definition, we want to update the probability of Y happening, given that we have observed X. X can be viewed as the observable data or the evidence we want to incorporate.

In the Bayesian viewpoint of probability, $P(Y)$ is called the **prior** probability of Y since it reflects our belief about Y before observing any data. $P(Y|X)$ is called the **posterior** probability of Y, as it reflects an update on our belief about \$Y after incorporating observed data.

Let us go back to the standard deck of cards example. Let us find $P(B|A)$, the probability that our card is a picture card, given that we know the card is a black suit. Visually, we can use the definition of conditional probability using Figure 2.6 below.

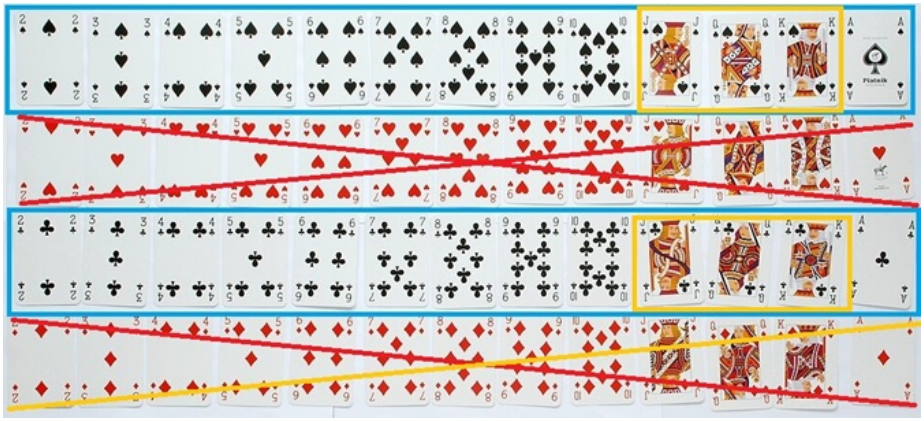

Figure 2.6: Events A, given B

We are told that our card is a black suit, so we have only 26 possible outcomes to consider, as the red cards are eliminated and are crossed out in Figure 2.6. out of these 26 outcomes, how many are picture cards? So this probability $P(B|A)$ is $\frac{6}{26}$.

Figure 2.6 represents the frequentist viewpoint of conditional probability: $P(B|A)$ represents the long run proportion of picture cards among cards that are black suits.

We can also apply equation (2.3): $P(B|A) = \frac{\frac{6}{52}}{\frac{1}{2}} = \frac{6}{26}$ which gives the same answer.

Thought question: work out the probability that the card drawn is a black suit, given that we know the card is a picture card.

We can see from this example that in general $P(Y|X) \neq P(X|Y)$. This informs us that we need to be extremely careful when writing out our conditional probabilities and interpreting them, and knowing which one matters to our analysis. For example, the probability that I feel unwell given that I have the flu is close to 1, but the probability that I have the flu given that I feel unwell is not close to 1 (since there are many things that can me me feel unwell). This confusion regarding conditional probabilities is sometimes called the confusion of the inverse or the prosecutor's fallacy. This fallacy wrongly assumes that if the probability of a fingerprint match given that the person is innocent is small, it means that the probability that the person is innocent given a fingerprint match must also be small. Before going over this fallacy in more detail, we need to cover a few more concepts.

2.3.2 Multiplication Rule

From equation (2.3), we have the **multiplication rule** in probability

$$P(Y \cap X) = P(Y|X) \times P(X) = P(X|Y) \times P(Y). \tag{2.4}$$

The multiplication rule is useful in finding the probability of multiple events happening, aespecially if the events happen sequentially. As an example, consider drawing two cards, without replacement, from a standard deck of cards. Without replacement means that after drawing the first card, it is not returned to the deck, so there will be 51 cards remaining after the first draw. Let D_1 and D_2 denote the events that the first draw is a diamond suit and the second draw is a diamond suit respectively. We want to find the probability that both cards drawn are diamond suits. This probability can be written as $P(D_1 \cap D_2) = P(D_1) \times P(D_2|D_1) = \frac{13}{52} \times \frac{12}{51} = \frac{156}{2652}$.

2.3.3 Independent Events

Events are independent if knowledge about whether one event happens or not does not change the probability of the other event happening. This implies that if X and Y are independent events, then the definition of conditional probability simplifies to $P(Y|X) = P(Y)$. Likewise $P(X|Y) = P(X)$. Applying this to the multiplication rule, we have the following for multiplication rule for independent events

$$P(Y \cap X) = P(Y) \times P(X). \tag{2.5}$$

The probability of all events happening is just the product of the probabilities for each individual event, if the events are all independent.

Going back to our example with the standard deck of cards, where A denotes the event that I draw a card with a black suit (spades or clubs), and B denotes

the event I draw a picture card (Jack, Queen, or King). We had earlier found that $P(B) = \frac{12}{52}$ and that $P(B|A) = \frac{6}{26}$. Notice that these two probabilities are numerically equal, which informs us that the events are independent. Knowing whether the card is a black suit or not does not change the probability that the card is a picture card. This makes sense intuitively since the proportion of cars that are picture is the same for black and red suits.

2.3.4 Bayes' Rule

The definition of conditional probability in equation (2.3) and the multiplication rule in equation (2.4) give us **Bayes' rule**

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}. \tag{2.6}$$

Bayes' rule is useful if we want to find $P(Y|X)$ but we only have information regarding $P(X|Y)$ available. A fairly popular model is called linear discriminant analysis, and it models the conditional probability using Bayes' rule.

2.3.5 Odds

The **odds** of an event Y are

$$odds(Y) = \frac{P(Y)}{P(Y^c)}. \tag{2.7}$$

You may realize that the left hand side of equation @ref{eq:odds} is equal to the left hand side of a logistic regression equation.

Using equation (2.7), we can switch from odds to probability easily

$$P(Y) = \frac{odds(Y)}{1 + odds(Y)}. \tag{2.8}$$

2.3.6 Odds Form of Bayes' Rule

Using Bayes' rule in equation (2.6) and the definition of odds in equation (2.7), we have the **odds form for Bayes' rule**

$$\frac{P(Y|X)}{P(Y^c|X)} = \frac{P(X|Y)}{P(X|Y^c)} \frac{P(Y)}{P(Y^c)}. \tag{2.9}$$

2.3.7 Law of Total Probability

Let $Y_1, Y_2, \cdots, Y_n$ be a partition of the sample space ($Y_1, Y_2, \cdots, Y_n$ are disjoint and their union is the sample space), with $P(Y_i) > 0$) for all i. Then

$$\begin{aligned} P(X) &= \sum_{i=1}^{n} P(X|Y_i) \times P(Y_i) \\ &= P(X|Y_1) \times P(Y_1) + P(X|Y_2) \times P(Y_2) + \cdots + P(X|Y_n) \times P(Y_n). \end{aligned} \tag{2.10}$$

The law of total probability informs us of a way to find the probability of X. We can divide the sample space in disjoint sets Y_i, find the conditional probability of X within each set, and then take a weighted sum of these conditional probabilities, weighted by $P(Y_i)$. This is useful if the conditional probability for each set is easy to obtain.

The law of total probability in equation (2.10) can be applied to the denominator of Bayes' rule in equation (2.6) to have the following variation of Bayes' rule:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{\sum_{i=1}^{n} P(X|Y_i) \times P(Y_i)}. \tag{2.11}$$

2.3.8 Worked Example

2.3.8.1 Approach 1: Using Bayes' Rule

We consider this worked example on how to apply Bayes' rule and the law of total probability. Suppose my email can be divided into three categories: E_1 denotes spam email, E_2 denotes important email, and E_3 denotes not important email. An email must belong to only one of these categories. Let F denote the event that the email contains the word "free". From past data, I have the following information:

- $P(E_1) = 0.2, P(E_2) = 0.5, P(E_3) = 0.3$.
- The word "free" appears in 99% of spam email, so $P(F|E_1) = 0.99$.
- The word "free" appears in 10% of important email, so $P(F|E_2) = 0.1$.
- The word "free" appears in 5% of important email, so $P(F|E_3) = 0.05$.

