Section 1.1: Introduction and Probability Concepts

Jared S. Murray
STA-371
McCombs School of Busines
The University of Texas at Austin

Suggested Reading:
OpenIntro Statistics, Chapters 2.1, 2.2, 2.4 (also review Ch 1)
R/RStudio resources on the course webpage

Getting Started

- Syllabus
- General Expectations
  1. Be on time and prepared
  2. Participate in discussions
  3. Complete homework assignments
  4. Get familiar with R and RStudio

Course Overview

Section 1: Probability & statistics review, introduction to simulation, and decision making under uncertainty

Section 2: Simple Linear Regression

Section 3: Multiple Linear Regression

Section 4: Forecasting and Time Series

Section 5+: Additional topics in modeling/simulation

Statistical computing

- We will use R for statistical analysis throughout the course
- This is industrial-strength, state-of-the-art, and free software for statistical computing
- Industrial-strength means industrial-strength: Google, J.P. Morgan, Whole Foods, Facebook, and even Microsoft use R
- We will access R through RStudio, a graphical interface for R

Getting Started with R and RStudio

Your first homework assignment (nothing to turn in!)

1. Visit the course website and complete the first three tutorials listed under "R/RStudio" resources

2. Run the R code accompanying these lecture notes. Try changing things and see what happens!

Let's start with a question...

My entire portfolio is in U.S. equities. How would you describe the possible outcomes for my returns in 2017?

Another question... (Targeted marketing)

Suppose you are deciding whether or not to target a customer with a promotion (or an ad)...

It will cost you \$.80 (eighty cents) to run the promotion and a customer spends \$40 if they respond to the promotion.

Should you do it? What if it cost \$80? Or \$35?

Introduction

Probability and statistics let us talk meaningfully about uncertain events.

- How likely is President Trump to finish a four year term?
- How much will Amazon sell next quarter?
- What will the return of my retirement portfolio be next year?
- How often will users click on a particular Facebook ad?

All of these involve inferring or predicting unknown quantities!

Random Variables

- *Random Variables* are numbers that we are NOT sure about, but have sets of possible outcomes we can describe.
- Example: Suppose we are about to toss a coin twice. Let X denote the number of heads we observe.

  Here X is the **random variable** that stands in for the number about which we are unsure.

Probability

Probability is a language designed to help us talk and think about random variables. The key idea is that to each **event** (one or more possible outcomes) we will assign a number between 0 and 1 which reflects how likely that event is to occur. For such an immensely useful language, it has only a few basic rules.

1. If an event A is certain to occur, it has probability 1, denoted $P(A) = 1$.
2. $P(A^C) = 1 - P(A)$. (A^C is "not-A")
3. If two events A and B are mutually exclusive (both cannot occur simultaneously), then $P(A \text{ or } B) = P(A) + P(B)$.
4. $P(A \text{ and } B) = P(A)P(B|A) = P(B)P(A|B)$

Probability Distribution

- We describe the behavior of random variables with a probability distribution, which assigns probabilities to events.
- Example: If X is the random variable denoting the number of heads in two *independent* coin tosses, we can describe its behavior through the following probability distribution:

$$X = \begin{cases} 0 & \text{with prob.} \quad 0.25 \\ 1 & \text{with prob.} \quad 0.5 \\ 2 & \text{with prob.} \quad 0.25 \end{cases}$$

- X is called a *Discrete Random Variable* as we are able to list all the possible outcomes
- Question: What is $Pr(X = 0)$? How about $Pr(X \geq 1)$?

Probability Distributions via Simulation

- This is a simple example, so we can compute the relevant probability distribution
- What if we couldn't do the math? Could we still understand the distribution of X?
- Yes - by simulaiton!

Quick intro to R

We can do more efficient simulations in R.

I'll show you some code today, but don't worry if it's hard to follow right now - we will get lots of practice.

