

[Tan et al. \(2018\)](#) were interested in how pet dogs bond with complete strangers. In one of their studies, they placed a food treat under one of two cups (the dog did not know which bowl had the food) and had a stranger point the dog toward the cup with the food. Although the researchers did many trials to see how trust evolved, one of the results they analyzed was whether dogs would pick up the cue from the stranger on the very first trial. A total of 53 dogs were tested.

In the study, 27 of the 53 dogs picked the cup with the food on the first trial.

**Question #1** According to the Neyman-Pearson paradigm we have been working with in Labs 8 and 9, should we accept the null hypothesis  $p = 0.5$  or the alternative hypothesis  $p = 0.7$ ? Why?

We accept the null because the amount of dogs who went to the food does not lie in the critical region determined in Labs 8 and 9.

If dogs do not trust strangers initially, they would be expected to choose between the two cups more-or-less at random on the first trial. However, if dogs tend to trust strangers, they would be expected to choose the cup with food more often than would be suggested by chance alone.

**Question #2** Using the parameter  $p$ , write the null and alternative hypothesis according to the “null hypothesis significance testing” paradigm. Label which one is the null hypothesis ( $H_0$ ) and which is the alternative ( $H_a$ ).

Null Hypothesis: when the dogs pick the food cups on their own

Alternative: when the dogs follow the stranger’s indication where the food is

Now we will use R to set up the distribution of our test statistic under the null hypothesis. Open up a new script and assign the following variables:

```
> n <- # the sample size from question 3
> p <- # the value of p if the null hypothesis in Question 2 is correct
```

In the experiment, 27 of the 53 dogs picked the cup with the food on the first trial. Let’s define that now:

```
> X <- 27 # the number of dogs who got the food on the first trial
```

**Question #3** To compute the p-value according to the “null hypothesis significance testing” paradigm, which of the events below should we find the probability of? Justify your answer.

- a) Exactly 27 dogs getting the cup with food
- b) 27 or fewer dogs getting the cup with food
- c) 27 or more dogs getting the cup with food

The more dogs that go, there is even more evidence that the statement is true.  
To find the p-value, type one of the following lines:

If you chose answer (a) to **Question #5**:

```
> dbinom(X, n, p) # probability of getting exactly X successes out of n trials with probability p
```

If you chose answer (b) to **Question #5**:

```
> pbinom(X, n, p, lower.tail = TRUE) # probability of getting X or fewer successes
```

If you chose answer (c) to **Question #5**:

```
> pbinom(X, n, p, lower.tail = FALSE) + dbinom(X, n, p) # probability of getting more than X successes, plus probability of getting exactly X successes = probability of X or more successes
```

**Question #4** According to R, what is the p-value?

The value from R is 0.5

**Question #5** Using a significance level of 0.05, should you reject the null hypothesis or fail to reject the null hypothesis? Justify your answer.

We should fail to reject the null hypothesis because there is not enough evidence to change our minds.

**Question #6** Can you conclude that dogs do not trust strangers? Can you conclude that dogs do trust strangers? Or can you not make a conclusion either way? Justify your answer.

We did not find any evidence to reject the null hypothesis.

**Question #7** Would you make the same conclusion using the Neyman-Pearson testing paradigm? Why or why not?

You cannot make the same conclusion because you did not explicitly reject the null hypothesis because the alternative cannot be proven true.

Now we will use the `binom.test` command to “automatically” do a binomial hypothesis test. Any time you do a binomial test in R, you should first separately define the variables representing the number of successes, the number of trials, and the probability of success on a trial. In this lab, we’ve already defined those variables as `X`, `n`, and `p`, respectively.

```
> binom.test(X, n, p, alternative = "one letter goes here") # inside the quotation marks, type l if your alternative hypothesis has a < sign or g if it has a > sign
```

**Question #8** Paste the console output below.

```
data: X and n
number of successes = 27, number of trials = 53, p-value = 0.6081
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.6290759
sample estimates:
probability of success
 0.509434
```

**Question #9** Identify the value of the test statistic and the p-value from the output. Are these values the same as you got using the probability calculations?

The value of the test statistic is 27 and the value of the p-value is 0.5 . These are the same values found above.

For Lab 11 and 12, you will collect your own data. For Lab 11, we will analyze the colors of Starburst. For Lab 12, we will analyze the personality types of statistics students, but we will collect the data now.

To collect the personality type data, please open the following link in your browser window: <http://www.16personalities.com/free-personality-test>, and complete the 10-15 minute personality test. Once you complete the test, your personality will be classified into one of 16 “types,” but we will consider only the 4 main personality categories: Analysts, Diplomats, Sentinels, and Explorers. The figure below shows someone classified as an “Architect,” which is one of the Analyst types as shown by the highlighting. Note that you may have to click on “Start Reading” to find this type. While you are waiting for the rest of the class to finish, feel free to peruse your “results,” or to click on the Personality Types tab to look at all of the possible types.



**Question #1** Once you obtain your individual results, complete the Questionnaire on Titanium for your personality type, major, and the number of Starbursts in each color you received.

In this lab, we will test the claim that all four main colors are equally likely in the population of all Starburst.

**Question #2** Write out the null hypothesis for this goodness-of-fit test.

There should be an equally likelihood of obtaining each color.

**Question #3** If the null hypothesis is true, how many of each Starbursts type would we expect to see in our sample?

There should be a 25% chance of being a certain color Starbursts.

**Question #4** If the sample size assumptions are met, what would be the sampling distribution of your test statistic (i.e., what is the type of distribution and the degrees of freedom)?

The degrees of freedom would be number of categories - 1 which is going to be  $4 - 1 = 3$ . There are 3 degrees of freedom.

Let's put in our data and running the goodness-of-fit test in R.

```
> Starburst <- c(11, 13, 17, 21) # fill this in with the actual numbers
> probs <- c() # fill this in with your probabilities under H0
```

**Question #5** Are the sample size assumptions are met for a chi-square goodness of fit test? Show how you checked the assumptions.

The sample size assumptions are met because simulation does not provide a degree of freedom and because the expected count of each color is more than 5.

If the sample size assumptions are met, run the following code to perform a chi-square goodness of fit test:

```
> chisq.test(Starburst, p = probs)
```

Otherwise, add the following argument to perform a goodness of fit test by simulation:

```
> chisq.test(Starburst, p = probs, simulate.p.value = TRUE)
```

**Question #6** Copy the RStudio output below.

Chi-squared test for given probabilities

data: star

X-squared = 7.4545, df = 3, p-value = 0.05874

**Question #7** What is the value of the chi-square test statistic as computed by R?

The value of chi-squared will be 7.4545

**Question #8** What is the p-value for this test?

The p-value is 0.05874.

