Lecture 6

Paul Holaway, Abhi Thanvi

June 29th, 2022

Lecture 5 Review

Before we move on to the new material, we will do a quick review of Lecture 5 content. Last time we learned about plots and sample spaces. Creating plots is a good tool to have because visualizing your data is important. Many people (myself included) prefer to see a type of data visualization instead of reading walls of text or output. Let's quickly review the three types of plots we covered last time.

Example 1; Plots Review

Here we are using the UIUC GPA data set. This has the GPAs for different courses at UIUC from Spring 2010 to Spring 2020. There are a few columns that are imported funny, but we can fix those because we know how to rename columns (#DataCleaningReview).

```
library(tidyverse)
```

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

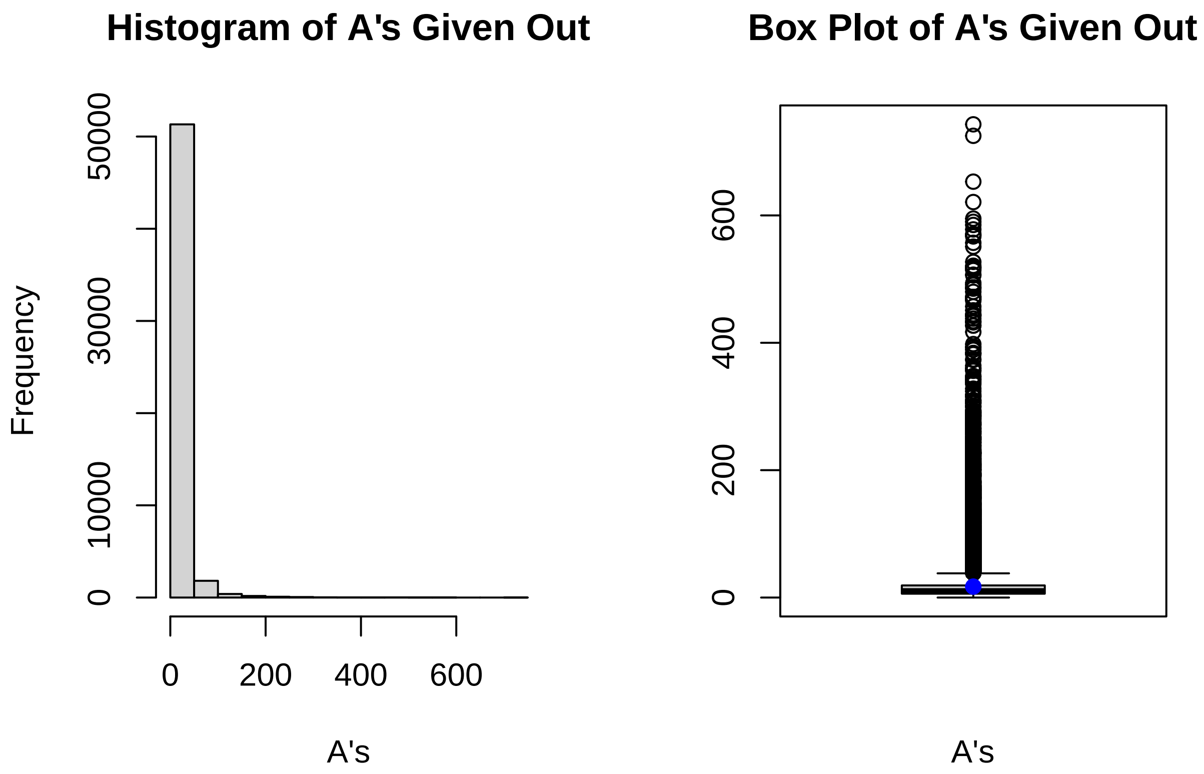
```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

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