I receive an email that has the word free. What is the probability that it is spam? So we want to find $P(E_1|F)$. Using equation (2.11), we have

$$\begin{aligned} P(E_1|F) &= \frac{P(E_1 \cap F)}{P(F)} \\ &= \frac{P(F|E_1) \times P(E_1)}{P(F|E_1) \times P(E_1) + P(F|E_2) \times P(E_2) + P(F|E_3) \times P(E_3)} \\ &= \frac{0.99 \times 0.2}{0.99 \times 0.2 + 0.1 \times 0.5 + 0.05 \times 0.3} \\ &= 0.7528517 \end{aligned}$$

2.3.8.2 Approach 2: Using Tree Diagrams

A tree diagram is useful in finding conditional probabilities and probabilities involving intersections. It is a visual way of displaying the information you have at hand, when you have conditional probabilities over disjoint sets and probabilities for each disjoint set. In our toy example, the disjoint sets are the type of email I receive, E_1, E_2, E_3, and the conditional probabilities we have are over these disjoint sets, i.e. $P(F|E_1), P(F|E_2)$ and $P(F|E_3)$. We can put this information visual by first splitting our sample space into the disjoint sets E_1, E_2, E_3, and then splitting each disjoint set on whether the email has the word "free" (F) or not (F^c). This information is displayed in a tree diagram as in Figure 2.7.

Each split is represented by a branch, and we write the corresponding probability on each branch. We want to find the probability that a received email is spam given that it contains the word "free", $P(E_1|F)$, and using the definition of conditional probability in equation (2.3)

$$P(E_1|F) = \frac{P(E_1 \cap F)}{P(F)}.$$

Looking at the tree diagram in Figure 2.7, we can label the branches that lead to the numerator $P(E_1 \cap F)$, the probability that the email is spam and contains the word free. This is shown on the tree diagram below in Figure 2.8 below by highlighting the corresponding branches in blue.

So $P(E_1 \cap F) = 0.2 \times 0.99 = 0.198$. We then need to find the denominator $P(F)$. Looking at Figure 2.7, we can see three branches that lead to an email containing the word free: $P(E_1 \cap F)$ or $P(E_2 \cap F)$ or $P(E_3 \cap F)$. This is shown on the tree diagram below in Figure 2.9 below by highlighting the corresponding branches in gold.

We know the probability for each branch, and we add them up to obtain the denominator $P(F) = 0.2 \times 0.99 + 0.5 \times 0.1 + 0.3 \times 0.05 = 0.263$. Putting the pieces together, we have

$$P(E_1|F) = \frac{P(E_1 \cap F)}{P(F)} = \frac{0.198}{0.263} = 0.7528517.$$

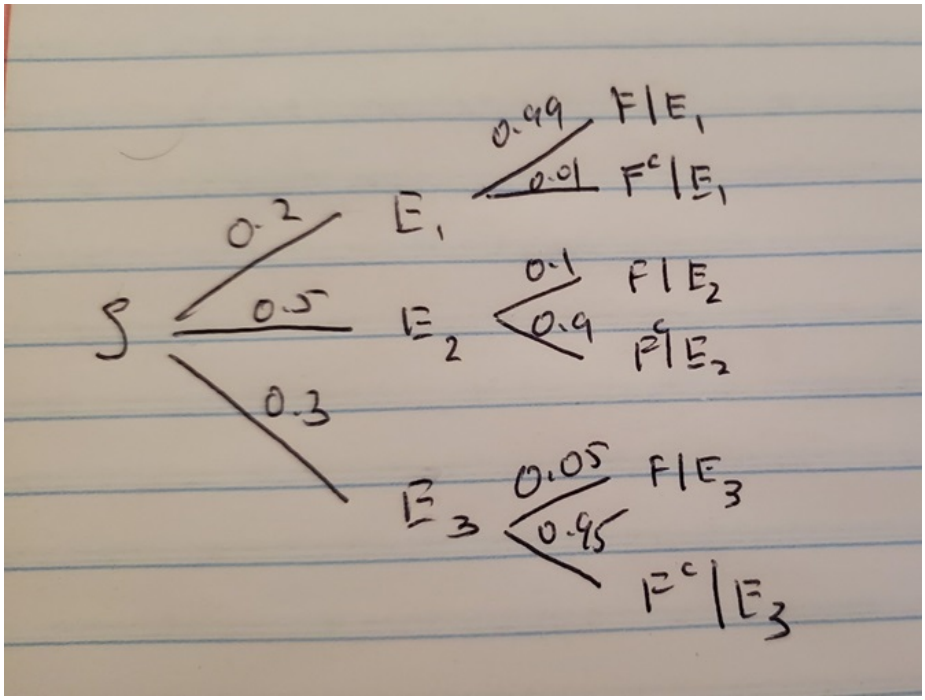

Figure 2.7: Tree Diagram for Email Example

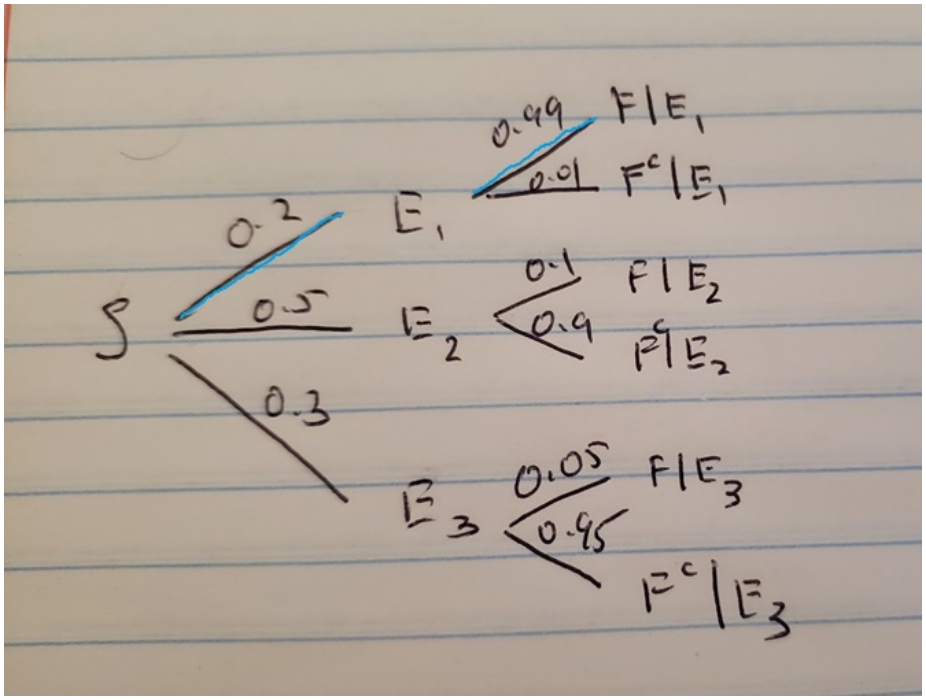

Figure 2.8: Tree Diagram for Email Example, Branch for Numerator in Blue

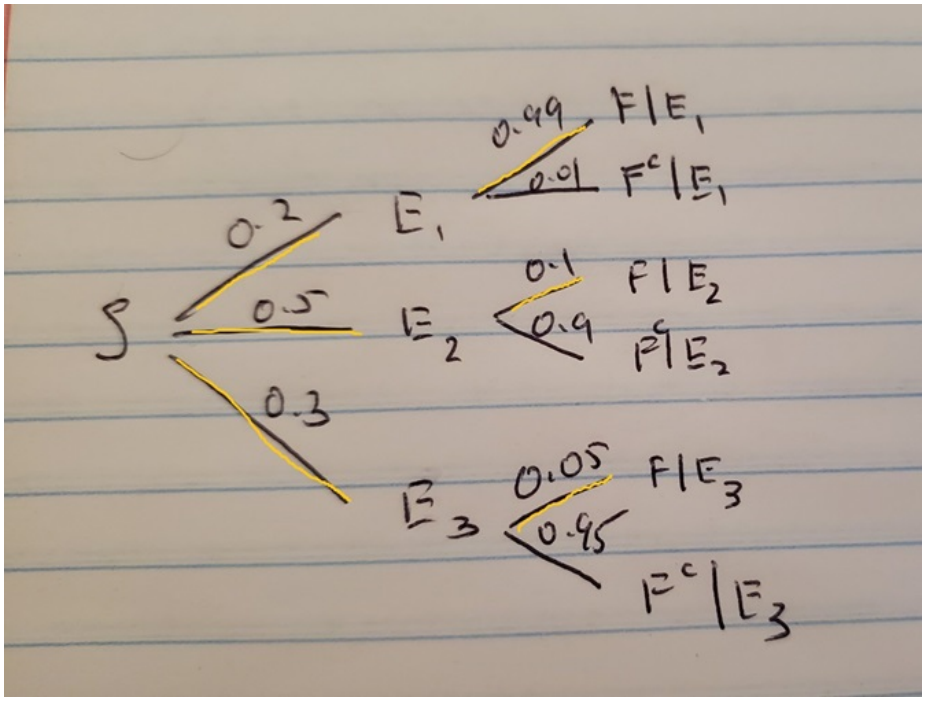

Figure 2.9: Tree Diagram for Email Example, Branches for Denominator in Gold

Note: If you compare the intermediate calculations in approach, you end up using the calculations in approach 1, without referring to any of the associated equations.