R can be used as a calculator:

```
1+3

## [1] 4

sqrt(5)

## [1] 2.236068
```

Quick intro to R

We can save values for later, in specially named containers called **variables**

```
x = 5
print(x)

## [1] 5

x+2

## [1] 7
```

Quick intro to R

Variables can be numbers, vectors, matrices, text, and other special data types. We will only worry about a few of these.

```
y = "Hello"
print(y)

## [1] "Hello"

z = c(1, 3, 4, 7)
print(z)

## [1] 1 3 4 7

s = rep(1, 3)
print(s)

## [1] 1 1 1
```

Probability Distributions via Simulation in R

R has extensive capabilities to generate random numbers. The `sample` function simulates discrete random variables, by default giving equal probability to each outcome:

```
sample(c(1, 4, 5), size=4, replace=TRUE)

## [1] 1 4 4 5
```

Probability Distributions via Simulation

Let's simulate flipping a fair coin twice:

```
sample(x = c(0,1), size = 2, replace = TRUE)
```

```
## [1] 0 1
```

And a few more times:

```
sample(x = c(0,1), size = 2, replace = TRUE)
```

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

```
sample(x = c(0,1), size = 2, replace = TRUE)
```

```
## [1] 1 0
```

```
sample(x = c(0,1), size = 2, replace = TRUE)
```

Probability Distributions via Simulation

To approximate the probability distribution of X, we can repeat this process MANY times and count how often we see each outcome.

A “for loop” is our friend here (for now):

```
num.sim = 10000
num.heads.sample = rep(x = NA, times = num.sim)
for (i in 1:num.sim) {
    coinflips.result = sample(x = c(0, 1),
        size = 2, replace = TRUE)
    num.heads.sample[i] = sum(coinflips.result)
}
```

Aside: For Loops and Efficient Computing in R

- For speed and code readability reasons, R gurus usually recommend against for loops (including me!)
- **But** they are often more accessible to new users than the efficient alternatives.
- You can find more efficient implementations of some in-class simulations posted online; time permitting we will revisit them at the end of the course.

Aside: Packages in R

One powerful reason to use R is the number of user contributed packages that extend its functionality. (Rohit and I have both contributed R packages for general use!)

We'll use the `mosaic` package in R to simplify some common tasks, like simple repeated simulation:

```
library(mosaic)
num.heads.sample = do(num.sim) * {
  coinflips.result = sample(x = c(0, 1),
                            size = 2, replace = TRUE)
  sum(coinflips.result)
}
```

Probability Distributions via Simulation

Results (first 10 samples):

```
head(num.heads.sample, 10)

##    result
## 1       1
## 2       1
## 3       1
## 4       2
## 5       1
## 6       1
## 7       1
## 8       1
## 9       0
## 10      0
```

Probability Distributions via Simulation

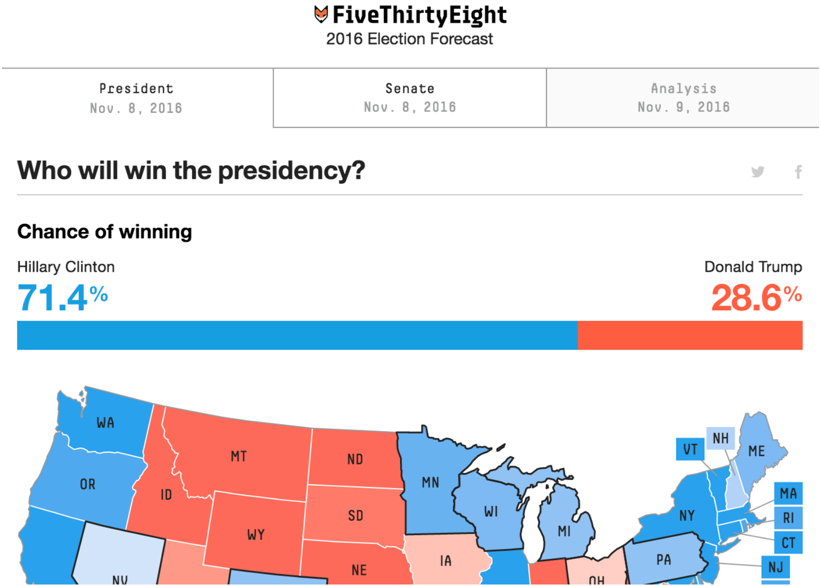
Results (summary):

```
table(num.heads.sample)

## num.heads.sample
##    0    1    2
## 2513 5015 2472

table(num.heads.sample)/num.sim

## num.heads.sample
##      0      1      2
## 0.2513 0.5015 0.2472
```

What have we done here? We:

- Set up a **model** of the world (The coin is fair, so $P(Heads) = 0.5$, and the tosses are independent)
- Understood the implications of that model through:
    1. Mathematics (probability calculations)
    2. Simulation

When we add the ability to incorporate **learning** about **uncertain** model parameters (statistics!) we have a powerful new toolbox for making **inference, predictions, and decisions**.



https://projects.fivethirtyeight.com/2016-election-forecast/

Pete Rose's Hitting Streak

Pete Rose of the Cincinnati Reds set a National League record of hitting safely in 44 consecutive games...

