Lecture 7

Paul Holaway, Abhi Thanvi

June 30th, 2022

Lecture 6 Review

Before we move on to the new material, we will do a quick review of Lecture 6 content. Last time we introduced basic probability, how to calculate it by hand, and how to calculate it in `RStudio`.
Probability is defined as the likelihood that an event will occur. If the probability is close to 0, then the event is not likely to occur. If the probability is close to 1, then the event is likely to occur. If the probability is 0, then the event will never occur. If the probability is 1, then the event will occur.
We also learned some properties…

1. For any event A, $0 \leq P(A) \leq 1$.
2. For any event A, $P(A)$ is the sum of all probabilities of all of the outcomes in A.
3. $\sum_{i=1}^{n} P(A_i) = 1$, or the sum of the probabilities of all possible outcomes in a sample space is 1.
4. The probability of the empty set is 0 because there are no possible outcomes that can occur $P(\{\}) = P(\emptyset) = 0$

And some terminology…

1. Complement (A' or A^C) = Not
2. Union ($\cap$) = And
3. Intersection ($\cup$) = Or
4. Disjoint = Nothing In Common

Okay, now let's do a quick example both by hand and in `RStudio`.

Example 1; Probability Part 1 Review

Question 1: You are playing cards with your friends. You are this close to having a winning hand. All you need is a Queen or a Heart. What is the probability that you draw a Queen or a Heart. You cannot see any other hand so therefore the chances of drawing any one specific card (4 of Spades for example) is $\frac{1}{52}$.

Answer: $P(Q \cup H) = P(Q) + P(H) - P(Q \cap H)$
$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}$

Question 2: Now let's do the same thing in `RStudio`, the sample space has been set up for you.

```
##Do NOT Delete This Code!!
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
suit = c(rep("S",13),rep("H",13),rep("C",13),rep("D",13))
num = c(seq(2,10,1),"J","Q","K","A",seq(2,10,1),"J","Q","K","A",seq(2,10,1),
        "J","Q","K","A",seq(2,10,1),"J","Q","K","A")
deck = data.frame(suit,num)
```

```
Q = deck %>% filter(num == "Q")
H = deck %>% filter(suit == "H")
QH = deck %>% filter(num == "Q" & suit == "H")
length(Q$suit)/length(deck$suit) + length(H$suit)/length(deck$suit) -
  length(QH$suit)/length(deck$suit)
```

```
## [1] 0.3076923
```

```
16/52
```

```
## [1] 0.3076923
```

```
QH2 = deck %>% filter(num == "Q" | suit == "H")
length(QH2$suit)/length(deck$suit)
```

```
## [1] 0.3076923
```

Excellent, we got the same answer as before. Again, you can calculate your probabilities anyway that you want. We are just giving you options to do so.

Okay, now we can move onto the next portion of lecture content.

Probability Part 2

Today will (hopefully) be less of an info-dump and more examples than last time. We will be covering three main topics.

1. The idea of independence.
2. Multiplication Rule
3. Conditional Probability

First we will start with the idea of independence.

Independence

The idea of independence is unique to probability. No other field has anything like it, so many fields have adopted it. Events are said to be independent if they are not affected by the occurrence of another event or events. The most classic example of this is flipping a coin. Whatever you flipped the first n times, the probability of the next flip being heads or tails is not going to change. This then leads up to half of the multiplication rule. We need conditional probability for the other half, so we will hold on that for now.

Multiplication Rule (Part 1)

Multiplication rule is when we want to look at more complicated probabilities, such as the probability that two things happen at the same time. The multiplication rule has two cases...

- When the events are independent.
- When the events are not independent.

Independent events with the multiplication rule are easy. If the events are independent, then $P(A \cap B) = P(A) * P(B)$. Or, if two (or more) events are independent, then you simply multiply the probabilities of each of them happening together.

Example 2; D&D

You are playing D&D with your friends and are currently facing a low-level boss. To determine if your attacks hit, you roll a 20-sided dice (d20). You are pretty certain it will take two more successful attacks to take down the boss. What is the probability that you and your friend who will attack next both hit? You need to roll higher than a 10 for your attack to be successful and your friend needs to roll higher than a 14 for their attack to be successful.

Answer: $P(H_1 \cap H_2) = P(H_1) * P(H_2)$ because each dice roll is independent.
Therefore $P(H_1 \cap H_2) = \frac{10}{20} * \frac{6}{20} = \frac{60}{400} = 0.15$

Pretty easy, but let's do one more example before moving onto conditional probability.

