Introduction to Probability, R, and Simulation

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

Introduction

Probability and statistics let us talk meaningfully about uncertain events.

- What will Amazon's revenue be 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. To each **event** (one or more possible outcomes) we 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. $P(\sim A) = 1 - P(A)$. ($\sim A$ 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. $P(A \text{ and } B) = P(A)P(B \text{ given } A) = P(B)P(A \text{ given } B)$.

Probability

A little notation:

1. $P(A \text{ and } B)$ is called a joint probability (the probability both A and B happen), and we often just write $P(A, B)$.
2. $P(A \text{ given } B)$ is called a conditional probability – the probability that A happens, given that B definitely happens. We will write $P(A \mid B)$ for this conditional probability.

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 simulation!

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:

```r
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:

```
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: Packages in R

One powerful reason to use R is the number of user contributed packages that extend its functionality.

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

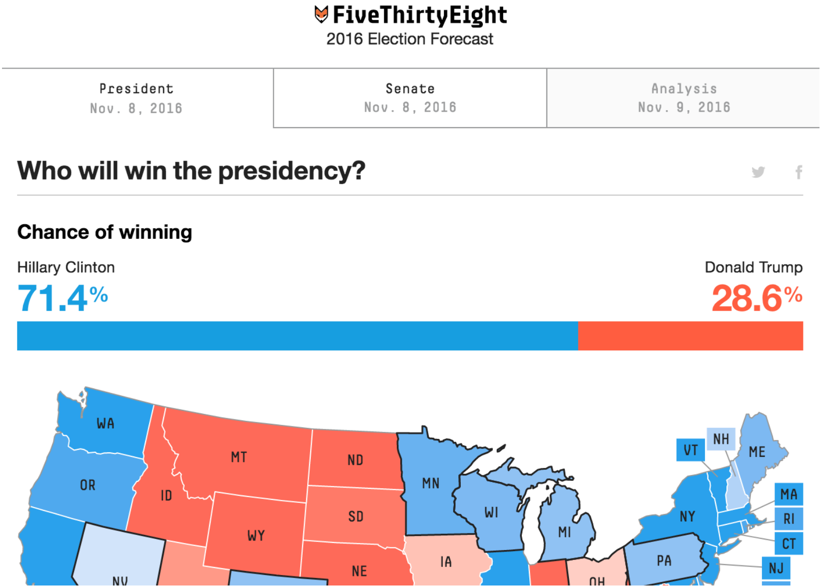
Results (summary):

```
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/