- Rose was a .300 hitter.
- Assume he comes to bat 4 times each game.
- Each at bat is assumed to be independent, i.e., the current at bat doesn't affect the outcome of the next.

What probability might reasonably be associated with a hitting streak of that length?

Pete Rose's Hitting Streak

Let A_i denote the event that "Rose hits safely in the i^{th} game"

Then P(Rose Hits Safely in 44 consecutive games) $=$
$P(A_1$ and $A_2 \ldots$ and $A_{44}) = P(A_1)P(A_2)...P(A_{44})$

We now need to find $P(A_i)$... It is easier to think of the complement of A_i, i.e., $P(A_i) = 1 - P(\text{not } A_i)$

$$\begin{aligned} P(A_i) &= 1 - P(\text{Rose makes 4 outs}) \\ &= 1 - (0.7 \times 0.7 \times 0.7 \times 0.7) \\ &= 1 - (0.7)^4 = 0.76 \end{aligned}$$

So, for the winning streak we have $(0.76)^{44} = 0.0000057$!!!
(Why?) (also, Joe DiMaggio's record is 56!)

New England Patriots and Coin Tosses

For the past 25 games the Patriots won 19 coin tosses!

What is the probability of that happening?

Let T be a random variable taking the value 1 when the Patriots win the toss or 0 otherwise.

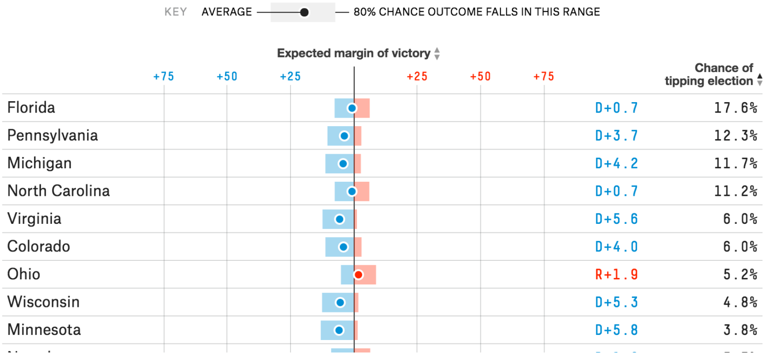
It's reasonable to assume $Pr(T = 1) = 0.5$, right??

Now what? It turns out that there are 177,100 different sequences of 25 games where the Patriots win 19... it turns out each potential sequence has probability 0.5^{25} (why?)

Therefore the probability for the Patriots to win 19 out 25 tosses is $177,100 \times 0.5^{25} = 0.005$

Trump's Victory: A Surprise?

Simplifying things: Trump had to win 5 states: Florida, Pennsylvania, Michigan, North Carolina, and Wisconsin.



Based on this info, what was the probability of a Trump victory? (538 said 0.29 - why?)

Conditional, Joint and Marginal Distributions

In general we want to use probability to address problems involving more than one variable at a time. We need to think carefully about *dependence* between the random variables!

Think back to our first question on the returns of my portfolio... if we know that the economy will be growing next year, does that change the assessment about the behavior of my returns?

We need to be able to describe what we think will happen to one variable relative to another...

Conditional, Joint and Marginal Distributions

Here's an example: We want to answer questions like: How are my sales impacted by the overall economy?

Let E denote the performance of the economy next quarter... for simplicity, say $E = 1$ if the economy is expanding and $E = 0$ if the economy is contracting (what kind of random variable is this?)
Let's assume $Pr(E = 1) = 0.7$

Conditional, Joint and Marginal Distributions

Let S denote my sales next quarter... and let's suppose the following probability statements:

S	$Pr(S \mid E=1)$	S	$Pr(S \mid E=0)$
1	0.05	1	0.20
2	0.20	2	0.30
3	0.50	3	0.30
4	0.25	4	0.20

These are called *Conditional Distributions*

Conditional, Joint and Marginal Distributions

s	$Pr(S = s \mid E = 1)$	S	$Pr(S = s \mid E = 0)$
1	0.05	1	0.20
2	0.20	2	0.30
3	0.50	3	0.30
4	0.25	4	0.20