**Question #9** Using a 5% significance level, what can you conclude about the distribution of Starburst?

We just barely crossed the threshold of it being reasonable so we can still keep the assumption that it is equally likely to pick one of four colors which is 25%.

**Question #10** Do you believe that your conclusion (from **Question #9**) applies to the population of all Starburst? (HINT: Think about the sample we used and the way we collected the data)

This will most likely not represent the entire population of all Starburst because there is not enough elements in the sample.

## Lab 12

Jared Dyreson

TR @ 11:30 - 14:15

1. Write the null hypothesis for this test of independence.
  - **Someone's personality is strictly agnostic from their major, contributing in no way, shape or form**
2. If the sample size assumptions are met (all expected counts  $> 5$ ), what would be the sampling distribution of your test statistic (i.e., what is the type of distribution and the degrees of freedom)?
  - **The sampling distribution of my test statistic would be  $\chi^2$  and the degrees of freedom is  $(4 - 1)(3 - 1) = 6$**
3. If the null hypothesis is true, calculate the expected number of analysts who are Computer Science majors. If your number is not an integer, round it to at least one decimal place.
  - $\frac{20}{48} \times 7 = 2.916666667$ , **where 20 out of the 48 students are computer science majors and there are 7 total diplomats in the class**
4. If the null hypothesis is true, calculate the Pearson residual and contribution to the chi-squared statistic for analysts who are Computer Science majors.
  - $\frac{O-E}{\sqrt{E}} = \frac{5-2.92}{\sqrt{2.92}} = 1.217$
5. To obtain the p-value, can we use the sampling distribution from Question #2, or do we have to simulate a sampling distribution? Explain your reasoning. (HINT: look at your answer to Question #3)

```
majors <- read.csv("~/Downloads/majors.csv")
majors.table <- xtabs(~ Major + Personality.Type, data = majors)
```

- **We would need to simulate because the result from question 3 is 2.9 which is less than 5. We need to proceed with method two.**

6. Copy the RStudio output below.

```
chisq.test(majors.table, simulate.p.value = TRUE)
```

Pearson's Chi-squared test with simulated  
p-value (based on 2000 replicates)

```
data: majors.table
X-squared = 7.1918, df = NA, p-value = 0.3088
```

7. What is the value of the chi-square test statistic as computed by R?
  - **The value of  $\chi^2$  is 7.1918**
8. What is the p-value for this test?
  - **The p-value is 0.3088**
9. Using a 5% significance level, can you conclude that people's personality type affects their choice of major?
  - **Since the p-value greatly exceeds the cut off of 5 percent, we can fail reject the null hypothesis.**
10. Do you believe that your answer (from Question #9) applies to all students at Cal State Fullerton? (HINT: Think about the sample we used and the way we collected the data)
  - **This will most likely not represent the population because we only had a population consisting of natural science majors rather than a broader distribution of majors.**


This is an intentionally short lab to get you familiar with the basics of importing data into R and Rguroo.

Go to <https://www.rguroo.com>, click Register, then click Student and register for an Rguroo account using your student e-mail address(your\_login@csu.fullerton.edu). If you use your student e-mail address you will not have to pay for Rguroo access. Then, go into your student e-mail and complete the registration process.

Once you have registered for an Rguroo account, introduce yourself to six other students in the class. Ask them to tell you:

- Their name
- Their major
- Their height in inches
- Their favorite color
- Approximately how far away from campus they live
- How many units they are taking this semester

Record your responses in an Excel or Google Sheets spreadsheet. Follow the principles of tidy data: put each student in its own row, each variable in its own column, and each value in its own cell.

Save/export the file with a .csv extension. Then, go to the [Data](#) section of Rguroo and select [Data Import](#)  [Data Frame](#). Click on [Browse](#) and find the file on your desktop. Leave all other options as is for now, except maybe check the [Strip Blanks](#) box to make sure nothing weird is going on. If you do not like the default name, type a name in the [Name](#) section. Click on [Upload](#), and upon doing so the file will appear under the [Datasets](#) window on the left, and a table will appear in the middle of the interface, displaying the [Summary](#) of the dataset.

**Question #1** Paste below a screenshot of the Summary table that automatically appears when a dataset is uploaded to Rguroo.



---


## Summary of Data set 'Untitled spreadsheet - Sheet1'



### Numerical Variables

Variable	No. read	No. observed	No. missing	Min	Q1	Q2	Q3	Max	Mean	Std. deviation	Variance	SE of mean
height	8	8	0	63	67	69.5000	71	72	68.7500	3.10530	9.64286	1.09789
commute	8	8	0	0	3.50000	7.50000	22	27	11.6250	10.5144	110.554	3.71742
units	8	8	0	11	12.5000	14	15	18	14	2.13809	4.57143	0.755929

### Categorical Variables

Variable	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7
X	person 1:1	person 2:1	person 3:1	person 4:1	person 5:1	person 6:1	(Other):2
name	abby:1	alex:1	angelica:1	asad:1	bryan:1	cody:1	(Other):2
major	bio:4	cs:3	math:1				
color	black:1	blue:5	purple:1	red:1			

Next, open R Studio on a lab computer (or on your computer if you have it already installed). In the top right pane, under *Environment*, click *Import Dataset*  *From Text (readr)*. Again, browse to find the file, and if you do not like the default name, type a name in the *Name* section. Once done, click *Import* to import the same file into R. Once the file is imported, you should see it in the Environment window.

Click *File*  *New File*  *R Script* (or press *Ctrl + Shift + N*) to open up a new script in the top left. In the script window, type:

```
summary(dataset)
# you should replace dataset with the actual name of the dataset
# in the Environment window
# comments are indicated with the # symbol - you don't actually
# have to type these commented lines
```

Next, put your cursor on that line (or highlight the code) and click *Run* (or press *Ctrl + Enter*). In the Console window in the bottom right, you should now see:

```
> summary(dataset)
```

followed by the output of the command.