2.4 Confusion of the Inverse

We are now ready to talk about the prosecutor's fallacy, or the **confusion of the inverse**, that we had earlier mention in section 2.3.1. In essence, the confusion happens when we falsely equate $P(X|Y)$ to be equal to $P(Y|X)$. In fact, a large value for $P(X|Y)$ does not necessarily imply that $P(Y|X)$ is also large. The term prosecutor's fallacy when this confusion is applied in a criminal trial, e.g. the probability that an abusive relationship ends in murder could be small, but the probability that there was abuse in a relationship that ended in murder could be a lot higher.

We will go over some examples that are based on real life.

2.4.1 Disease Diagnostics

Suppose we are testing a patient if he has a rare disease, which is estimated to be prevalent in 0.5% of all people. Suppose we have a medical test for this disease that is accurate. There can be a number of definitions of accuracy. In disease diagnostics, a couple of measures are sensitivity, which is the proportion of people with the disease who test positive, and specificity, the proportion of people without the disease who test negative. A positive test indicates the person has the disease. Suppose the sensitivity and specificity are both high: 0.95 and 0.9 respectively. Suppose the patient tests positive, what is the probability that the patient actually has the disease? Assume the test always indicates positive or negative.

For this example, let D denote the event the patient has the disease, and let $+$ denote the event the patient tests positive on the test, and - denote the event the patient tests negative on the test. Given the information, we have

- $P(D) = 0.005$.
- $P(+|D) = 0.95$.
- $P(-|D^c) = 0.9$.

We wish to find $P(D|+)$. Using Bayes rule and the Law of Total probability, this is

$$\begin{aligned} P(D|+) &= \frac{P(D \cap +)}{P(+)} \\ &= \frac{P(+|D) \times P(D)}{P(+|D) \times P(D) + P(+|D^c) \times P(D^c)} \\ &= \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.1 \times 0.995} \\ &= 0.04556355 \end{aligned}$$

which is a small probability, so the patient is highly unlikely to actually have the rare disease. So while the test has high sensitivity with $P(+|D) = 0.95$, this does not imply that a patient who tests positive actually has the disease, since $P(D|+)$ is low. The implication is that for a rare disease, a positive test does not imply you have a high probability of having the disease, even if the test is accurate.

Why does this result make sense? Essentially, a large proportion of a small population could still be numerically much smaller than a small proportion of a large population. The disease is rare, so we have a small population of people with the disease, and almost all of them are detected by the test. We also have an extremely large population of people without the disease, and even a small proportion of them who erroneously test positive could still be a fairly large number. So among all the positive tests, most of the people do not have the disease. We consider the following table based on a population of 20 thousand people.

	Positive	Negative	Total
Disease	95	5	100
No Disease	1990	17910	19900
Total	2085	17915	20000

Look at the first column, which shows number of people who test positive. A see that a large proportion of diseased people are detected, but since there are relatively few people with the disease, this number is small, 95. A small proportion of people who do not have the disease test positive for the disease, and a small proportion of this large population results in a relatively larger number, 1990. So most of the people who test positive, $95 + 1990 = 2085$ actually do not have the disease. Therefore $P(D|+) = \frac{95}{2085} = 0.04556355$.

We can also explain this result through the Bayes' viewpoint of probability. Without knowing any information about the results of the test, the prior probability $P(D) = 0.005$. However, upon seeing that the person positive, we updated the posterior probability $P(D|+) = 0.04556355$, which is an increase from 0.005 when we knew knowing. The updated posterior probability is about 9 times

the prior. So we believe the person is more likely to have the disease upon viewing the positive test, than if we knew nothing about the test result. The posterior probability is still small since its value depends on two pieces of information: the prior $P(D)$ and the sensitivity $P(+|D)$. The product of these values belong to the numerator when calculating $P(D|+)$. The denominator is $P(+|D) \times P(D) + P(+|D^c) \times P(D^c)$. If the prior $P(D)$ is extremely low, then $P(D^c)$ is extremely close to 1, since the person either has the disease or does not have the disease. With $P(D)$ belong extremely low, the numerator is close to 0, and the value of the denominator is close to $P(+|D^c) \times P(D^c)$, therefore $P(D|+)$ is small.

Notice how we have talking about rare diseases? This confusion of the inverse, thinking that a high sensitivity implies that a person likely to have the disease if they test positive, only applies to rare diseases. If the disease is more prevalent, a high sensitivity is more likely to imply the person has the disease if they test positive.

So why should we take such tests for rare diseases? What should we do? We should go through the test again. It turns out that if you test positive twice for a rare disease, the probability that you have the disease increases by a lot than if you only tested once and tested positive.

To perform this calculation, we will use the odds form for Bayes' rule, per equation (2.9)

$$\begin{aligned}\frac{P(D|T_1 \cap T_2)}{P(D^c|T_1 \cap T_2)} &= \frac{P(T_1 \cap T_2|D)}{P(T_1 \cap T_2|D^c)} \frac{P(D)}{P(D^c)} \\ &= \frac{0.95^2}{0.1^2} \frac{0.005}{0.995} \\ &= 0.4535176\end{aligned}$$

where T_1 and T_2 denote the events the person test positive in the first test and second test respectively. We also assume that the results from each test are independent with previous tests.

The odds of having the disease given that the person positive twice is 0.4535176. Therefore, using equation (2.8), the corresponding probability of having the disease given that the person tested positive twice is $P(D|T_1 \cap T_2) = \frac{0.4535176}{1+0.4535176} = 0.3120138$. See how this posterior probability has increased with two positive tests, from 1 positive test.

Thought question: perform the calculations to show that the posterior probability that the person has the disease if the person tests positive on 3 tests is 0.8116199.

Thought question: do you notice a certain pattern emerging when performing these calculations as the person undergoes more tests? Could you write either

a mathematical equation, or even a function in R, that allows us to quickly compute the probability the person has the disease given that the person tested positive k times, where k can denote any non negative integer?

2.4.2 Prosecutor's Fallacy

The confusion of the inverse is also called the prosecutor's fallacy (sometimes also called the defense attorney's fallacy depending on which side is making the mistake) when it occurs in a legal setting. Generally, the confusion comes from equating P(evidence|innocent) with P(innocent|evidence).

The book provides a discussion about this in Section 2.8, examples 2.8.1 and 2.8.2.

Chapter 3

Discrete Random Variables

This module is based on Introduction to Probability (Blitzstein, Hwang), Chapters 3 and 4. You can access the book for free at https://stat110.hsites.harvard.edu/ (and then click on Book). Please note that I cover additional topics, and skip certain topics from the book. You may skip Sections 3.4, 3.7, 3.9, 4.3, 4.9 from the book.

3.1 Random Variables

The idea behind random variables is to simplify notation regarding probability, enable us to summarize results of experiments, and make it easier to quantify uncertainty.

3.1.1 Example

Consider flipping a coin three times and recording if it lands heads or tails each time. The sample space for this experiment will be $S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$. Given that each outcome is equally likely, the probability associated with each outcome is $\frac{1}{8}$.

Suppose I want to find the probability that I get exactly 2 heads out of the 3 flips. I could express this as:

- P(two heads out of three flips), or
- $P(HHT \cup HTH \cup THH)$, or
- $P(A)$ where A denotes the event of getting two heads out of three flips.

Another way is to define a random variable X that expresses this event a bit more efficiently. Let X denote the number of heads out of three flips, so another way could be to write $P(X = 2)$. This is idea behind random variables: to assign events to a number.

3.1.2 Definition

A **random variable (RV)** is a function from the sample space to real numbers.

By convention, we denote random variables by capital letters. Using our 3 coin flip example, X could be 0, 1, 2, or 3. We assign a number to each possible outcome of the sample space.

Random variables provide numerical summaries of the experiment. This can be useful especially if the sample space is complicated. Random variables can also be used for non numeric outcomes.

3.1.3 Discrete Vs Continuous

One of the key distinctions we have to make for random variables is to determine if it is discrete or continuous. The way we express probabilities for random variables depends on whether the random variable is discrete or continuous.

A **discrete random variable** can only take on a countable (finite or infinite) number of values.

The number of heads in 3 coin flips, X is countable and finite, since we can actually list all of the values it can take as $\{0, 1, 2, 3\}$ and there are 4 such values. X must take on one of these 4 numerical values; it cannot be a number outside this list. So it is discrete.