- In blue is the conditional distribution of S given $E = 1$
- In red is the conditional distribution of S given $E = 0$
- We read: *the probability of Sales of 4 ($S = 4$)* **given(or conditional on)** *the economy is growing ($E = 1$) is 0.25*

Conditional, Joint and Marginal Distributions

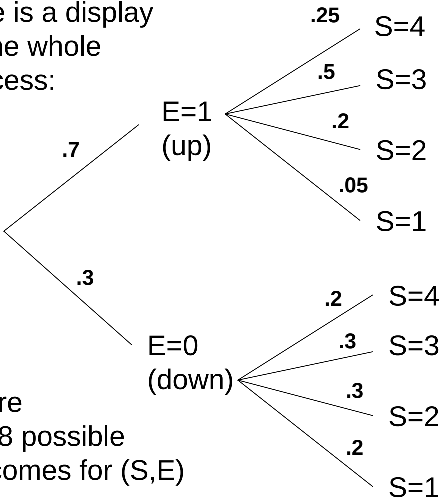
The conditional distributions tell us about about what can happen to S for a given value of E... but what about S and E jointly?

$$\begin{aligned} Pr(S = 4 \quad \text{and} \quad E = 1) &= Pr(E = 1) \times Pr(S = 4 | E = 1) \\ &= 0.70 \times 0.25 = 0.175 \end{aligned}$$

In English, 70% of the time the economy is growing. *1/4 of the times that the economy is growing* sales equals 4... 25% *of* 70% is 17.5%

Conditional, Joint and Marginal Distributions

here is a display of the whole process:



p(S=4 and E=1) = .7*.25 = .175

p(S=3 and E=1) = .7*.5 = .35

p(S=2 and E=1) = .7*.2 = .14

p(S=1 and E=1) = .7*.05 = .035

p(S=4 and E=0) = .3*.2 = .06

p(S=3 and E=0) = .3*.3 = .09

p(S=2 and E=0) = .3*.3 = .09

p(S=1 and E=0) = .3*.2 = .06

There are 8 possible outcomes for (S,E)

Conditional, Joint and Marginal Distributions

We call the probabilities of E and S together the joint distribution of E and S.

In general the notation is...

- $Pr(Y = y, X = x)$ is the joint probability that the random variable Y equals y **AND** the random variable X equals x.
- $Pr(Y = y|X = x)$ is the conditional probability that the random variable Y takes the value y **GIVEN** that X equals x.
- $Pr(Y = y)$ and $Pr(X = x)$ are the marginal probabilities of $Y = y$ and $X = x$ – that is, the probability that X equals x absent any knowledge of Y.

Conditional, Joint and Marginal Distributions

Why we call marginals marginals... the table represents the joint distribution and at the margins, we get the marginals.

		S				
		1	2	3	4	
E	0	.06	.09	.09	.06	.3
	1	.035	.14	.35	.175	.7
		.095	.23	.44	.235	1

Conditional, Joint and Marginal Distributions

Example... Given $E = 1$ what is the probability of $S = 4$?

		S				
		1	2	3	4	
E	0	.06	.09	.09	.06	.3
	1	.035	.14	.35	.175	.7
		.095	.23	.44	.235	1

$$Pr(S = 4|E = 1) = \frac{Pr(S = 4, E = 1)}{Pr(E = 1)} = \frac{0.175}{0.7} = 0.25$$

Conditional, Joint and Marginal Distributions

Example... Given $S = 4$ what is the probability of $E = 1$?

		S				
		1	2	3	4	
E	0	.06	.09	.09	.06	.3
	1	.035	.14	.35	.175	.7
		.095	.23	.44	.235	1

$$Pr(E = 1|S = 4) = \frac{Pr(S = 4, E = 1)}{Pr(S = 4)} = \frac{0.175}{0.235} = 0.745$$

Independence

Two random variable X and Y are *independent* if

$$Pr(Y = y|X = x) = Pr(Y = y)$$

for all possible x and y.

In other words,

knowing X tells you nothing about Y!

e.g., tossing a coin 2 times... what is the probability of getting H in the second toss given we saw a T in the first one?

Dependence (or "association") makes things more complicated...