```
gpa <- read.csv("~/Desktop/DPISu22/Data Sets/gpa.csv", stringsAsFactors=TRUE)
gpa = gpa %>% rename("A+" = "A.") %>% rename("A-" = "A..1") %>% rename("B+" = "B.")
gpa = gpa %>% rename("B-" = "B..1") %>% rename("C+" = "C.") %>% rename("C-" = "C..1")
gpa = gpa %>% rename("D+" = "D.") %>% rename("D-" = "D..1")
```

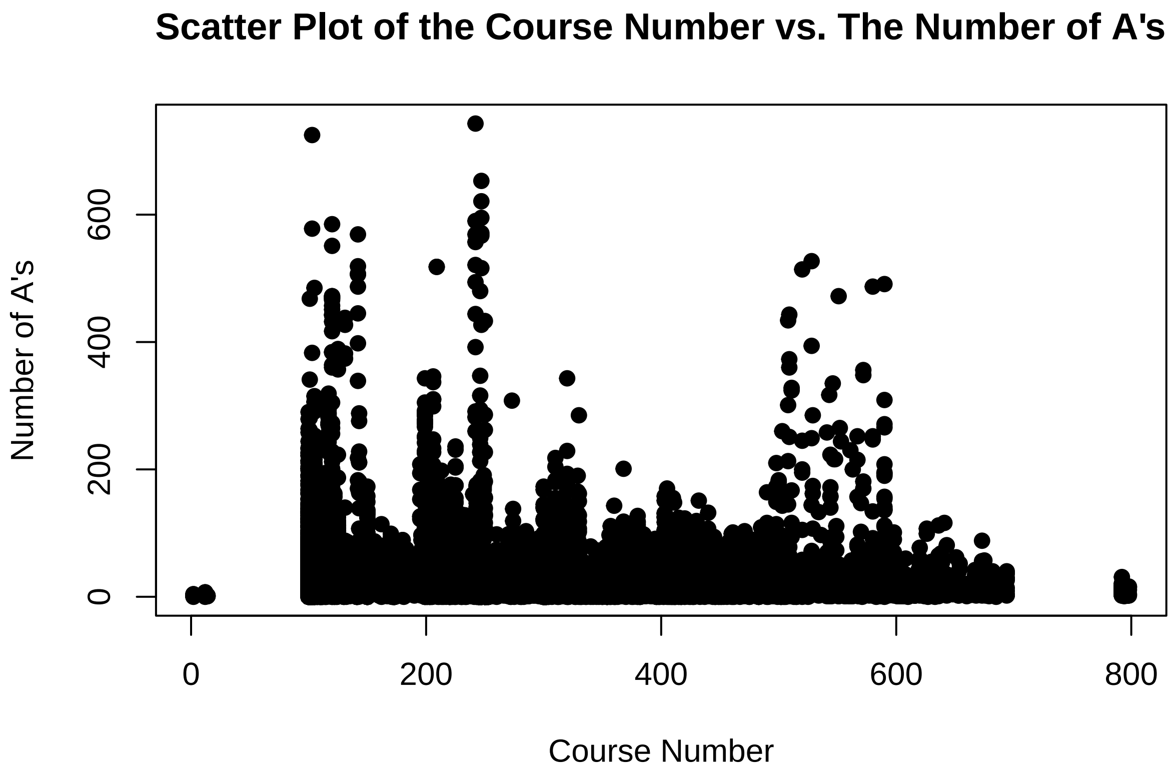
There we go, all fixed. Now we can start plotting. Let's start by making a histogram and box plot of the `A` column.

```r
par(mfrow = c(1,2))
hist(gpa$A, xlab = "A's", main = "Histogram of A's Given Out")
boxplot(gpa$A, xlab = "A's", main = "Box Plot of A's Given Out")
points(mean(gpa$A), col = "blue", pch = 19)
```



Both results are heavily right-skewed, or we can see that there are some, but very few courses that give out a large amount of A's. Now let's make a scatter plot for `A` against the course number. One would think there would be less A's given out in higher level courses because they should be harder.