A random variable is countable and infinite if we can list the values it can take, but the list has no end. For example, the number of people using a crosswalk over a 10 year period could take on the values $\{0, 1, 2, 3, \cdots\}$. The number could take on any of an infinite number of values, but values in between these whole numbers cannot occur. So the number of people using a crosswalk over a 10 year period is a discrete random variable.

A **continuous random variable** can take on an uncountable number of values in an interval of real numbers.

For example, height of an American adult is a continuous random variable, as height can take on any value in interval between 40 and 100 inches. All values between 40 and 100 are possible.

For this module, we will focus on discrete random variables.

The **support** of a discrete random variable X is the set of values X can take such that $P(X = x) > 0$, i.e. the set of values that have non zero probability of happening. Using our 3 coin flips example, where X is the number of heads out of the 3 coin slips, the support is $\{0, 1, 2, 3\}$. Usually, the support of discrete random variables are integers.

Thought question: Can you come of examples of discrete and continuous random variables on your own? Feel free to search the internet for examples as well.

Table 3.1: PMF for X

x	PMF
0	0.125
1	0.375
2	0.375
3	0.125

3.2 Probability Mass Functions (PMFs)

We use probability to describe the behavior of random variables. This is called the **distribution** of a random variable. It specifies the probabilities of all events associated with the random variable. For example, what is the probability of obtaining 3 heads in 3 coin flips, or what is the probability of obtaining at least one head on 3 coin flips?

For discrete random variables, the distribution is specified by the **probability mass function (PMF)**. The PMF of a discrete random variable X is the function $P_X(x) = P(X = x)$. It is positive when x is in the support of X, and 0 otherwise.

Note: In the notation for random variables, capital letters such as X denote random variables, and lower case letters such as x denote actual numerical values. So if we want to find the probability that we have 2 heads in 3 coin flips, we write $P(X = 2)$, where x is 2 in this example.

Going back to our example where we record the number of heads out of 3 coin flips, we can write out the PMF for the random variable X:

- $P_X(0) = P(X = 0) = P(TTT) = \frac{1}{8}$,
- $P_X(1) = P(X = 1) = P(HTT \cup THT \cup TTH) = \frac{3}{8}$,
- $P_X(2) = P(X = 2) = P(HHT \cup THH \cup HTH) = \frac{3}{8}$,
- $P_X(3) = P(X = 3) = P(HHH) = \frac{1}{8}$.

Fairly often, the PMF of a discrete random variable is presented in a simple table like in Table 3.1 below:

Or use a simple plot like the one below in Figure 3.1:

```
##support
x<-0:3
## PMF for each value in the support.
PMFs<-c(1/8, 3/8, 3/8, 1/8)
## create plot of PMF vs each value in support
plot(x, PMFs, type="h", main = "PMF for X", xlab="# of heads", ylab="Probability", ylim=c(0,1))
```

The PMF provides a list of all possible values for the random variable and the corresponding probabilities. In other words, the PMF describes the distribution

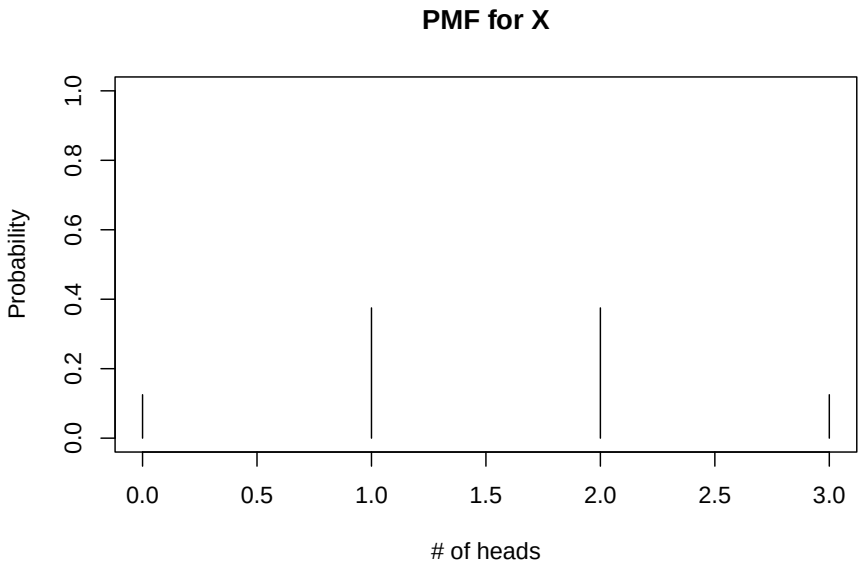

Figure 3.1: PMF for X

of the relative frequencies for each outcome. For our experiment, observing 1 or 2 heads is equally likely, and they occur three times as often as observing 0 or 3 heads. Observing 0 or 3 heads is also equally likely.

3.2.1 Valid PMFs

Consider a discrete random variable X with support $x_1, x_2, \cdots$. The PMF $P_X(x)$ of X must satisfy:

- $P_X(x) > 0$ if $x = x_j$, and $P_X(x) = 0$ otherwise.
- $\sum_{j=1}^{\infty} P_X(x_j) = 1$.

In other words, the probabilities associated with the support are greater than 0, and the sum of the probabilities across the whole support must add up to 1.

Thought question: based on Table 3.1, can you see why our PMF for X is valid?

3.2.2 PMFs and Histograms

Recall the frequentist viewpoint of probability, that it represents the relative frequency associated with an event that is repeated for an infinite number of times.

Consider our experiment where we flip a coin 3 times and count the number of

heads. The support of our random variable X, the number of heads, is $\{0, 1, 2, 3\}$. Imagine performing our experiment a large number of times. Each time we perform the experiment, we record the number of heads. If we performed the experiment one million times, we would have recorded one million values for the number of heads, and each value must be in the support of X. If we then create a histogram for the one million values for the number of heads, the shape of the histogram should be very close to the shape of the plot of the PMF in Figure 3.1. Figure 3.2 below shows the resulting histogram after performing the experiment 1 million times.

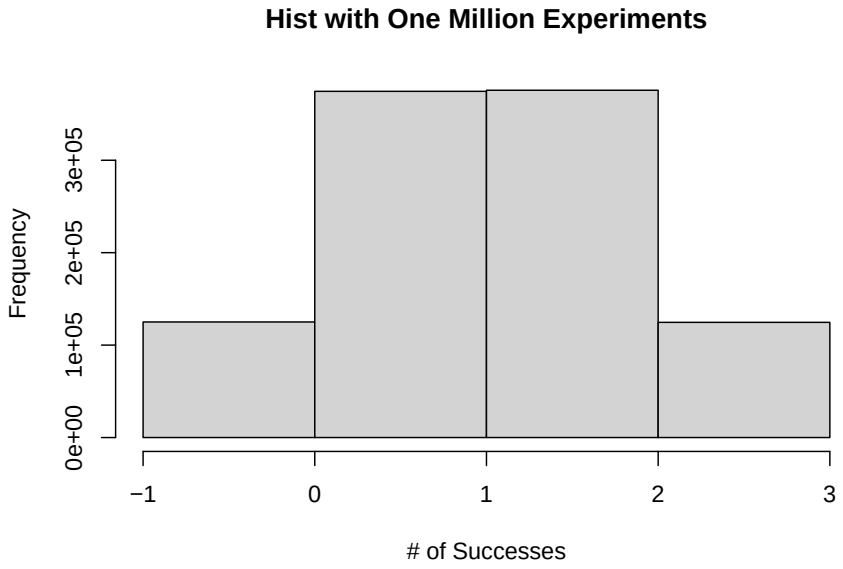

Figure 3.2: Histogram from Experiment Performed 1 Million Times

Note: What we have just done here was to use simulations to repeat an experiment a large number of times using code.

3.3 Cumulative Distribution Functions (CDFs)

Another function that is used to describe the distirbution of discrete random variables is the **cumulative distribution function (CDF)**. The CDF of a random variable X is $F_X(x) = P(X \leq x)$. Notice that unlike the PMF, the definition of CDF applies for both discrete and continuous random variables.

Going back to our example where we record the number of heads out of 3 coin flips, we can write out the CDF for the random variable X:

Table 3.2: CDF for X

x	CDF
0	0.125
1	0.500
2	0.875
3	1.000

- $F_X(0) = P(X \leq 0) = P(X = 0) = \frac{1}{8}$,
- $F_X(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}$,
- $F_X(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = \frac{1}{2} + \frac{3}{8} = \frac{7}{8}$,
- $F_X(3) = P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = \frac{7}{8} + \frac{1}{8} = 1$.

