BIOS 4120 Lab 3

Feburary 4 - 5, 2020

Objectives

In today's lab we will:

1. Use RStudio to create tables and graphics from a dataset
2. Discuss the relationship between hypothesis testing and confidence intervals

Common Plot Types and Their Functions

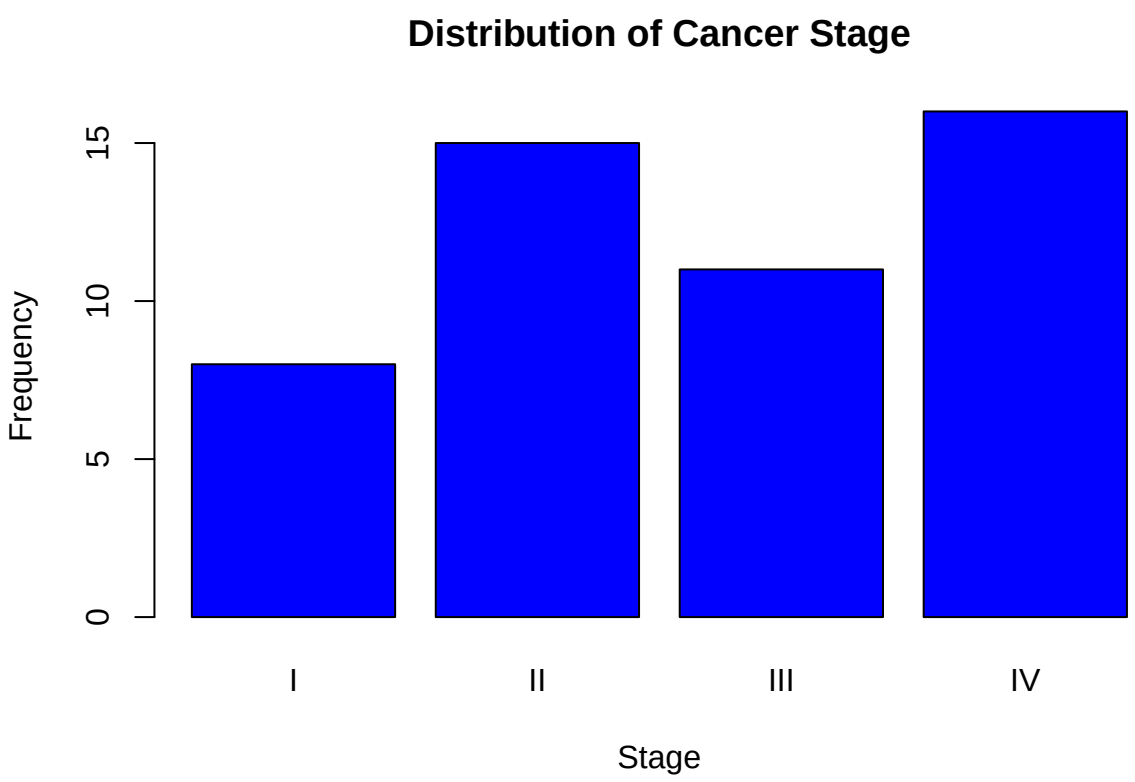
Creating a Bar Chart

For a bar chart, each column represents a group defined by a categorical variable. If you have data that is catagorical (like we see in the 'titanic' dataset), you will want to use a bar chart to display information.

Bar Plots – 'barplot()'

A simple way to create barplots is to use the barplot function.

```
counts <- c(8, 15, 11, 16)
barplot(height = counts,
        names.arg = c("I", "II", "III", "IV"),
        col = "blue",
        main = "Distribution of Cancer Stage",
        xlab = "Stage",
        ylab = "Frequency")
```



Determining Parameters for Plotting Functions

The easiest way to figure out what parameters are associated with these plotting functions is to use the "?" command and read the description of the function.

Note: There is no real limit on how many parameters you can put in one of these functions, but it is easiest to read if you separate them onto different lines.