**Question #2** Paste below the summary output from R.

```
vim r_program.r
Lab 1 done X
name      name      Age major      height      color      commute
person 1:1 abby      :1      bio :4      Min.       :63.00      black :1      Min.       : 0.00
person 2:1 alex       :1      cs  :3      1st Qu.    :67.50      blue  :5      1st Qu.    : 3.75
person 3:1 angelica :1      math:1      Median     :69.50      purple:1     Median     : 7.50
person 4:1 asad       :1      Mean       :68.75      red    :1      Mean       :11.62
person 5:1 bryan      :1      3rd Qu.    :71.00      3rd Qu.    :21.00
person 6:1 cody        :1      Max.Subm   :72.00      Max.       :27.00
(Other) :2      (Other) :2
units
Min.      :11.00      Not graded
1st Qu.   :12.75
Median    :14.00      Tuesday, August 27, 2019, 2:00 PM
Mean      :14.00
3rd Qu.   :15.00      10 mins 49 secs
Max.      :18.00
modified
Tuesday, August 27, 2019, 1:49 PM
Press ENTER or type command to continue
submissions
Lab 1 Completed by Jared Dyreson.pdf August 27 2019, 1:49 PM
```

**Question #3** What differences (if any) do you see between the default *Rguroo* summary and the default *R* summary?

One is represented via *Rguroo* is similar to the original Excel spreadsheet which is more graphical. The text based version directly reads the CSV file and outputs it to a row column format in text.

Let's consider the probability experiment in which we roll a fair six-sided die.

**Question #1** What is the sample space,  $S$ , for our experiment? That is, list all the possible outcomes.

Our sample space is all of the possible numbers that can be rolled, which range from 1 to 6.

**Question #2** What is the probability associated with each outcome?

There is a 1 in 6 chance to roll any given number, as this is not a physically weighted dice.

**Question #3** What is the probability that you roll an even number?

There is a 1 in 2 chance to roll an even number as there are 2, 4, 6 for the possible outcomes.

**Question #4** What is the probability that you roll a number less than 5?

There is a 4 in 6 chance ( $2/3$ ) for rolling something less than 5 and not including 5.

**Question #5** What is the probability that you roll an even number AND a number less than 5?

There is a 2 in 6 chance. The numbers that are even are 2, 4, 6 and the numbers in that range less than 5 are 2 and 4.

**Question #6** What is the probability that you roll an even number OR a number less than 5?

There is a 5 in 6 chance.

Let  $a$  be an ordered list = [2, 4, 6]

Let  $b$  be an ordered list = [1, 2, 3, 4]

$a + b = [1, 2, 2, 3, 4, 4, 6]$

Make  $a + b$  into set as such: {1, 2, 3, 4, 6}

**Question #7** Are the events "roll an even number" and "roll a number less than 5" disjoint? Why or why not?

These two events can be disjoint because it is possible to roll an even number and it being less than five. Please see above solution for more details.

**Question #8** Are the events "roll an even number" and "roll a number less than 5" independent? Why or why not?

These events cannot be independent because a roll cannot discredit either event as they are agnostic from one another.

Now let  $X$  be the discrete random variable representing the outcome of rolling a fair six-sided die; that is,  $X = 1$  if you roll a 1,  $X = 2$  if you roll a 2, etc.

**Question #9** Fill in the table to describe the Probability Distribution of  $X$ :

$X = x$	$P(X = x)$
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$

**Question #10** Use the table to find  $P(1 < X \leq 4)$ .

Interval notation:  $(1, 4]$

We are interested in values 2, 3, 4 but not 1.

Since the value of the dice directly correspond to its chance of probability, we can simply sum up those probabilities which will be  $3/6$ .

Let  $X$  be the random variable representing the outcome of rolling a fair six-sided die; that is,  $X = 1$  if you roll a 1,  $X = 2$  if you roll a 2, etc.

Recall that in Lab 2, you found the probability mass function of  $X$  to be:

$X = x$	$P(X = x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6



**Question #1** Compute the expected value (mean) of  $X$ .

The expected value would  $1/6(1+2+3+4+5+6) = 3.5$

**Question #2** Compute the variance and standard deviation of  $X$ .

The variance would be  $1/6((1-3.5)^2+(2-3.5)^2+(3-3.5)^2+(4-3.5)^2+(5-3.5)^2+(6-3.5)^2) \approx 2.916666667$

The standard deviation would be  $\sqrt{2.916666667} \approx 1.707825128$

Now we are going to simulate the act of rolling a die. In RStudio, open a new Script ([File](#)  [New File](#)  [R Script](#) or [Ctrl + Shift + N](#)). The code below (in **this font**) will go in your script. It is best to type the lines (except for the introductory **>** or **+**) in the script window and then run them, rather than typing the lines directly into the Console. This will also allow you to save your script as a .R file at the end of the lab.

First, specify your sample space in a vector and create a corresponding vector of the associated probabilities:

```
> S <- c(1, 2, 3, 4, 5, 6)
```

```
> Prob <- c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)
```

Now, let's simulate rolling the die 10 times by randomly sampling a value of the sample space with probability specified in our probability vector:

```
> n_rolls <- 10

> set.seed(338)

> rolls <- sample(x = S, size = n_rolls, prob = Prob, replace = TRUE)

##note that by setting replace equal to TRUE, this replaces the value we
just sampled, thus making each experiment repetition INDEPENDENT.

> num_sixes <- sum(rolls == 6)

> mean(rolls)
```

Once all this code is written, highlight the entire set of code and click [Run](#) (or hit [Ctrl + Enter](#) to run it). You should see the code and the resulting output in the *Console* window.

**Question #3** What is the relative frequency of the number 6 (the proportion of the time you got 6)? Is it close to the probability you computed earlier?

The relative frequency is 2 and that is not close to the number computed earlier.

**Question #4** What is the mean value rolled? Is it close to the mean you computed earlier?

The mean value rolled from the Rscript program came out to be 3.7 which is very close to the calculated 3.5 in question 1.

Set the value of **n\_rolls** to 100. Highlight that line and the lines after it, then click *Run*. This will now create a simulation of 100 die rolls from the same probability model (because you didn't change **S** or **Prob**) and obtain the number of 6's and the mean value rolled. Repeat for **n\_rolls** = 1000, 10000, and 100000.

**Question #5** Convert each value of **num\_sixes** to a relative frequency and fill in table below:

Number of Rolls	Relative Frequency (Proportion of 6's)	Mean
10	2	3.7
100	18	3.65
1,000	169	3.558
10,000	1706	3.5094
100,000	16597	3.49271

**Question #6** Describe what happens to the relative frequency of the occurrence of observing a 6 as the number of rolls increases from 10 to 100,000.