Notice how these calculations were based on the PMF. To find $P(X \leq x)$, we summed the PDF over all values of the support that is less than or equal to x. Fairly often, the CDF of a discrete random variable is presented in a simple table like Table 3.2 below:

Or in a simple plot like in Figure 3.3 below:

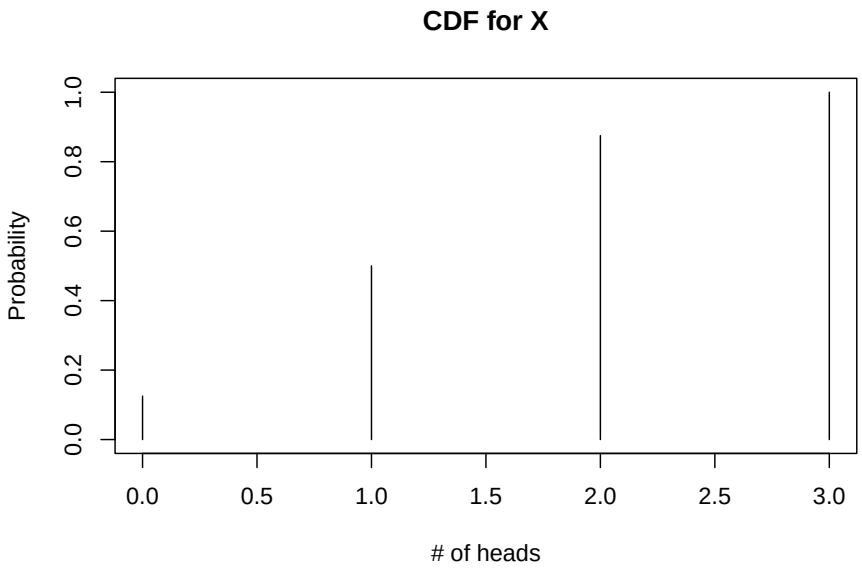

Figure 3.3: CDF for X

Thought question: do you see similarities between the CDF and the empirical

cumulative density function (ECDF) from section 1.3.3?

3.3.1 Valid CDFs

The CDF $F_X(x)$ of X must:

- be non decreasing. This means that as x gets larger, the CDF either stays the same or increases. Visually, a graph of the CDF should never decreases as x increases.
- approach 1 as x approaches infinity and approach 0 as x approaches negative infinity. Visually, a graph of the CDF should be equal to or close to 1 for large values of x, and it should be equal to or close to 0 for small values of x.

Thought question: Look at the CDF for our example in Figure 3.3, and see how it satisfies the criteria listed above for a valid CDF.

3.4 Expectations

In the previous section, we see how PMFs and CDFs can be used to describe the distribution of a random variable. As the PMF can be viewed as a long-run version of the histogram, it gives us an idea about the shape of the distribution. Similar to Section 1, we will also be interested in measures of centrality and spread for random variables.

A measure of centrality for random variables is the **expectation**. The expectation of a random variable can be interpreted as the long-run mean of the random variable, i.e. if we were able to repeat the experiment an infinite number of times, the expectation of the random variable will be the average result among all the experiments.

For a discrete random variable X with support $x_1, x_2, \cdots,$, the expected value, denoted by $E(X)$, is

$$E(X) = \sum_{j=1}^{\infty} x_j P(X = x_j). \tag{3.1}$$

We can use Table 3.1 as an example. To find the expected number of heads out of 3 coin flips, using equation (3.1),

$$\begin{aligned} E(X) &= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} \\ &= 1.5 \end{aligned}$$

What we did was to take the product of each value in the support of the random variable with its corresponding probability, and add all these products.

We can see another interpretation of the expected value of a random variable from this calculation: it is the weighted average of the values for the random variable, weighted by their probabilities.

Intuitively, this expected value of 1.5 should make sense. If we flip a coin 3 times, and the coin is fair, we expect half of these flips to land heads, or 1.5 flips to land heads.

3.4.1 Linearity of Expectations

We have seen how to calculate the expected value of a random variable X using equation (3.1). What we need is the PMF of X. Sometimes our random variable can be viewed as a sum (or difference) of other random variables, or it could involve a multiplication and / or adding a constant to the random variable. Consider some of these scenarios:

- Suppose my friend and I are fisherman. Let Y be the random variable describing the number of fish I catch on a workday, and let W be the random variable describing the number of fish my friend catches on a workday. We can let $T = Y + W$ be the random variable describing the total number of fish we catch on a workday.
- Suppose that I sell each fish for \$10 and my friend sells each fish for \$15. We can let $R = 10Y + 15W$ be the random variable that describes the revenue we generate on a workday.
- Suppose that my friend and I rent out a space at the market to sell our fish, and it costs \5 a day to rent out the space. We can let $G = 10Y + 15W - 5$ be the random variable that describes our gross income for the day.

All of these examples involve new random variables, T, R, G that can be based on previously defined random variables, Y, W. It turns out that to find the expectations of the new random variables, all we need is the expectations of the previously defined random variables. We do not need to find the PMFS for T, R and S and then apply equation (3.1).

These can be done through the **linearity of expectations**: Let X and Y denote random variables, and a, b, c denote some constants, then

$$E(aX + bY + c) = aE(X) + bE(Y) + c. \tag{3.2}$$

Applying equation (3.2) to the fishing examples:

- $E(T) = E(Y) + E(W)$,
- $E(R) = 10E(Y) + 15E(W)$,
- $E(G) = 10E(Y) + 15E(W) - 5$.

All we need to find the expected values for the total number of fish, revenue generated, and gross income were the expected values for the number of fish

each of us caught. We do not need the PMFs for T, R, G.

3.4.1.1 Visual Explanation

For a visual explanation of why equation (3.2) makes sense, we go back to our previous example where X denotes the number of heads in 3 coin flips. Figure 3.1 displays the PMF for this random variable. Let us create the PMF for a new random variable $Y = 2X$ and display it in Figure 3.4 below:

```
##support of X
x<-0:3
## PMF for each value in the support.
PMFs<-c(1/8, 3/8, 3/8, 1/8)
EX<-1.5

##support of Y
y<-2*x
## PMF for each value in the support.
PMFs<-c(1/8, 3/8, 3/8, 1/8)
EY<-2*EX

par(mfrow=c(2,1))

## create plot of PMF vs each value in support
plot(x, PMFs, type="h", main = "PMF for X", xlab="X", ylab="Probability", ylim=c(0,1))
##overlay a line representing EX in red
abline(v=EX, col="red")

## create plot of PMF vs each value in support
plot(y, PMFs, type="h", main = "PMF for Y", xlab="Y", ylab="Probability", ylim=c(0,1))
##overlay a line representing EY in red
abline(v=EY, col="red")
```

Note that the red vertical lines represent the expected value for the random variable, and since the PMFs are symmetric, the expected value lies right in the middle of the support. Comparing the PMFs in Figure 3.4, we get Y by multiplying X by 2. So the support of Y is now $\{0, 2, 4, 6\}$ but the associated probabilities are unchanged, so the heights of the probabilities on the vertical axis are unchanged. Therefore, the center, the expected value, is multiplied by the same constant.

Consider another random variable $W = X + 3$. We create the PMF for W and display it in Figure 3.5 below:

```
##support of X
x<-0:3
## PMF for each value in the support.
PMFs<-c(1/8, 3/8, 3/8, 1/8)
```

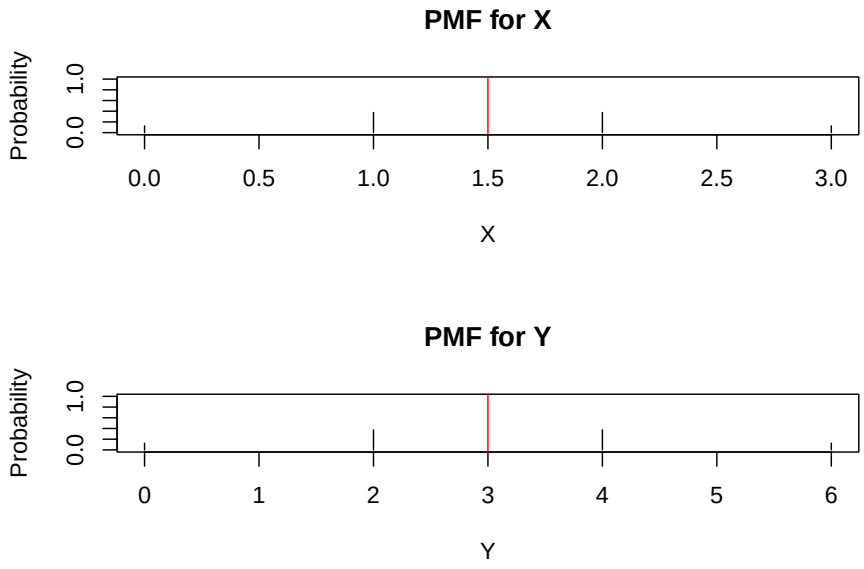

Figure 3.4: PMF for X and Y=2X

```
EX<-1.5

##support of W
w<-x+3
## PMF for each value in the support.
PMFs<-c(1/8, 3/8, 3/8, 1/8)
EW<-EX+3

par(mfrow=c(2,1))

## create plot of PMF vs each value in support
plot(x, PMFs, type="h", main = "PMF for X", xlab="X", ylab="Probability", ylim=c(0,1))
##overlay a line representing EX in red
abline(v=EX, col="red")

## create plot of PMF vs each value in support
plot(w, PMFs, type="h", main = "PMF for w", xlab="W", ylab="Probability", ylim=c(0,1))
##overlay a line representing EW in red
abline(v=EW, col="red")
```