The Berkeley gender bias case (Simpson's paradox)

Let $A =$ "admitted to Berkeley". In 1973 it was noted that $P(A \mid \text{male}) = 0.44$ while $P(A \mid \text{female}) = 0.35$. Meanwhile, individual departments showed no signs of discrimination. Suppose for now Berkeley has only chemistry and biology departments

	Chemistry	Psychology
$P(A \mid \text{female})$	0.6	0.3
$P(A \mid \text{male})$	0.5	0.25

What is going on?

The Law of Total Probability

We need one more useful probability tool, the law of total probability:

$$P(A) = P(A \text{ and } B) + P(A \text{ and } B^c)$$

where B^c is the *complement* of B (the event that "B doesn't happen")

The Berkeley gender bias case

For Females:

$$P(A) = P(A \mid \text{chem})P(\text{chem}) + P(A \mid \text{psych})P(\text{psych})$$

$$0.35 = 0.6P(\text{chem}) + 0.3(1 - P(\text{chem}))$$

$$\text{(hence)} \quad P(\text{chem}) = 0.167$$

For Males:

$$P(A) = P(A \mid \text{chem})P(\text{chem}) + P(A \mid \text{psych})P(\text{psych})$$

$$0.44 = 0.5P(\text{chem}) + 0.25(1 - P(\text{chem}))$$

$$\text{(hence)} \quad P(\text{chem}) = 0.76$$

The probability a student applies to chem *depends* on their gender.

The Berkeley gender bias case

The explanation for the apparent overall bias was that women have a higher probability of applying to Psychology than to Chemistry (assuming for simplicity that these are the only two options) and overall Psychology has a lower admissions rate!

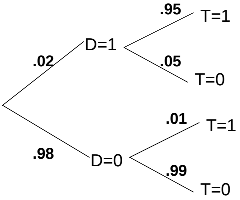
This is a cautionary tale! Before we can act on a apparent association between two variables (for example, sue Berkeley) we need to account for potential lurking variables that are the real cause of the relationship. We will talk a lot more about this... but keep in mind, association is NOT causation!
(Question: Should we conclude there was no gender bias in favor of men in admissions to Berkeley?)

Flipping the script: Inverting conditional probability

Often we have one conditional probability (say $P(A \mid B)$) but we want the other one (i.e. $P(B \mid A)$). Say we're testing for a disease.

Let $D = 1$ indicate you have the disease

Let $T = 1$ indicate that you **test** positive



$\Longrightarrow$

		D = 0	D = 1
T	0	.9702=.98*.99	.001
	1	.0098	.019

If you take the test and the result is positive, you are really interested in the question: Given that you tested positive, what is the probability you have the disease?

Disease Testing Example

		D	
		0	1
T	0	.9702	.001
	1	.0098	.019

$$Pr(D = 1 | T = 1) = \frac{0.019}{(0.019 + 0.0098)} = 0.66$$

Bayes theorem

The computation of $Pr(X = x|Y = y)$ from $Pr(X = x)$ and $Pr(Y = y|X = x)$ is called Bayes theorem...

$$\begin{aligned} Pr(X = x|Y = y) &= \frac{Pr(Y = y \text{ and } X = x)}{Pr(Y = y)} \\ &= \frac{Pr(Y = y \text{ and } X = x)}{\sum_{i=1}^{n} Pr(Y = y, X = x_i)} \\ &= \frac{Pr(X = x)Pr(Y = y|X = x)}{\sum_{i=1}^{n} Pr(X = x_i)Pr(Y = y|X = x_i)} \end{aligned}$$

In the disease testing example:

$\Pr(D = 1|T = 1) = \frac{p(T=1|D=1)p(D=1)}{p(T=1|D=1)p(D=1)+p(T=1|D=0)p(D=0)}$

$\Pr(D = 1|T = 1) = \frac{0.019}{(0.019+0.0098)} = 0.66$

Bayes theorem as population frequencies (aside)

- Try to think about this intuitively... imagine you are about to test 100,000 people.
- we assume that about 2,000 of those have the disease.
- we also expect 1% of the disease-free people to test positive, ie, 980, and 95% of the sick people to test positive, ie 1,900. So, we expect a total of 2,880 positive tests.
- Choose one of the 2,880 people at random... what is the probability that he/she has the disease?

$$\Pr(D = 1 | T = 1) = 1,900/2,880 = 0.66$$

We get the same answer!

Probability and Decisions

- So you've tested positive for a disease. Now what?
- Let's say there's a treatment available. Do you take it?
- What additional information do you need?