The relative frequency of the occurrence of observing a 6 increases exponentially as it approaches infinity.

**Question #7** Describe what happens to the mean value of the rolls as the number of rolls increases from 10 to 100,000.

The mean value of the rolls as the number of rolls approaches infinity gets lower and lower.

On Titanium, you will see a dataset called *HairEyeColor*. Download the dataset and then import it to RStudio following the same procedure you used in Lab 1.

Once the data is input, open a new script (*Ctrl + Shift + N*) and type the following code in the script window, then highlight it and press *Ctrl + Enter* to run it:

```
summary(apply(HairEyeColor, 2, as.factor))
```

There's a lot going on in this command. If you remember from Lab 1, your categorical variables were imported as character variables (i.e. text), which is very nice in a lot of situations but a bit annoying here. This command converts every column in our dataset to a factor variable (remember, all factors are categorical!), then summarizes each column individually.

As a side note, actually type the commands in the script – don't use Copy/Paste. I do my best to eliminate any weird Word autoformatting, but sometimes it slips through and you might get a weird error message if you paste directly from Word.

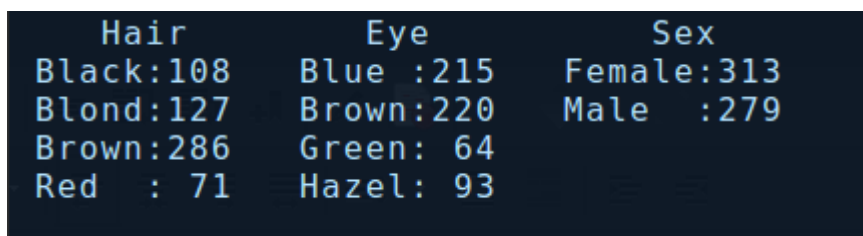
When you run the code (highlight it and then either click *Run* or press *Ctrl + Enter*), you should see it duplicated in the Console window:

```
> summary(apply(HairEyeColor, 2, as.factor))
```

For clarity, example code in the RStudio labs will include the **>** prompt to indicate separate lines of code. If a **+** sign appears at the beginning of a code line, that indicates that the line is continuing the previous command.

Below the code, you should see the output of the command.

**Question #1** Copy or screenshot the Summary output and paste it below.



```
      Hair      Eye      Sex
Black:108  Blue :215  Female:313
Blond:127  Brown:220  Male  :279
Brown:286  Green: 64
Red  : 71    Hazel: 93
```

**Question #2** How many variables are there in this dataset? Are the variables numerical or categorical? Specifically name one of the categorical variables and state its levels.

There are three main variables with sub variables. The variables are categorical as they describe the data.

Sex has two options; Male or Female. These sub variables have a quantity as well.



To find the number of rows in our dataset, we can use either of the following commands:

```
> nrow(HairEyeColor)
```

```
> dim(HairEyeColor)[1] # dim() gives the dimensions of the data frame; the  
first number gives the number of rows and the second gives the number of  
columns
```

**Question #3** How many cases are there in this dataset?

There are 592 cases in this dataset.

Now let's graphically depict the eye colors using a bar graph. There are many different ways to produce plots in R. Most of these methods require the use of *packages*, which contain additional commands that extend the functionality of R (you can think of them as free DLC). To create our plots, we're going to use the ggplot2 package, which is commonly used in "data science." Check the [Packages](#) tab to see if ggplot2 is already installed. If not, click [Install](#) in the tab and type [ggplot2](#) in the prompt. Then type the commands (in the script window, then [Ctrl + Enter](#) to run them so the Console looks like the text below):

```
> library(ggplot2) # loads the ggplot2 package; run this at the beginning  
of the console session, or write it at the top of the script  
> eye_plot <- qplot(HairEyeColor$Eye, geom= "bar") # creates a barplot of  
the values of Eye color  
> ?qplot # brings up the help file for the qplot command, in the Help tab  
in the bottom right
```

Let's stop and examine what's going on in this code. The first command we ran loads the [ggplot2](#) package, which we need to do before we can use any of its functions, and there's a comment to remind us to load it at the top of the script if we use any ggplot2 functions in the script.

Next, the [<-](#) sequence is used to assign a variable name to the output of the code. In this case, we use [<-](#) to store our counts in the variable [eye\\_plot](#). Typing those two characters over and over can get annoying, so we can use the shortcut [Alt + -](#) instead. There are a bunch of different conventions for naming variables in R. The most common and accepted convention is to write a descriptive name for the variable, and if the name has multiple words, connect the words with either an underscore ([\\_](#)) or a period ([.](#)).


Next, the [\\$](#) indicates to select the [Eye](#) column from HairEyeColor. There are a variety of ways to use the [\\$](#) operator, but they mostly correspond to "find the variable with this specific name in my data set or output."