Notice the PMFs for X and W look almost exactly the same. The only difference is that every value in the support for X is shifted by 3 units. The probabilities

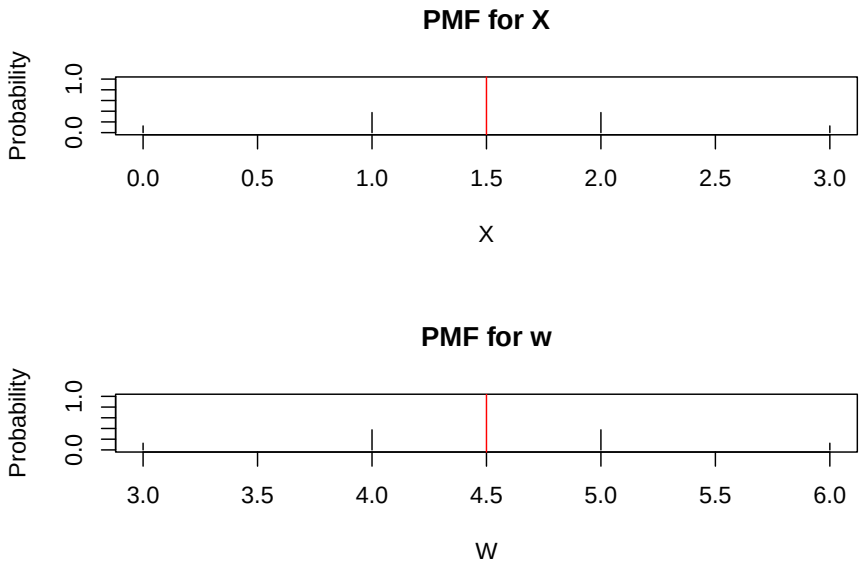

Figure 3.5: PMF for X and W=X+3

stay the same, so the heights in the PMFs are unchanged. So if every value is shifted by 3 units, the expected value, the long-run average, also gets shifted by 3 units. Adding a constant to a random variable shifts the expected value accordingly.

3.4.1.2 One More Example

We look at one more example to illustrate the usefulness of the linearity of expectations. Consider a drunk man who has to walk on a one-dimensional number line and starts at the 0 position. For each step the drunk man takes, he either moves forward, backward, or stays at the same spot. He steps forward with probability p_f, backward with probability p_b, and stays at the same spot with probability p_s, where $p_f + p_b + p_s = 1$. Let Y be the position on the number line of the drunk man after 2 steps. What is the expected position of the drunk man after two steps, i.e. what is $E(Y)$? Assume that each step is independent.

Using brute force, we can find the PMF of Y, and find $E(Y)$ using equation (3.1). First, we need to find the sample space for Y. With two steps, the sample space is $\{-2, -1, 0, 1, 2\}$. Next, we need to find the probabilities associated with each outcome in the sample space.

- For $Y = -2$, the man must move backward on each step. This probability will be $P(Y = -2) = p_b^2$.

- Likewise, for $Y = 2$, the man must move forward on each step. This probability will be $P(Y = 2) = p_f^2$.
- For $Y = -1$, the man could stay on the first step, then move back on the second, or move back on the first step, and stay on the second. This probability will be $P(Y = -1) = p_s p_b + p_b p_s = 2p_b p_s$.
- Likewise, for $Y = 1$, the man could stay on the first step, then move forward on the second, or move forward on the first step, and stay on the second. This probability will be $P(Y = 1) = p_s p_f + p_f p_s = 2p_f p_s$.
- For $Y = 0$, the man could move forward, then backward, or move backward then forward, or stay on both steps. So $P(Y = 0) = p_f p_b + p_b p_f + p_s^2 = p_s^2 + 2p_b p_f$.

Using equation @ref(eq:3_EX),

$$\begin{aligned} E(Y) &= -2 \times p_b^2 + -1 \times 2p_b p_s + 0 \times p_s^2 + 2p_b p_f + 1 \times 2p_f p_s + 2 \times p_f^2 \\ &= 2(p_f - p_b) \end{aligned}$$

Note: I skipped a lot of messy algebra to get to the end result. Even with skipping this step, setting up the PMF was quite a bit of work.

Suppose we use the linearity of expectations in equation (3.2). Let Y_1, Y_2 denote the distance the man moves at step 1 and 2 respectively. Then $Y = Y_1 + Y_2$. The sample of Y_1 and Y_2 are the same: $\{-1, 0, 1\}$. Both Y_1 and Y_2 have the following PMF:

- $P(Y_i = -1) = p_b$
- $P(Y_i = 0) = p_s$
- $P(Y_i = 1) = p_f$

And using equation (3.1),

$$\begin{aligned} E(Y_i) &= -1 \times p_b + 0 \times p_s + 1 \times p_f \\ &= p_f - p_b \end{aligned}$$

So therefore $E(Y) = E(Y_1) + E(Y_2) = 2(p_f - p_b)$. Notice how much simpler the solution becomes using linearity of expectations? Imagine if we wanted to find the expected position after 500 steps? Writing out the sample space for 500 steps will be extremely long.

3.4.2 Law of the Unconscious Statistician

Suppose we have the PMF of a random variable X, and we want to find $E(g(X))$, where g is a function of X (you can think of g as a transformation performed on X). One idea could be to find the PMF of $g(X)$ and then use the definition

of expectation in equation (3.1). But we have seen in the previous subsection that finding the sample space after transforming the random variable can be challenging. It turns out we can find $E(g(X))$ based on the PMF of X, without having to find the PMF of $g(X)$.

This is done through the **Law of the Unconscious Statistician (LOTUS)**. Let X be a discrete random variable with support $\{x_1, x_2, \cdots\}$, and g is a function applied to X, then

$$E(g(X)) = \sum_{i=j}^{\infty} g(x_j)P(X = x_j). \tag{3.3}$$

An application of LOTUS is in finding the variance of a discrete random variable.

3.4.3 Variance

We have talked about the shape of the distribution of a discrete random variable, and its expected value. One more measure that we are interested in is the spread associated with the distribution. One common measure is the variance of the random variable.

The **variance** of a random variable X is

$$Var(X) = E[(X - EX)^2] \tag{3.4}$$

and the **standard deviation** of a random variable X is the squareroot of its variance

$$SD(X) = \sqrt{Var(X)}. \tag{3.5}$$

Looking at equation (3.4) a little more closely, we can see a natural interpretation of the variance of a random variable: it is the average squared distance of the random variable from its mean, in the long-run. An equivalent formula for the variance of a random variable is

$$Var(X) = E(X^2) - (EX)^2. \tag{3.6}$$

Equation (3.6) is easier to work with than equation (3.4) when performing actual calculations.

Let us now go back to our original example, where X denotes the number of heads out of 3 coin flips. Earlier, we found the PMF of this random variable, per Table 3.1, and we found its expectation to be 1.5. To find the variance of X using equation (3.6), we find $E(X^2)$ first using LOTUS in equation (3.3)

$$\begin{aligned} E(X^2) &= 0^2 \times \frac{1}{8} + 1^2 \times \frac{3}{8} + 2^2 \times \frac{3}{8} + 3^2 \times \frac{1}{8} \\ &= 3 \end{aligned}$$

so $Var(x) = 3 - 1.5^2 = \frac{3}{4}$.

Thought question: Try to find $Var(X)$ using equation (3.4) and LOTUS. You should arrive at the same answer but the steps may be a bit more complicated.

3.4.3.1 Properties of Variance

Variance has the following properties:

- $Var(X + c) = Var(X)$, where c is a constant. This should make sense, since if we add a constant to a random variable, we shift it by c units. As shown earlier in Figure 3.5, the expected value also gets shifted by c units. Variance measures the average squared distance of a variable from its mean. So the distance, and the squared distance, of X from its mean is unchanged.
- $Var(cX) = c^2 Var(X)$. Look at Figure 3.4, notice the distance between each value in the support from its expected value gets multiplied by 2 (since $Y = 2X$). So if we multiply a random variable by c, the distance between each value in the support on its expected value is multiplied by c. Since variance measures squared distance, the variance gets multiplied by c^2.
- If X and Y are independent random variables, then $Var(X + Y) = Var(X) + Var(Y)$.