```
plot(gpa$Number,gpa$A,xlab = "Course Number", ylab = "Number of A's",
     main = "Scatter Plot of the Course Number vs. The Number of A's", pch = 19)
```



Well this is unexpected. There seems to be no correlation whatsoever between the course number and the number of A's. What could be a lurking variable here? Something to think about.

Example 2; Sample Space Review

Now let's quickly review sample space. Remember, the sample space is just a list of all the possible outcomes something can have. For the GPA data set, what would be the sample space for the grades?

Answer: $S = \{A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-, F\}$ We would not include W because it is not a grade, just a transcript note that the student withdrew from the course. Whenever you are figuring out the sample space, sometimes background knowledge is required. There may come times that you will have to look up information to help you figure out the sample space.

For example, the `Year`. If I had not told you earlier what the range was, you would have had to either look up the data set or use what you currently know about data science to figure it out. What would the sample space be for `Year`.

Answer: $S = \{2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020\}$

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

Probability Part 1

What is Probability?

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. How how do we calculate this? In the most simple form possible…

$$P(Event) = \frac{No.of Occurrences}{Total No.of Outcomes}$$

To figure out both the numerator and the denominator, we first need to figure out what the sample space is. Once you have the sample space, you can easily figure out the total number of outcomes and the number of occurrences. Now before we can go onto calculating probabilities, we first need to know some properties about probability.

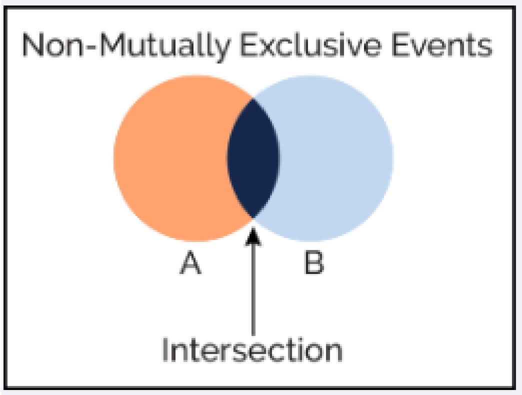
1. For any event A, $0 \leq P(A) \leq 1$.
2. For any event A, $P(A)$ is the sum of all probabilities of all of the outcomes in A.
3. $\sum_{i=1}^{n} P(A_i) = 1$, or the sum of the probabilities of all possible outcomes in a sample space is 1.
4. The probability of the empty set is 0 because there are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

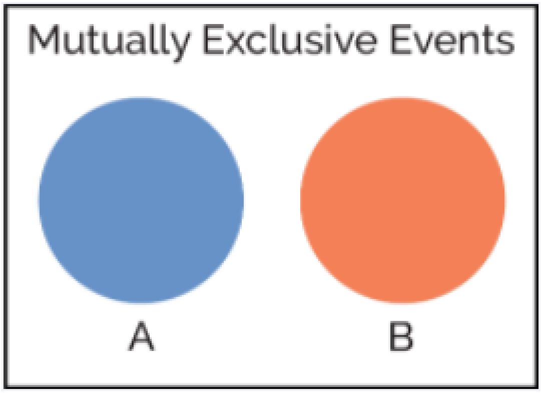
Now let's learn some more probability terminology. This will be useful later today and heavily tomorrow.

1. The event ***A Complement***, denoted by A' or A^C, consists of all outcomes in the sample space S that are *not* in A.
2. The event ***A Union B***, denoted by $A \cup B$, consists of all outcomes that are in A *or* B.
3. The event ***A Intersection B***, denoted by $A \cap B$, consists of all outcomes that are in A *and* B.
4. If A and B have no elements in commons, they are ***Disjoint*** or ***Mutually Exclusive***, denoted by $A \cap B = \{\}$.

That's a bit wordy, so let me give you some tips to make it easier.

1. Complement = Not
2. Union = And
3. Intersection = Or
4. Disjoint = Nothing In Common





Okay, that was a lot of information just now. Let's start doing some examples before I go onto the next portion of lecture content.

Example 3; Sales Person Selection Probability

Let's suppose you are in charge of doing a performance review for your company on the sales people in your department. There are five people in total. You are required to randomly select one of the five (to be fair) this month, the others will be reviewed in later months. However, your boss has altered the selection probabilities such that the ones who are not doing as well are less likely to be chosen at the moment. He wants to give them more time to try to get up to the company's standards. The table of selection probabilities is below.

Sales Person	Able	Baker	Charlie	Diana	Evelyn
Probability of Selection	0.30	0.20	0.10	0.05	0.35

Question 1: What is the probability that Baker is selected?

Answer: $P(Baker) = 0.20$

Question 2: What is the probability that a man is selected?

Answer: $P(Man) = 0.30 + 0.20 + 0.10 = 0.60$

Question 3: What is the probability that a woman is selected?

Answer: $P(Woman) = 0.05 + 0.35 = 0.40 \ OR \ P(Woman) = 1 - P(Woman^C) = 1 - P(Man) = 1 - 0.60 = 0.40$

Example 4; Unfair Die

A mafia boss has a game in his casino called "No One Wins". If you roll a 1, you lose everything. If you roll a 2, you lose 50%. If you roll a 3, you lose 25%. If you roll a 4, you lose 10%. If you roll a 5, you lose 5%. If you roll a 6, you win 10X as much as you bet. Everyone flocks to the casino to play this not realizing the mob boss weighted the die (thus making it unfair). The table of probabilities is listed below.

DISCLAIMER: Paul and Abhi do not endorse gambling.

Value	1	2	3	4	5	6
Expected Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
Actual Probability	0.20	0.20	0.18	0.18	0.16	0.08

Question: What is the probability of losing money?

Answer: $P(L) = 0.20 + 0.20 + 0.18 + 0.18 + 0.16 = 0.92 \ OR \ P(L) = 1 - P(W) = 1 - 0.08 = 0.92$

Addition Rule

The previous two examples were when each event was mutually exclusive. However, how do we calculate probabilities when the events are not mutually exclusive? Well, it makes things a little bit complicated, but not much. We just need to make sure that we do not double count events if the two are not mutually exclusive. This is where the addition rule comes in. It is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Before when the events were mutually exclusive, $P(A \cup B) = P(A) + P(B)$ because $P(A \cap B) = 0$ if two events are mutually exclusive.

Example 3; Continued

Let's suppose your boss's boss found out about the unfair selection process and is not happy. Thankfully you are not in trouble. Now, you have to review two people randomly this month and all much have equal selection probabilities. However, if you randomly select the same person twice, you will have to do a full audit on that person's sales records. You will be fired if your boss's boss finds out you did not follow these instructions. I have created the sample space below.