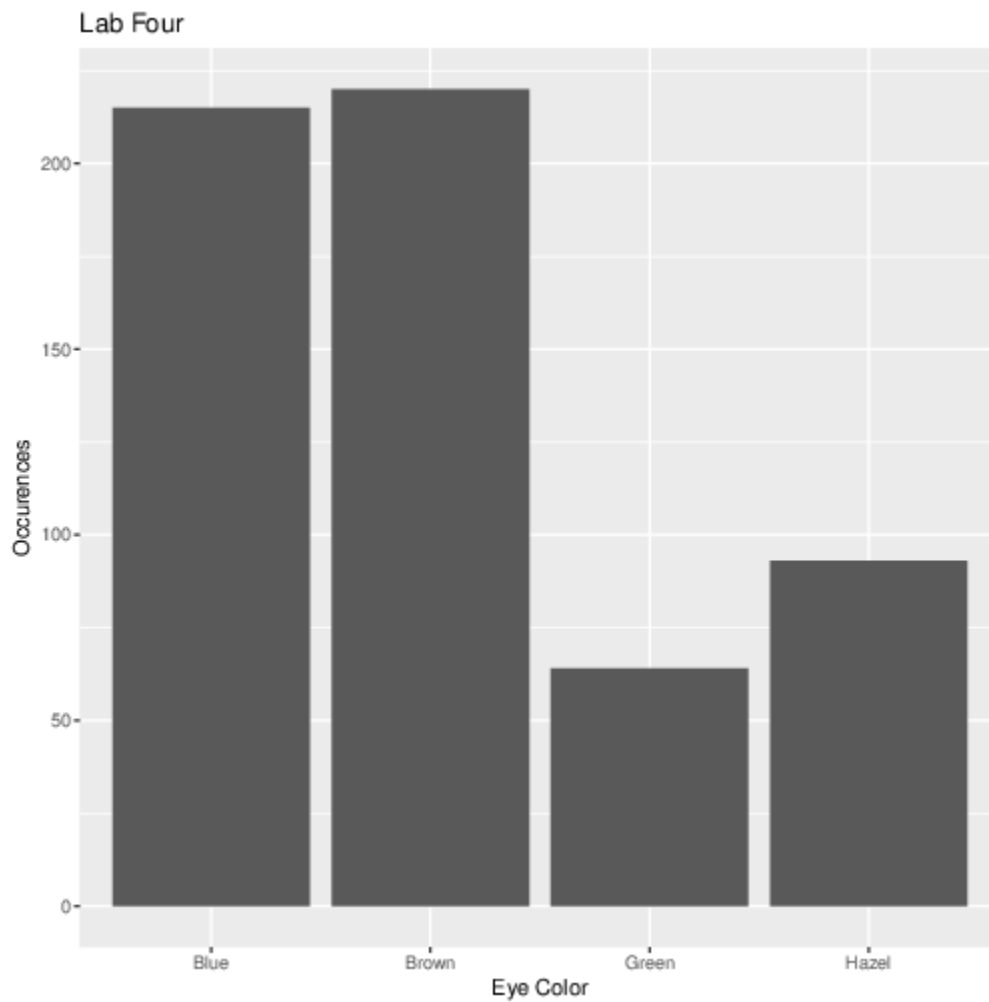
Lastly, you will notice that no plot was actually created. If we didn't store the plot as a variable, we would see a plot in the [Plots](#) window in the bottom right. But because of how the ggplot2 package works, we can make modifications to the plot before we actually see it. We're going to modify the plot by adding a title and axis labels:

```
> eye_plot_labeled <- eye_plot + labs(title= "Plot Title", x = "x-axis  
label", y = "y-axis label") This text is here to make sure you get an error  
message if you're copy/pasting; fix your title and axis labels, then delete  
or comment out this sentence.
```

Once you're done adding and modifying things in your plot, print it to the [Plots](#) tab in the bottom right:

```
> print(eye_plot_labeled)
```

**Question #4** Copy the graph (In the Plots tab, [Export](#)  [Copy to Clipboard](#)) and paste it below.



### Question #5 Which category has the most people? Which has the least?

There are more people who have brown eyes with a 37.20% and green eyes with 10.8%

Now we're going to add some extra stuff to the plot. First, let's get the actual counts. We'll do this by installing the dplyr library and using some functions:

```
> library(dplyr) # the function we need is in the dplyr library, which may  
already be installed. If not, install it using the Packages tab.  
> eye_counts <- count(HairEyeColor, Eye)
```

Now let's get the actual percentages, rounded to the nearest tenth of a percent:

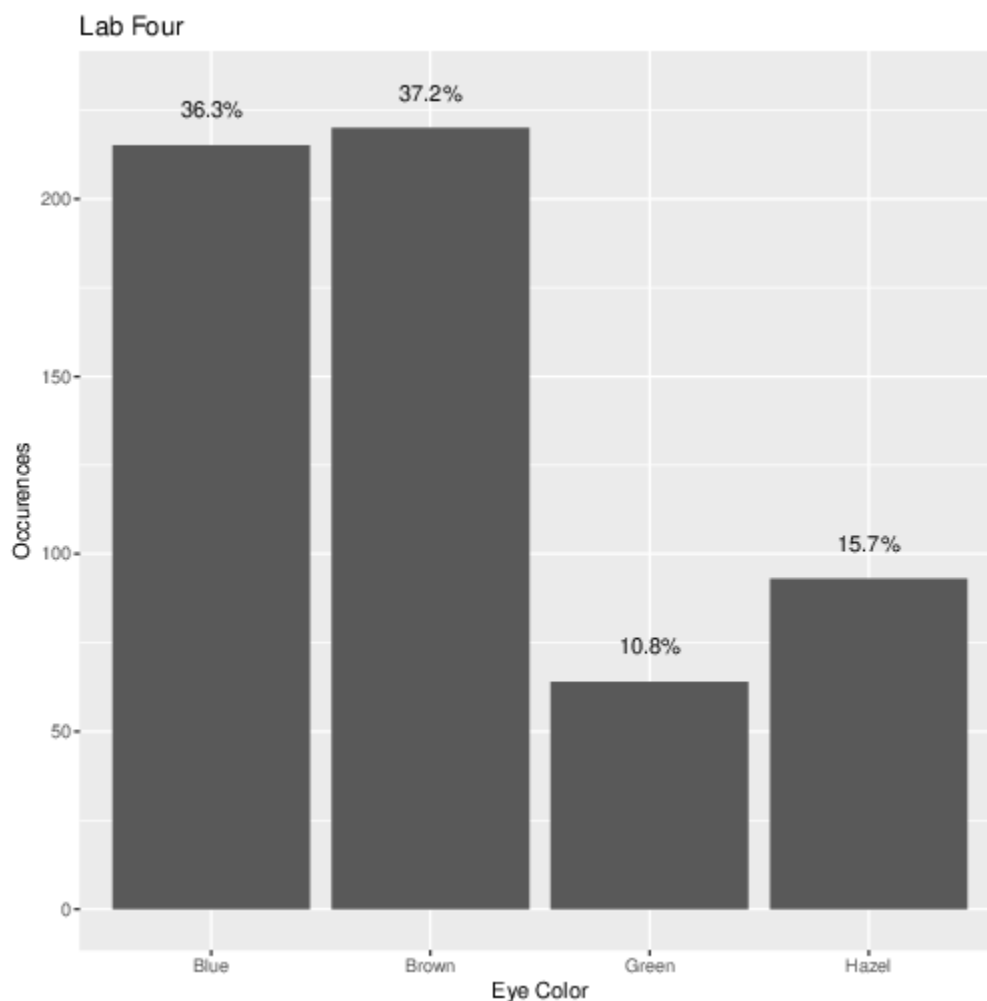
```
> eye_pcts <- eye_counts$n/sum(eye_counts$n) * 100  
> eye_pcts <- round(eye_pcts, 1)
```

And add the text percentages to the plot:

```
> eye_plot_pcts <- eye_plot_labeled + annotate("text", x = seq(1,4), y =  
eye_counts$n+10, label = paste(eye_pcts, "%", sep = ""))
```

Once again, there's a whole lot going on in these commands. Try to figure out as best you can what's going on by looking at the help files for the [annotate](#), [seq](#), and [paste](#) commands (for instance, [?annotate](#)).