Common Graphical Parameters for Plot Functions

- xlab, ylab – the label on the x-axis and y-axis, respectively
- xlim, ylim – a vector representing the x and y domains, respectively
- pch – an integer representing the type of plot points. You can also create your own plot points with quotes. e.g. "|"
  - See http://www.endmemo.com/program/R/pchsymbols.php for more plot point options
- lty – an integer representing the type of line
- main – sets a title for the plot
- col – sets the color for points, lines, or graphics for your plot

See http://www.statmethods.net/advgraphs/parameters.html for point, color, and line options and there corresponding codings

Note: **Not all plots have the same arguments.** The help function is a great way to see what arguments are part of a function

Adding Lines or Text to Plots

Once you run your plotting command and the figure pops up, you can still add either lines or text to the plot using these functions:

```
abline(a, b, h, v)

text(x, y, "Text")
```

Creating Legends

The legend function in R can be used to add a legend to a plot. There are a variety of arguments that this legend function can take. Some of the most commonly used are described below.

- you can set x equal to a specified keyword ("bottomright", "bottom", "bottomleft", "left", "topleft", "top", "topright", "right") to indicate placement of the legend or you can use the x and y arguments to specify x and y coordinates for legend position.
- the legend argument takes a character expression to be included in the legend
- lwd takes an integer to indicate the line width
- the col argument dictates the color of the points or lines that appear in the legend

```
legend(x = "bottomright",
       legend = example,
       lty = 1,
       lwd = 5,
       col = "blue")
```

Storing Plots

Usually you can right-click on a plot in RStudio or R to copy them and then easily paste them onto a Google document. If this is giving you trouble, you can always save a plot to a pdf using the form below (after you have set a working directory):

```
pdf("Name_of_Plot.pdf", height = 3.5, width = 5.25)

## PLOT CODE HERE

dev.off()
```

This will save a pdf of the plot you set up in the working directory you had set up at the beginning.

Using ggplot

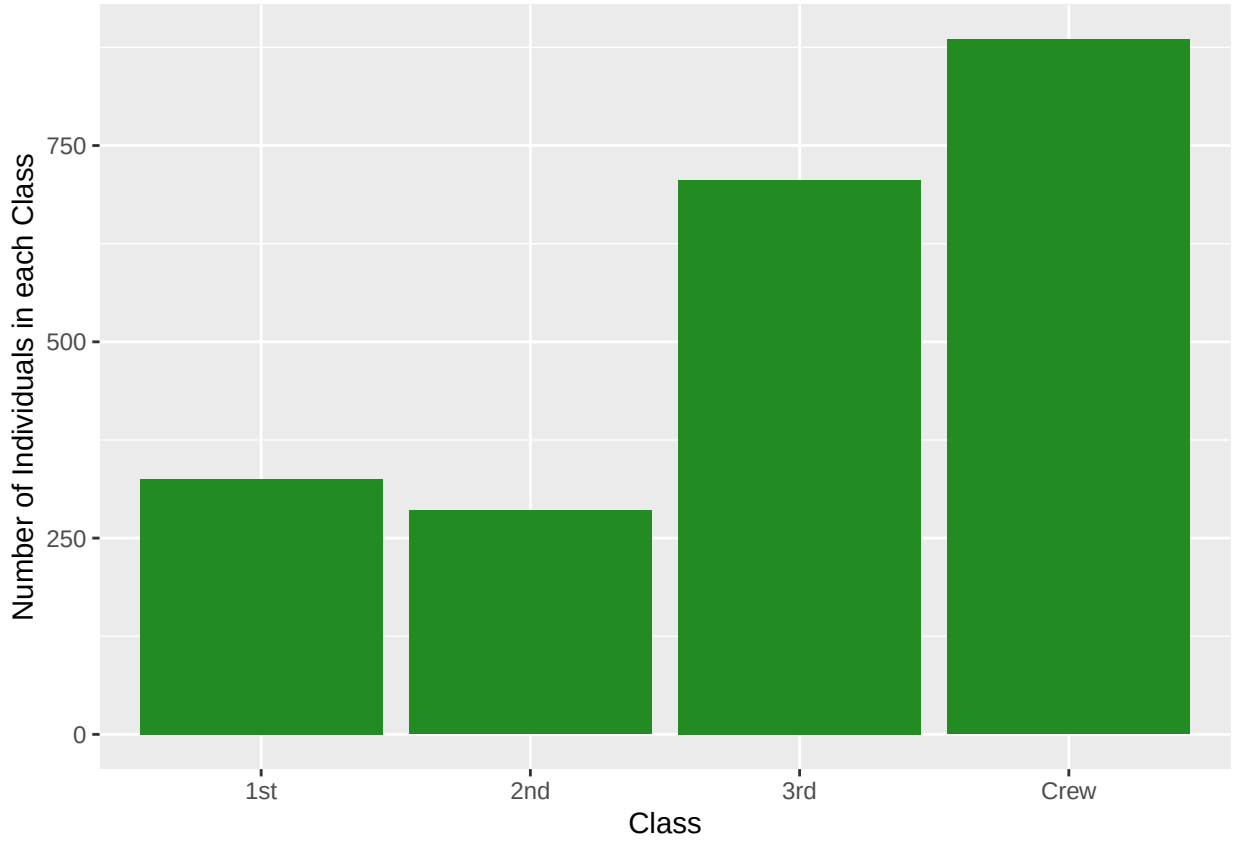
When can also use ggplot to create barplots. Using ggplot is currently a popular way to create graphics. To use the ggplot function you will need to install and load the ggplot2 package which can be done with the following code.

```
#install.packages(ggplot2)
library(ggplot2)
```