3.5 Common Discrete Random Variables

Next, we will introduce some commonly used distributions that may be used for discrete random variables. A number of common statistical models (for example, logistic regression, Poisson point process) are based on these distributions.

3.5.1 Bernoulli

The Bernoulli distribution might be the simplest discrete random variable. The support for such a random variable is $\{0, 1\}$. In other words, the value of a random variable that follows a Bernoulli distribution is either 0 or 1. A Bernoulli distribution is also described by the parameter p, which is the probability that the random variable takes on the value of 1.

More formally, a random variable X follows a **Bernoulli distribution** with parameter p if $P(X = 1) = p$ and $P(X = 0) = 1 - p$, where $0 < p < 1$. Using mathematical notation, we can write $X \sim Bern(p)$ to express that the

random variable X is distributed as a Bernoulli with parameter p. The PMF of a Bernoulli distribution is written as

$$P(X = k) = p^k(1-p)^{1-k} \tag{3.7}$$

for $k = 0, 1$.

It is not enough to specify that a random variable follows a Bernoulli distribution. We need to also clearly specify the value of the parameter p. Consider the following two examples which describe two different experiments:

- Suppose I flip a fair coin once. Let $Y = 1$ if the coin lands heads, and $Y = 0$ if the coin lands tails. $Y \sim Bern(\frac{1}{2})$ in this example since the coin is fair.
- Suppose I am asked a question and I am given 5 multiple choices, of which only 1 is the correct answer. I have no idea about the topic, and the multiple choices do not help, so I have to guess. Let $W = 1$ if I answer correctly, and $W = 0$ if I answer incorrectly. $W \sim Bern(\frac{1}{5})$.

$P(Y = 1)$ and $P(W = 1)$ are not the same in these examples.

Fairly often, when we have a Bernoulli random variable, the event that results in the random variable being coded as 1 is called a **success**, and the event that results in the random variable being coded as 0 is called a **failure**. In such a setting, the parameter p is called the **success probability** of the Bernoulli distribution. An experiment that has a Bernoulli distribution can be called a Bernoulli trial.

If you go back to the second example in section 0.1.2, we were modeling whether a job applicant receives a callback or not. In this example, we could let V be the random variable that an applicant receives a callback, with $V = 1$ denoting the applicant received a callback, and $V = 0$ when the applicant did not receive a callback. We used logistic regression in the example. It turns out that logistic regression is used when the variable of interest follows a Bernoulli distribution.

3.5.1.1 Properties of Bernoulli

Consider X is a Bernoulli distribution with parameter p. The expectation of a Bernoulli distribution is

$$E(X) = p \tag{3.8}$$

and its variance is

$$Var(X) = p(1-p). \tag{3.9}$$

Thought question: Use the definition of expectations for discrete random variables, equation (3.1), and the PMF of a Bernoulli random variable, and LOTUS to prove equations (3.8) and (3.9).

The expected value being equal to p for a Bernoulli distribution should make sense. Remember that the support for such a random variable is 0 or 1, with $P(X = 1) = p$. Using the frequentist viewpoint, if we were to flip a coin and record heads or tails, and repeat this coin flipping many times, we will have record a bunch of 0s and 1s to represent the result for all the coin flips. The average of this bunch of 0s and 1s is just the proportion of 1s.

The equation for the variance of a Bernoulli distribution exhibits an interesting and intuitive behavior. Figure 3.6 below shows how the variance behaves as we vary the value of p:

```
p<-seq(0,1,by = 0.001) ##sequence of values for p
Bern_var<-p*(1-p) ##variance of Bernoulli
##plot variance against p
plot(p, Bern_var, ylab="Variance", main="Variance of Bernoulli as p is Varied")
```

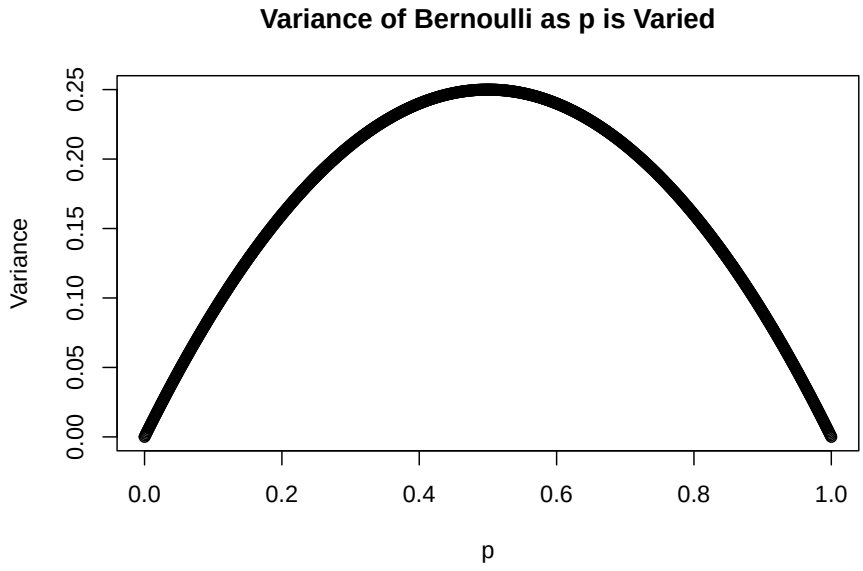

Figure 3.6: Variance of Bernoulli

Notice the variance is at a maximum when $p = 0.5$, and the variance is minimum (in fact it is 0) when $p = 0$ or $p = 1$. If we have a biased coin such that it always lands heads, every coin flip will land on heads with no exception. There is no variability in the result, and we have utmost certainty in the result of each coin

flip. On the other hand, if the coin is fair such that $p = 0.5$, we have the least certainty in the result of each coin flip, and so variance is maximum when the coin is fair.

Another application of this property is during election results (assuming 2 candidates, but the same idea applies for more candidates). For swing states where the race is closer (so p is closer to half), projections on the winner have more uncertainty and so we need to get more data and wait longer for the projections. For states that primarily vote for one candidate (so p is closer to 0 or 1), projections happen a lot quicker as projections have less uncertainty.

3.5.2 Binomial

Suppose we have an experiment that follows a Bernoulli distribution, and we perform this experiment n times (sometimes called trials), each time with the same success probability p. The experiments are independent from each other. Let X denote the number of successes out of the n trials. X follows a **binomial distribution** with parameters n and p (number of trials and success probability). We write $X \sim Bin(n, p)$ to express that X follows a binomial distribution with parameters n and p, with $n > 0$ and $0 < p < 1$. The PMF of a Binomial distribution is written as

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.10}$$

for $k = 0, 1, 2, \cdots, n$, which is also the support of the binomial distribution.

In equation (3.10), $\binom{n}{k}$ is called the binomial coefficient, and is a number that represents the number of combinations that result in k successes out of the n trials. The binomial coefficient can be found using

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{3.11}$$

$n!$ is called n-factorial, and is the product of all positive integers less than or equal to n. So $n! = n \times (n-1) \times (n-2) \times \cdots \times 1$. As an example $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$, or using R:

```
factorial(5)
```

```
## [1] 120
```

Note: A fairly common model, the logistic regression model with aggregated data, is based on the binomial distribution. We mentioned logistic regression earlier. The difference between these two (with and without aggregated data) is based on the structure of the data frame. If you are interested in these differences, please read https://www.r-bloggers.com/2021/02/how-to-run-logistic-regression-on-aggregate-data-in-r/.

We go back to our first example of counting the number of heads out of three coin flips follows a binomial distribution. Each coin flip is either heads or tails. The success probability, the probability of heads, is 0.5 and is the same for each flip. The result of each flip is independent of other flips since other flips do not affect the outcome. The number of trials (flips in this example) is $n = 3$ is specified as a fixed value. We let x denote the number of heads in 3 coin clips, so we write $X \sim Bin(3, 0.5)$.