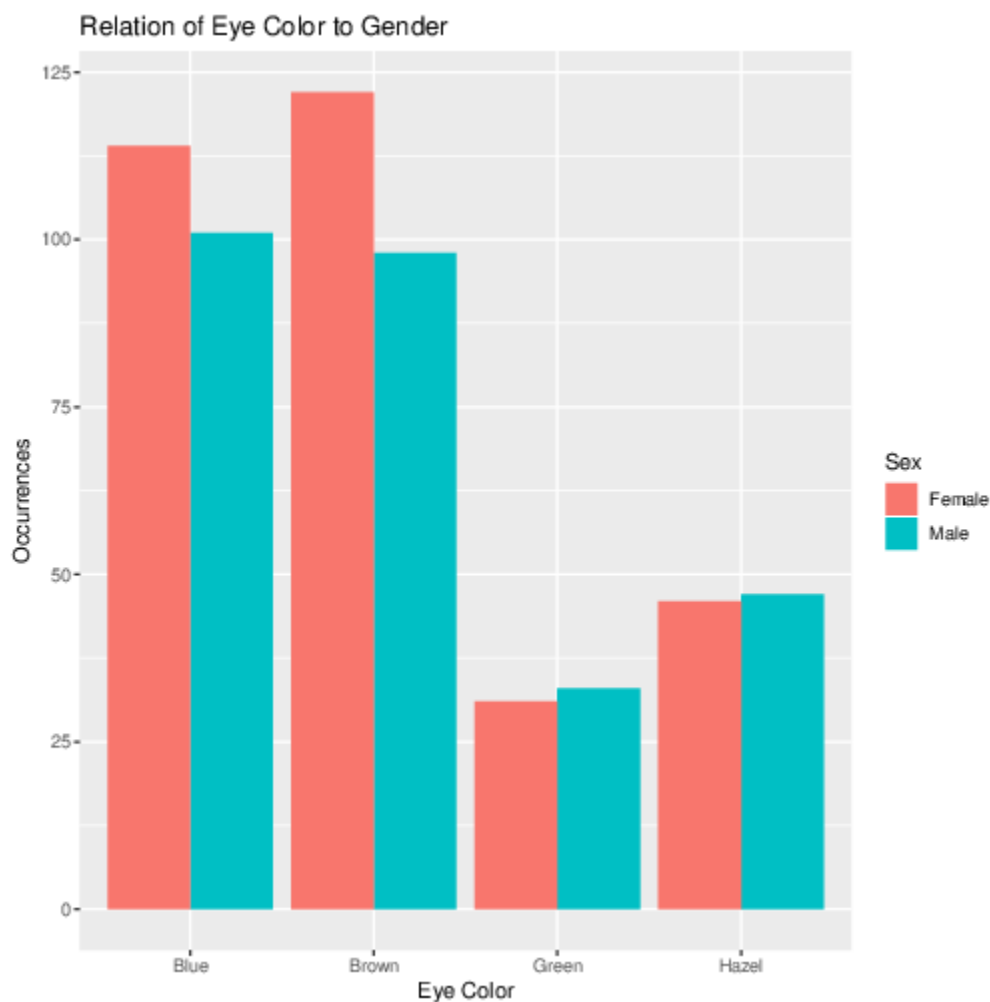
### Question #6 Print the new graph to the [Plots](#) tab, then copy it and paste it below.



Now let's start looking at the other plotting command in the ggplot2 library: **ggplot**. This command is a lot more complicated but also a lot more versatile. We'll use the example of looking at the relationship between eye color and gender, so we want to visually compare the distribution of Eye Color of males and females. Type the code below, but don't forget to change the plot and axis titles to something useful before running it!

```
> eye_gender_plot <- ggplot(data = HairEyeColor, mapping = aes(x = Eye))  
> eye_gender_barplot <- eye_gender_plot + geom_bar(aes(fill = Sex),  
position = "dodge") + labs(title = "Plot Title", x = "x-axis label", y =  
"y-axis label") AGAIN, IF YOU ARE COPY/PASTING, FIX THE LABELS AND THEN  
DELETE OR COMMENT THIS SENTENCE!  
> print(eye_gender_barplot)
```

**Question #7** Copy the new graph and paste it below.



**Question #8** Which color is most prevalent for females; which color for males?

Most prevalent for females is brown eyes and for males it is blue.

Suppose that we are randomly guessing on a 30-question multiple-choice exam. We're guessing a little better than chance alone would predict – we estimate we have a 35% chance of getting each question correct.

**Question #1** Let  $X$  be the number of questions we get correct. Evaluate the BINS assumptions for this scenario to explain why  $X$  can be modeled as a binomial random variable. As part of your evaluation, identify the parameters  $n$  and  $p$ .

**B:** we either get the question right or wrong

**I:** we assume that previous questions do not hint at answers in future problems

**N:** 30 questions which will be our  $N$  variable

**S:** We assume that we will answer each of the questions on the test.  $P=0.35$

Open a new R script and enter the values of  $n$  and  $p$ .

```
> n <- # number of trials
```

```
> p <- # probability of success on one trial
```

To find the exact probability of  $x$  successes in  $n$  trials, run the command:

```
> dbinom(x, size = n, p = p) # actually put in value of x; P(X = x) given  
the parameters n and p specified above
```

**Question #2** What is the probability of getting exactly 15 questions correct?

The probability will be 0.03510604 to get 15 questions correct

For **Questions #3-6**, use the **pbinom** and **dbinom** commands as needed.

```
> pbinom(x, size = n, p = p) # prob. of at most x successes in n trials,  
P(X <= x)
```

```
> pbinom(x, size = n, p = p, lower.tail = FALSE) # prob. of more than x  
successes in n trials, P(X > x)
```

Note that you may have to add and subtract probabilities as produced by **pbinom** and **dbinom**. It may be useful to first define the probability statement mathematically to figure out what to add/subtract. For

example, `pbinom(15, n, p) - dbinom(15, n, p)` gives  $P(X \leq 15) - P(X = 15) = P(X < 15)$ , the probability of strictly fewer than 15 successes in  $n$  trials.