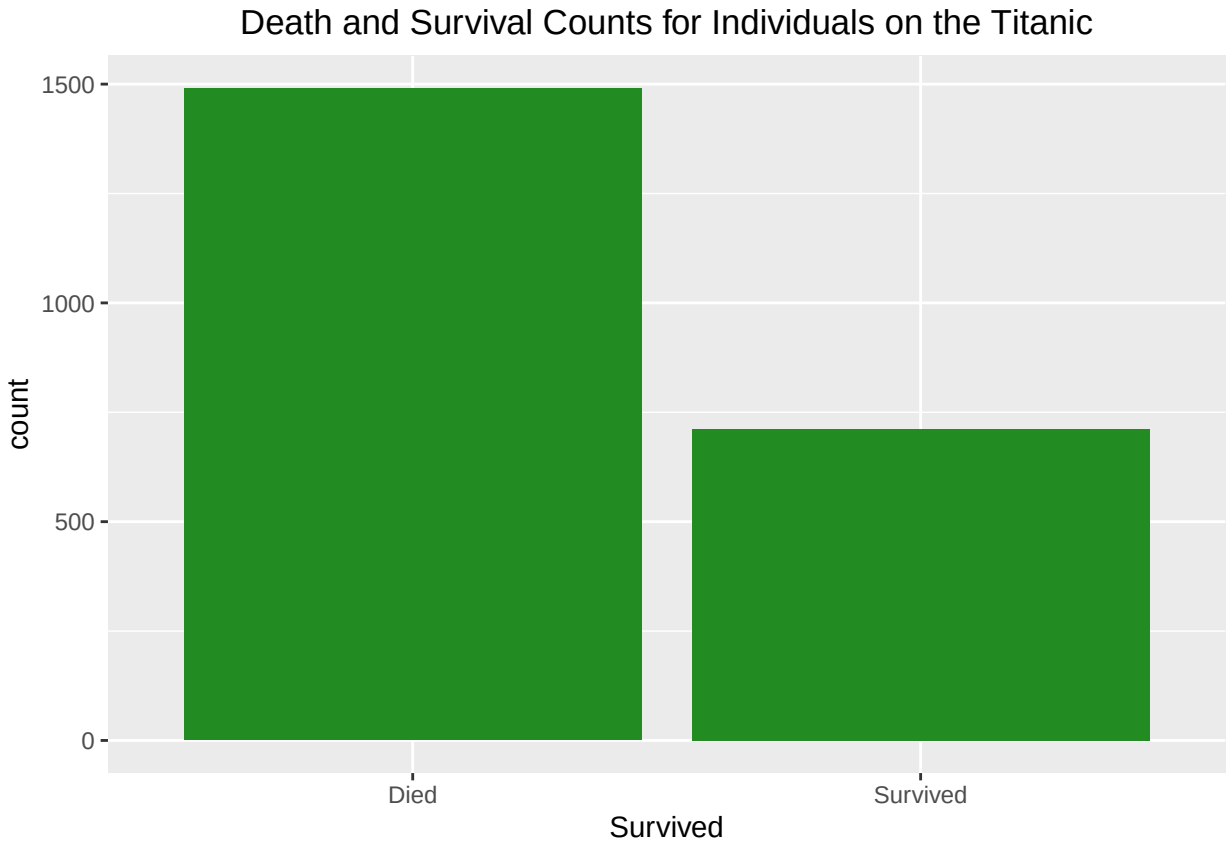
```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

Barchart using ggplot