Suppose we want to calculate $P(X = 2)$ using equation (3.10):

$$\begin{aligned} P(X = 2) &= \binom{3}{2}(0.5)^2(0.5)^1 \\ &= \frac{3!}{2!1!}(0.5)^2(0.5)^1 \\ &= 3 \times \frac{1}{8} \\ &= \frac{3}{8}. \end{aligned}$$

In this example, the binomial coefficient equals to 3. Which indicates there were 3 combinations to obtain 2 heads in 3 coin flips. $P(X = 2$ can be written as $P(HHT \cup HTH \cup THH)$. Solving for $P(HHT \cup HTH \cup THH)$, we have

$$\begin{aligned} P(HHT \cup HTH \cup THH) &= P(HHT) + P(HTH) + P(THH) \\ &= 0.5^3 + 0.5^3 + 0.5^3 \\ &= 3 \times \frac{1}{8} \\ &= \frac{3}{8}. \end{aligned}$$

so we could have solved this using basic probability rules from the previous module, without using the PMF of the binomial distribution in equation (3.10). Of course, the PMF of the binomial distribution gets a lot more convenient if n gets larger, as the number of combinations and sample space get a lot larger.

We can also use R to find $P(X = 2)$:

```
dbinom(2,3,0.5) ##specify values of k, n, p in this order
```

```
## [1] 0.375
```

3.5.2.1 Relationship Between Binomial and Bernoulli

Looking at the description of the Bernoulli and binomial distributions, you may notice that a Bernoulli random variable is a special case of a binomial random variable when $n = 1$, i.e. when we have only 1 trial.

The binomial random variable is also sometimes viewed as the sum of n independent Bernoulli random variables, all with the same value of p.

3.5.2.2 Properties of Binomial

If $X \sim Bin(n, p)$, then

$$E(X) = np \tag{3.12}$$

and

$$Var(X) = np(1-p). \tag{3.13}$$

These results should make sense when we note the relationship between a binomial random variable and Bernoulli random variable. Suppose we have random variables $Y_1, Y_2, \cdots, Y_n$ and they are all Bernoulli random variables with parameter p and are independent. Then $Y = Y_1 + Y_2 + \cdots + Y_n \sim Bin(n, p)$. Therefore, using the linearity of expectations in equation (3.2), $E(Y) = E(Y_1) + E(Y_2) + \cdots + E(Y_n) = np$. Since $Y_1, Y_2, \cdots, Y_n$ are independent, $Var(Y) = Var(Y_1) + Var(Y_2) + \cdots + Var(Y_n) = np(1-p)$.

3.5.2.3 PMFs of Binomial

We take a look at the PMFs of a few binomials, all with $n = 10$ but we vary p to be 0.2, 0.5, and 0.9, in Figure 3.7:

From figure 3.7, we can see that the distribution of the binomial is symmetric when $p = 0.5$, as middle values of k have higher probabilities, and the probabilities decrease as we go further away from the middle. When $p \neq 0.5$, we see that the distribution gets skewed. When the success probability is small, smaller number of successes are likelier, and when the success probability is large, larger number of successes are likelier, which is intuitive. If the probability of success is small, we expect most outcomes to be failures.

3.5.3 Poisson

One more common distribution used for discrete random variables is the Poisson distribution. This is often used when the variable of interest is what we call count data (the support is non negative integers), for example, the number of cars that cross an intersection during the day.

A random variable X follows a **Poisson distribution** with parameter λ, where $\lambda > 0$. Using mathematical notation, we can write $X \sim Pois(\lambda)$ to express that the random variable X is distributed as a Poisson with parameter p. The PMF of a Poisson distribution is written as

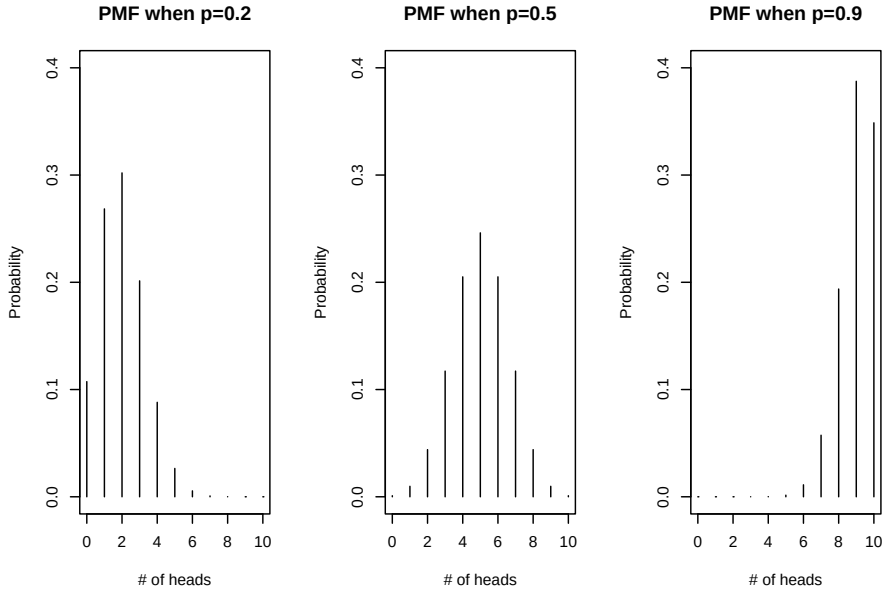

Figure 3.7: PMF for X, n=10, p varied

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{3.14}$$

for $k = 0, 1, 2, \cdots$. λ is sometimes called a rate parameter, as it is related to the rate of arrivals, for example, the number of that cross an intersection during a period of time.

3.5.3.1 Properties of Poisson

If $X \sim Pois(\lambda)$, then

$$E(X) = \lambda \tag{3.15}$$

and

$$Var(X) = \lambda. \tag{3.16}$$

These imply that larger values of a Poisson random variable are associated with larger variances. This is a common feature for count data. Consider the number of cars that cross an intersection during a one-hour time period. Consider the average number of cars during rush hour, say between 5 and 6pm. This average

number is large, but the number could be a lot smaller due to inclement weather, or the number could get a lot larger is there is a convention occurring nearby. On the other hand, consider the average number of cars between 3 and 4am. This average number is small, and is likely to be small all the time, regardless of weather conditions and whether special events are happening.

Another interesting property of the Poisson distribution is that it is skewed when λ is small, and approaches a bell-shaped distribution as λ gets bigger. Figure 3.8 below illustrates this:

```
##calculate probability of Poisson with these values on the support
x<-0:20
lambda<-c(0.5, 1, 4, 10) ##try 4 different values of lambda

##create PMFs of these 4 Poissons with different lambdas
par(mfrow=c(2,2))
for (i in 1:4)

{
  dens<-dpois(x,lambda[i])
  plot(x, dens, type="l", main=paste("Lambda is", lambda[i]))
}
```

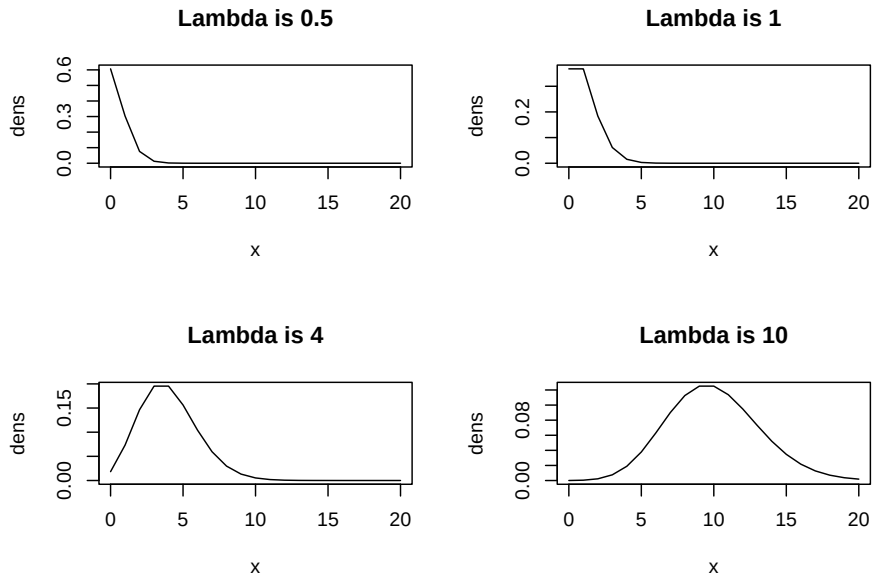

Figure 3.8: PMF for Poissons as Rate Parameter is Varied

3.5.3.2 Poisson Approximation to Binomial

If $X \sim Bin(n, p)$, and if n is large and p is small, then the PMF of X can be approximated by a Poisson distribution with rate parameter $\lambda = np$. In other words, the approximation works better as n gets larger and np gets smaller.

There are several rules of thumbs that exist to guide as to how large n should be and how small np should be. According to the National Institute of Standards and Technology (NIST), $n \geq 20$ and $p \leq 0.05$, or $n \geq 100$, and $np \leq 10$.

One of the main for using this approximation, instead of directly using the binomial distribution, is that the binomial coefficient can become compuationally expensive to compute when n is large.