**Question #3** What is the probability of getting at most 10 questions correct?

The probability would be 0.5077582

**Question #4** What is the probability of getting 18 or more questions correct?

The probability would be 0.001447268

**Question #5** What is the probability of getting more than 10 questions correct, but at most 18?

The probability would be 0.5063109

**Question #6** What is the probability of getting strictly fewer than 5 questions correct?

The probability would be 0.007517715

We can also think of this problem in terms of sample proportions. Remember that the sample proportions do not have a binomial distribution, but we can use the formula  $\hat{p} = X/n$  to convert between sample proportions  $\hat{p}$  and sample counts of success  $X$ .

**Question #7** What is the probability of getting exactly 60% on this 30-question test?

The probability would be 0.003056097

**Question #8** What is the probability of getting a score of 45% or lower?

The probability would be 0.8736881

**Question #9** What is the probability of getting a score lower than 50%?

The probability would be 0.9348103

**Question #10** Suppose that we pass the test (70% or above). Is this consistent with our assumption that we are randomly guessing with a 35% chance on each question? Why or why not?

Although it is theoretically possible to achieve a score of 70% or greater, the likelihood of that occurring by guessing alone is very low. The more questions we try to answer correctly, our overall probability drops exponentially.

**Clinical trials** are studies that provide intervention to human subjects and attempt to determine whether the intervention is safe and effective. Investigate the clinical trial, “A Behavioral Intervention to Improve Hypertension Control in Veterans,” which can be found online at

<https://clinicaltrials.gov/ct2/show/study/NCT00286754>

**Question #1** For this study, what is the real-world question of interest?

Which method of intervention will help the best to reduce blood pressure.

**Question #2** What is the population of interest in this study? What was the sample being studied (what are the experimental units, and how many are there)?

The population of interest is 533 veterans and the sample being studied are the three groups. These three groups were equally randomized and categorized based on the intervention method use for the respective group. The group names included SMI, HEI and UC.

**Question #3** What was the factor (experimental/explanatory) variable in this study? What were the primary outcome (response) variables? Classify each variable as categorical (qualitative) or numerical (quantitative).

The experimental variable was the type of treatment and the explanatory variable was the group of randomly selected veterans. Primary outcome variables included blood pressure, adherence (to diet, exercise and medications), quality of life, acceptability, cost/cost effectiveness.

Categorical: adherence, quality of life, cost effectiveness

Quantitative: blood pressure, cost, acceptability

**Question #4** What was the control group in this study? What was/were the treatment group(s)?

The control group on this study is the UC (usual care) group as there is no new change in the veteran's way of life. The treatment groups were SMI and HEI because they did introduce something new.

**Question #5** Did the experimenters avoid confounding due to the placebo effect? If so, how? If not, why not?

The experimenters managed to avoid confounding variables because they randomly selected their population and there is blinding on both the veteran and their respective social worker. These measures will mitigate bias and inaccurate data.

**Question #6** The experimenters included a number of “exclusion criteria.” People meeting one of these criteria would not be included in the study. For one of those criteria, explain why including people meeting that criterion might bias the results.

One of the criterion that would be introduce bias would be “Unable to follow the study protocol”. This is hard to ensure because people can/will lie if they know they’re in a study. That is called the Hawthorne Effect.

“A/B testing” is fancy name for randomized controlled experiments that manipulate only a single factor. Typically, most A/B testing is now done via the Internet, including websites, apps, e-mail marketing, etc. Open the paper “Controlled Experiments on the Web: Survey and Practical Guide,” which can be found at <http://www.exp-platform.com/Documents/controlledExperimentDMKD.pdf> and read sections 2 and 3.

**Question #7** Why is an A/B test called an A/B test?

There are two different versions of an end product but one will be well received and the other will suffer.

**Question #8** What is an A/A test and why would researchers use it?

You split your sample group in two but expose them to the same experience. Researchers would use it to test the experimentation system and assess its variability for power calculations.

**Question #9** For one of the examples discussed in section 2 (pages 143-148 of the document), explain what the treatment and control groups are.

For the Doctor FootCare UI example, people who had the coupon text input field had more of an adverse reaction to completing their purchase. This was because they felt there might be a better deal out there and they have not found it yet. When removed in later iterations, the sales went up by 6.5%.

**Question #10** For the example you chose in **Question #9**, what was the Overall Evaluation Criterion (OEC)? Is the Overall Evaluation Criterion based on an experimental (explanatory) variable or an outcome (response)?

The OEC in this example was the initial decrease in sales when the coupon text field was present and the overall increase when it was subsequently removed.



The ELISA test was an early test used to screen blood donations for antibodies to HIV. A study (Weiss et. al. 1985) found that the conditional probability that a person would test positive given that they had HIV was 0.97, and the conditional probability that a person would test negative given that they did NOT have HIV was 0.926. The World Almanac gives an estimate of the probability of a person in the USA of having HIV to be 0.0026.

**Question #1** Given the numbers in the paragraph above, identify the base rate (prevalence), sensitivity, and specificity of the test.

Prevalence: 0.0026

Sensitivity: 0.97

Specificity: 0.926

**Question #2** Suppose 10000 random people are tested. How many of them do you expect to actually have HIV? How many do you expect not to have HIV?

26 people would be expected to have HIV and 9974 not to have HIV

**Question #3** Of those with HIV, how many do you expect to test positive?

Of the 26, 25 would be expected to have HIV

**Question #4** Of those without HIV, how many do you expect to test negative?

9235.924 out of the 9974 would be expected to test negative

**Question #5** Draw a two-way table to represent this situation. Fill in your answers to **Questions #2-4** in the appropriate cells, then fill in the rest of the table. **Do not solve for a probability yet.**

Yessir.

**Question #6** Draw a tree diagram to represent this situation. Fill in the probabilities on each branch of the tree. **Do not solve for a probability yet.**

Yessir.

Recall that Bayes' Theorem says:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

**Question #7** In the situation of testing for HIV, what should Event A be? What should Event B be? How do you know?

Event "A" is having HIV and event "B" is testing for having HIV.

**Question #8** Rewrite Bayes's Theorem for this situation, plugging in the correct numbers for each probability. **Do not solve for a probability yet.**

The amount of people in the testing who have tested positive over the total amount of people you are testing.

**Question #9** Suppose a random person is tested and they test positive. Using the method (two-way table, tree diagram, Bayes's Theorem) that makes the most sense to you, find the conditional probability that this person has HIV given that they test positive.