```
#This creates the sample space for you to view. DO NOT DELTE THIS CODE!!
review1 = c("A","A","A","A","A","B","B","B","B","B","C","C","C","C","C","D","D",
            "D","D","D","E","E","E","E","E")
review2 = c("A","B","C","D","E","A","B","C","D","E","A","B","C","D","E",
            "A","B","C","D","E","A","B","C","D","E")
review = data.frame(review1,review2)
review = review %>% rename(Person1 = "review1") %>% rename(Person2 = "review2")
```

Question 1: Charlie and Diana are clearly struggling to meet the company's standards. What is the probability that both will be selected? It might be useful to look at the sample space for this one.

Answer: $P(C \cap D) = \frac{2}{25} = 0.08$, either can be selected first or second.

Question 2: You like both Charlie and Diana and do not want to review either of them. What is the probability that either will be selected (and most likely lose their jobs)? Again, looking at the sample space may help.

Answer: $P(C \cup D) = P(C) + P(D) - P(C \cap D) = 0.2 + 0.2 - 0.08 = 0.32$

Example 4; Continued

The mafia boss has heard that a government official not yet in his pocket has heard about people complaining that the games at his casino are fixed. After numerous attempts to bribe him fail, he decides to change the games to be fair, but much harder to win anything. His right-hand man recommends they change the "No One Wins" game to two dice instead of one. If a 1 is rolled on either roll, then lose, otherwise they win. The boss wants to know if this game benefits him or not. I have created the sample space below again.