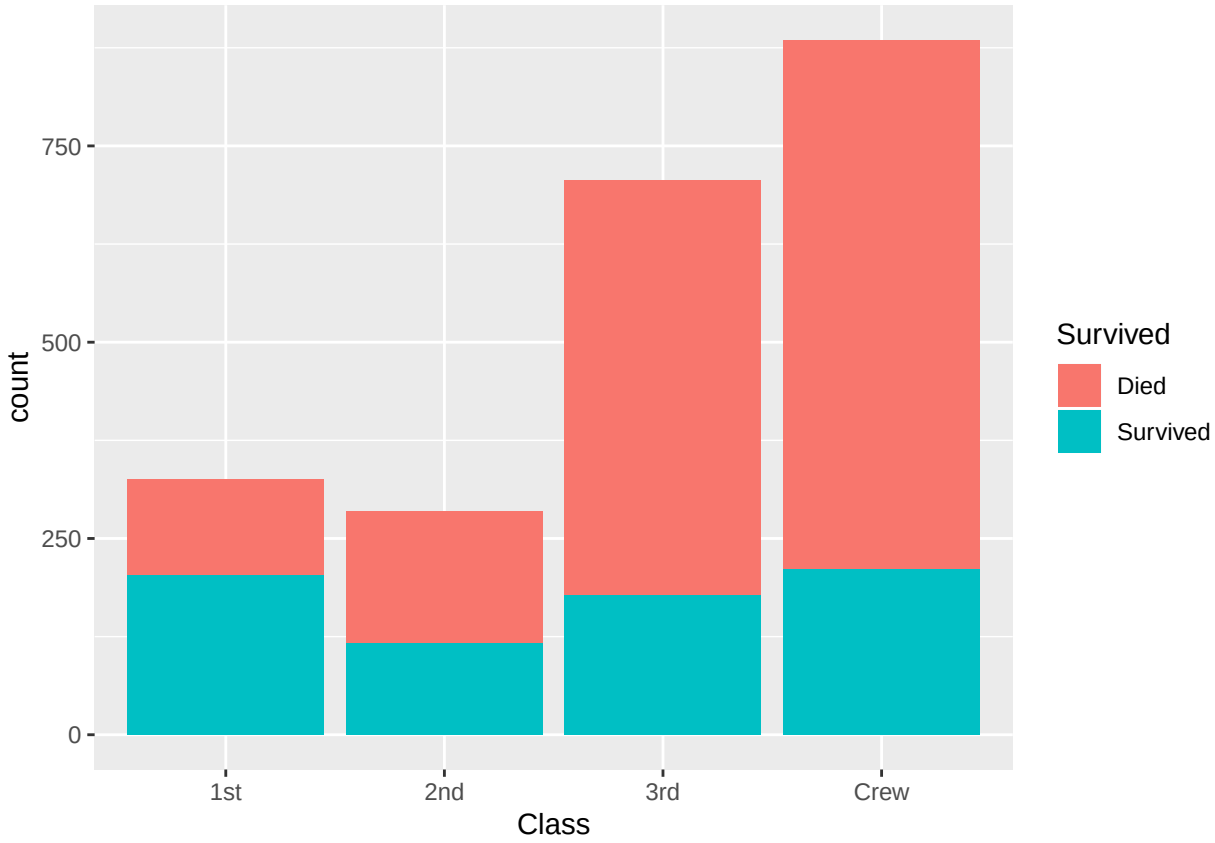
```
titanic <- read.delim("http://myweb.uiowa.edu/pbreheny/data/titanic.txt")