**Question #10** Based on your results, would you recommend that this be the first and only test used to detect HIV? Why or why not?

This should not be the only test you take to confirm the results given by the aforementioned test. There is significant margin of error in testing false positive.

[Tan et al. \(2018\)](#) were interested in how pet dogs bond with complete strangers. In one of their studies, they placed a food treat under one of two cups (the dog did not know which bowl had the food) and had a stranger point the dog toward the cup with the food. Although the researchers did many trials to see how trust evolved, one of the results they analyzed was whether dogs would pick up the cue from the stranger on the very first trial. A total of 53 dogs were tested.

**Question #1** Assuming that the tested dogs were independent (i.e., the dogs did not communicate anything to each other), check the other three assumptions of a binomial setting.

**B:** If the dog picks the cup that has food in

**I:** This is already assumed

**N:** 53

**S:** 0.5

**Question #2** What is the sample size in this study?

**Our sample size would be 53**

**Question #3** What is the parameter  $p$  about which we would like to make inference?

**The proportion of dogs that could pick up a cup with food and without food.**

According to the researchers, if dogs do not trust strangers, they would be just as likely to pick the “correct” bowl as the “incorrect” bowl. Let’s make this our “nothing unexpected is happening” condition.

**Question #4** What is the value of the parameter  $p$  under the null hypothesis?

**The value of parameter  $p$  is 0.5 because they could either trust the stranger or not trust the stranger.**

**Question #5** What type of “test statistic” can we compute from sample data in the binomial setting? What type of sampling distribution does it have, and what are the parameters of that distribution under our null hypothesis?

**Using the binomial setting we can calculate the number of successes.**

**The test statistic is the number of dogs who pick up the food.**

**Under the Null Hypothesis, we have a binomial distribution**

Now we will use R to set up the distribution of our test statistic under the null hypothesis. Open up a new script and assign the following variables:

```
> n <- # the sample size from question 3  
> p0 <- # the value of p if the null hypothesis is correct
```

Suppose that under our “something unexpected is happening” condition, where dogs *do* trust strangers, a “practically significant” value of  $p$  is assumed to be 0.7.

Assign the appropriate variable in the R script:

```
> p1 <- 0.7 # the value of p if the alternative hypothesis is correct
```

Since the value under  $H_1$  is higher than the value under  $H_0$ , we will look in the upper tail of the distribution for our critical region.

**Question #6** According to convention, what is our desired maximum probability of committing a Type I Error?

**The desired probability is 0.01 of committing a Type 1 Error because we want a small value**

Again, assign the appropriate variable in the R script:

```
> alpha <- # the conventional maximum value of alpha
```

Finally, we use our binomial probability calculator:

```
> qbinom(alpha, size = n, prob = p0, lower.tail = FALSE) # FALSE because we are looking in the upper tail of the distribution
```

**Question #7** If you did everything right, you will get an output of **32**. Does this correspond to a critical region of  $X \geq 32$  or a critical region of  $X > 32$  (or, equivalently,  $X \geq 33$ )? Why? (Hint: review your **pbinom** commands from Lab 5).

**The critical region of  $X > 32$  because the tail does not include the value of 32.**

Now we will assign the output to a variable:

```
> crit.value <- # either 32 or 33 depending on your answer to Question 7
```

Finally, to find the power, compute:

```
> power <- pbinom(crit.value, size = n, prob = p1, lower.tail = FALSE) + dbinom(crit.value, size = n, prob = p1)
```

**Question #8** According to R, what is the power of our test at our sample size and specific alternative hypothesis?

**According to R, the value of power is going to be 0.950508**

**Question #9** What is the probability of committing a Type II Error, given our sample size, alternative hypothesis, and  $\alpha$  value?

**Our Beta value is going to be 0.049492**

[Save](#) your script – it will make Lab 9 much easier!

**Question #10** Based on the work you have done in this lab, do you believe that the researchers have a high enough power to detect a practically significant effect of “dog trust in strangers”?

The researchers do have high enough power because it is well above the 80% threshold.

# Lab Assignment 9

Jared Dyreson

TR @ 11:30 - 14:15

MATH-338, Dr. Wynne

1. What is the power under the new alternative  $H_1: p = 0.6$ ?
  - **The power would be 0.425878**
2. What is the power under the new alternative  $H_1: p = 0.8$ ?
  - **The power would be 0.9991866**
3. How does the power change as the alternative value of  $p$  gets further from the null value of 0.5?
  - **The power increases as it deviates upward from the null value of 0.5**
4. What is the critical region for  $\alpha = 0.01$ ? Is it a larger or smaller critical region compared to  $\alpha = 0.05$ ?
  - **The new critical region is going to be 36 which is higher than the original value of 32 which was initially found with an alpha of 0.05**
5. Change `crit.value` to the endpoint of the new critical region (from Question #4). Then, run that line and the lines below it to compute power. What is the new power?
  - **The new power is going to be 0.5788821 when you change the critical value**
6. Repeat the steps using  $\alpha = 0.10$ . What is the power of the test at this new  $\alpha$  value?
  - **The new power is going to be 0.09055912 when the value of alpha changes to 0.10**
7. How does the power change as the probability of Type I Error increases? Why do you suspect it changes in that direction?
  - **When the  $\alpha$  changes, the critical value range will subsequently increase as well. This can be seen in the chart below**
  - $\alpha = 0.01 \rightarrow x \geq 36$
  - $\alpha = 0.05 \rightarrow x \geq 33$
  - $\alpha = 0.1 \rightarrow x \geq 32$
  - **This change can be attributed to the Type I Error because choosing lower values of  $\alpha$  make it harder to reject a null hypothesis.**
  - **The act of rejecting a true null hypothesis is considered a Type I Error**
8. What is the critical region for  $n = 30$ ?
  - **The critical region is 25**
9. Change `crit.value` to the endpoint of the new critical region (from Question #8). Then, run that line and the lines below it to compute power. What is the new power?
  - **The new power is 0**
10. Repeat the steps using  $n = 100$ . What is the power of the test using this new sample size?
  - **The new power for the new sample size is going to be 1**
11. How does the power change as sample size increases? Why do you suspect it changes in that direction? (Hint: think about the critical regions in terms of sample proportions!)
  - **As the sample size increases, the power increases as well. This makes sense because as you get more individuals for a study, the data becomes that more accurate. In terms of a critical region, there is a broader range of values you have access to.**

## External Links

- [Consequences of errors and significance \(STATS\)](#)