Example 3; Beach Visit

A CEO is trying to relax on the weekends. His employees are constantly calling him asking questions about projects they are working on. At first, he tries to avoid them by going to restaurants, but they know which ones he frequents so that did not work. His new plan is to go to the beach on Saturdays (at least for the summer). To make sure his employees cannot find him easily, he decides to randomly select a beach each week. One of the company projects uses the `bwq` data set, so he decides to randomly select one row and whatever beach it is go there.

```
bwq <- read.csv("~/Desktop/DPISu22/Data Sets/bwq.csv", stringsAsFactors=TRUE)
```

Question 1: What are the probabilities of each beach being selected?

```
table(bwq$Beach.Name) #To remind us what the beach names are.
```

```
##
## 63rd Street Beach     Calumet Beach    Montrose Beach Ohio Street Beach
##              3420              7570              7269              9343
##    Osterman Beach     Rainbow Beach
##              4023              3298
```

```r
length(bwq$Beach.Name) #How many observations we have (denominator).
```

```
## [1] 34923
```

We can take a bit of a shortcut here. Using the `table()` function we can see how many of each beach there are and using the `length()` function, we can see how many observations there are. This means we can easily calculate the probabilities.

```r
3420/34923 #63rd St.
```

```
## [1] 0.09792973
```

```r
7570/34923 #Calumet
```

```
## [1] 0.2167626
```

```r
7269/34923 #Montrose
```

```
## [1] 0.2081436
```

```r
9343/34923 #Ohio St.
```

```
## [1] 0.2675314
```

```r
4023/34923 #Osterman
```

```
## [1] 0.1151963
```

```r
3298/34923 #Rainbow
```

```
## [1] 0.09443633
```

Question 2: The plan works, for a bit. However, at the end of June, the CEO's employees overhear his plan from his assistant. Two of his employees (Greg and Sally) each come up with their own idea on what to do for July. Greg decides to alternate between Ohio St. and Calumet as they are the most frequent in the data set. Sally decides to just go to Ohio St. every Saturday. What are the probabilities that Greg's and Sally's selections would happen? There are **5** weekends in July.

```r
#Greg
(9343/34923) * (7570/34923) * (9343/34923) * (7570/34923) * (9343/34923)
```

```
## [1] 0.0008996904
```

```r
#Sally
(9343/34923)^5
```

```
## [1] 0.001370484
```

Note that this is **NOT** the probability of them seeing their boss, this is the probability that they see their boss all 5 days. However, because the events are all independent, then we can use the rules of probability to figure the probability that they will see their boss at least once.
First, let's do the math...
$P(X \geq 1) = 1 - P(X = 0)$ We want the probability that there are at least one or more correct guesses, which then would translate to 1 - the complement of this, which is no correct guesses.

```
#Greg
1 - (1 - 9343/34923) * (1 - 7570/34923) * (1 - 9343/34923) * (1 - 7570/34923) *
  (1 - 9343/34923)
```

```
## [1] 0.7589241
```

```
#Sally
1 - (1 - 9343/34923)^5
```

```
## [1] 0.7891639
```

Okay, so there is a pretty good chance they will find their boss at least once (not good for him). However, notice how Sally's strategy has a bit higher of a chance to work. Why is this? Something to think about for later.

Conditional Probability

Okay, now for the big one. Conditional probability is the probability that an event A occurs given that an event B has already occurred. The formula to calculate this is $P(A|B) = \frac{P(A \cap B)}{P(B)}$. Conditional probability is important to know when your events are *not* independent. If they were independent, whatever happened for event B would not matter. Let's do a quick example to illustrate this.

Example 4; Rolling the Dice

Let's say you are rolling a six-sided die and recording the number. What is the probability of rolling a 1 given you rolled an odd number on the last roll?

Answer: $P(R_2 = 1|R_1 = Odd) = \frac{P(R_2=1 \cap R_1=Odd)}{P(R_1=Odd)} = \frac{1/6*3/6}{3/6} = \frac{1}{6}$

Example 5; The Changing Labor Force

Over the past several decades, the nature of the US labor force has changed dramatically. More women are searching for jobs, more men are staying home with children, and senior citizens are remaining in their jobs longer. According to the US Census Bureau, for married-couple family groups, 84.9% of all fathers are employed; and in 57.5% of these households, both parents are employed. Suppose a married couple family group is selected at random. If the father is employed, what is the probability that the mother is also employed?
Answer: $P(M|F) = \frac{P(M \cap F)}{P(F)} = \frac{0.575}{0.849} \approx 0.6733$

Multiplication Rule (Part 2)

Okay, now how does this relate to the multiplication rule? Well, let's look at what happens when we manipulate the equation for conditional probability.

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

Okay, we now have the equation for multiplication rule when events are dependent. One thing to note, when the events are independent, then remember that whatever happens in event B does not affect the probability of event A $\implies P(A|B) = P(A)$. When you plug that into the formula above, we get $P(A \cap B) = P(A)P(B)$ which is what we had earlier. Let's look at some examples.

Example 6; Card Drawing

Let's say now you want to draw two cards randomly from a deck.

Question 1: What is the probability that you will draw two Hearts?

Answer: $P(H \cap S) = P(H_2|H_1)P(H_1) = \frac{13}{52} * \frac{12}{51} \approx 0.0588$

Question 2: What is the probability that you draw an Ace then a King?

Answer: $P(K_2 \cap A_1) = P(K_2|A_1)P(A_1) = \frac{4}{52} * \frac{4}{51} \approx 0.0060$

Question 3: What is the probability of drawing two Aces?

Answer: $P(A_2 \cap A_1) = P(A_2|A_1)P(A_1) = \frac{4}{52} * \frac{3}{51} \approx 0.0045$

What we have done today is just the tip of the iceberg. There are entire courses on probability you can take if you decide to in college. There are many who can get their PhDs doing probability research. (Way beyond the scope of this course and way beyond the scope of me.) Hopefully this gets you the introduction that you need.

End of Lecture 7 Notes