```
#This creates the sample space for you to view. DO NOT DELTE THIS CODE!!
roll1 = c(seq(1,6,1),seq(1,6,1),seq(1,6,1),seq(1,6,1),seq(1,6,1),seq(1,6,1))
roll2 = c(rep(1,6),rep(2,6),rep(3,6),rep(4,6),rep(5,6),rep(6,6))
roll = data.frame(roll1,roll2)
```

Question: What is the probability of winning the game?

Answer: The probability of not rolling a 1 on either roll. $P(W) = 1 - P(L)$,
$P(L) = P(R_1 = 1 \cup R_2 = 1) = P(R_1 = 1) + P(R_2 = 1) - P(R_1 = 1 \cap R_2 = 1) = \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36}$
$\implies P(W) = 1 - \frac{11}{36} = \frac{25}{36}$

Okay, now let me show you another way you can do this. We made the sample spaces for you, so we can use `RStudio` and conditionals/filtering (#Lecture2) to do it for us. Then you do not have to type it into a really expensive TI-84 calculator (or buy one because `R`/`RStudio` is free).

Example 3; Continued Again

Let's again calculate the two probabilities using `RStudio`. We want to figure out the probability that we select Charlie and Diana. To do this in `RStudio`, we just need to know two symbols, `&` and `|`. This is the "and" symbol and the "or" symbol. When introduced into filtering with conditionals, we can do multiple at once.

Question 1: For the first question, we want to have Charlie or Diana be the first one selected and Charlie or Diana be the second person selected. This can be done using the following code.

```
ex3 = review %>% filter((Person1 == "C" | Person1 == "D") &
                           (Person2 == "C" | Person2 == "D")) %>% filter(Person1 != Person2)
#The second filter is to remove when they are selected twice (full audit).
ex3
```

```
##   Person1 Person2
## 1       C       D
## 2       D       C
```

Great, now we have a list of possible outcomes. Now we need to divide that by the total number of possible outcomes. You know it is 25, but let's say you do not know. We have saved the sample space as `review`, so we can then use a function to calculate how big our sample space is. It is the `length()` function. Let's use it below.
Note: If the subset of the sample space is a data frame, then you will have to use the following syntax for this to work. `length(sub$x)/length(samplespace$x)` where `x` is any common variable between the two. Also, notice how you are going back to the basic definition of probability.

```
PCD = length(ex3$Person1)/length(review$Person1)
PCD
```

```
## [1] 0.08
```

Question 2: Okay, now let's do the second question.

```
C = review %>% filter(Person1 == "C")
D = review %>% filter(Person1 == "D")
PC = length(C$Person1) / length(review$Person1)
PD = length(D$Person1) / length(review$Person1)
PC
```

```
## [1] 0.2
```

```
PD
```

```
## [1] 0.2
```

```
PCD
```

```
## [1] 0.08
```

```
PC + PD - PCD
```

```
## [1] 0.32
```

Hey look at that. We got the same probabilities.

Example 4; Continued Again

Let's again calculate the this probability using `RStudio` and conditionals. So we want to calculate the probability of rolling a 1 on either roll.

```
One = roll %>% filter(roll1 == 1 | roll2 == 1)
1 - length(One$roll1) / length(roll$roll1)
```

```
## [1] 0.6944444
```

```
25/36
```

```
## [1] 0.6944444
```

Again, same probabilities as before. The mafia boss has decided to change the game to "Every One Wins" where you win if one of the dice is a 1. That way the probability of winning is $\frac{11}{36}$.

Now you many have noticed that Example 3 was more complicated to code than calculate out, whereas Example 4 was the opposite. Sometimes it will be easier to do a problem by hand than to calculate it out, sometimes the other way around. We are simply showing you tools to calculate probabilities so you can use what works best for you. For all labs and the project you may use either or both methods for a question. If you do either method correctly you will get the same correct answer. Next time we will go into some more difficult probability basics.

End of Lecture 6 Notes