ggplot(titanic, aes(x = Class)) +
  geom_bar(position = "stack", fill="forest green") +
  ylab("Number of Individuals in each Class") +
  xlab("Class")
```



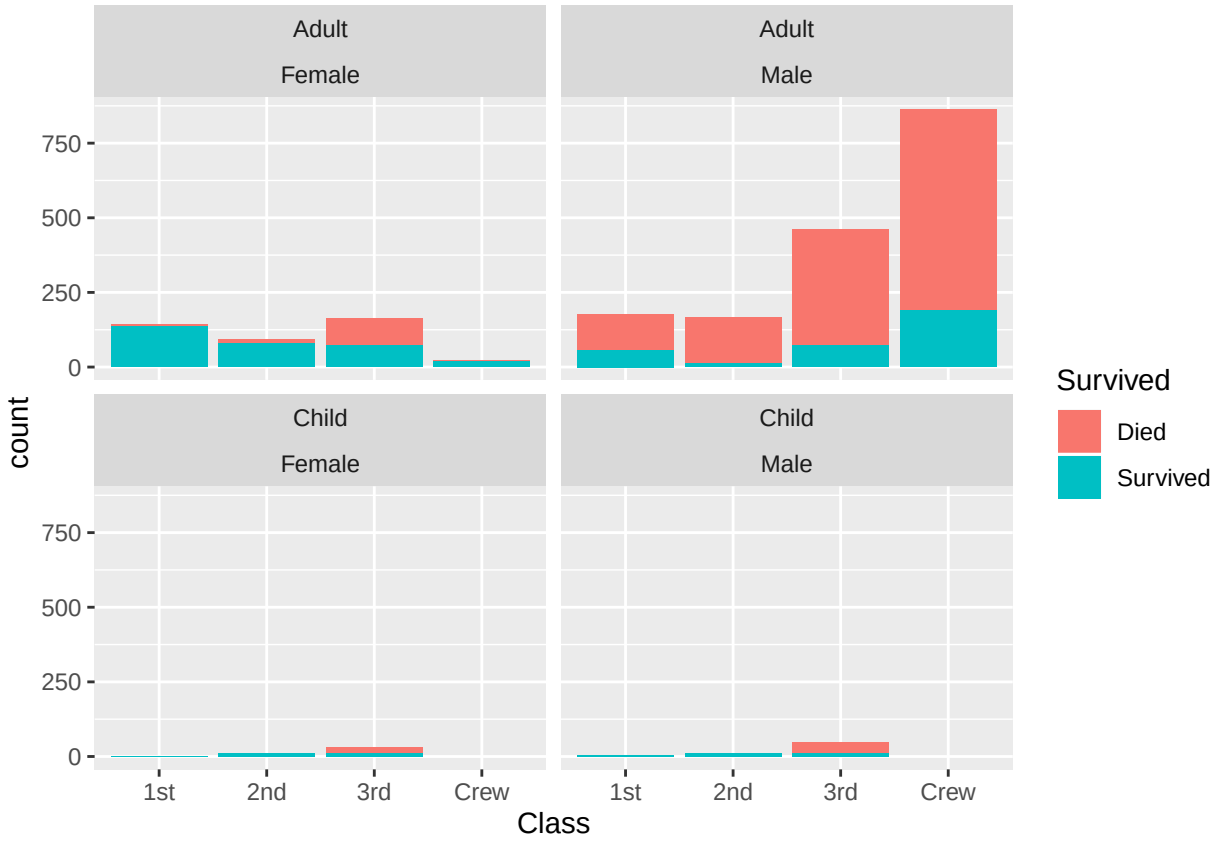
```r
ggplot(titanic, aes(x = Survived)) +
  geom_bar(position = "stack", fill="forest green") +
  ggtitle("Death and Survival Counts for Individuals on the Titanic") +
  theme(plot.title = element_text(hjust = 0.5))
```



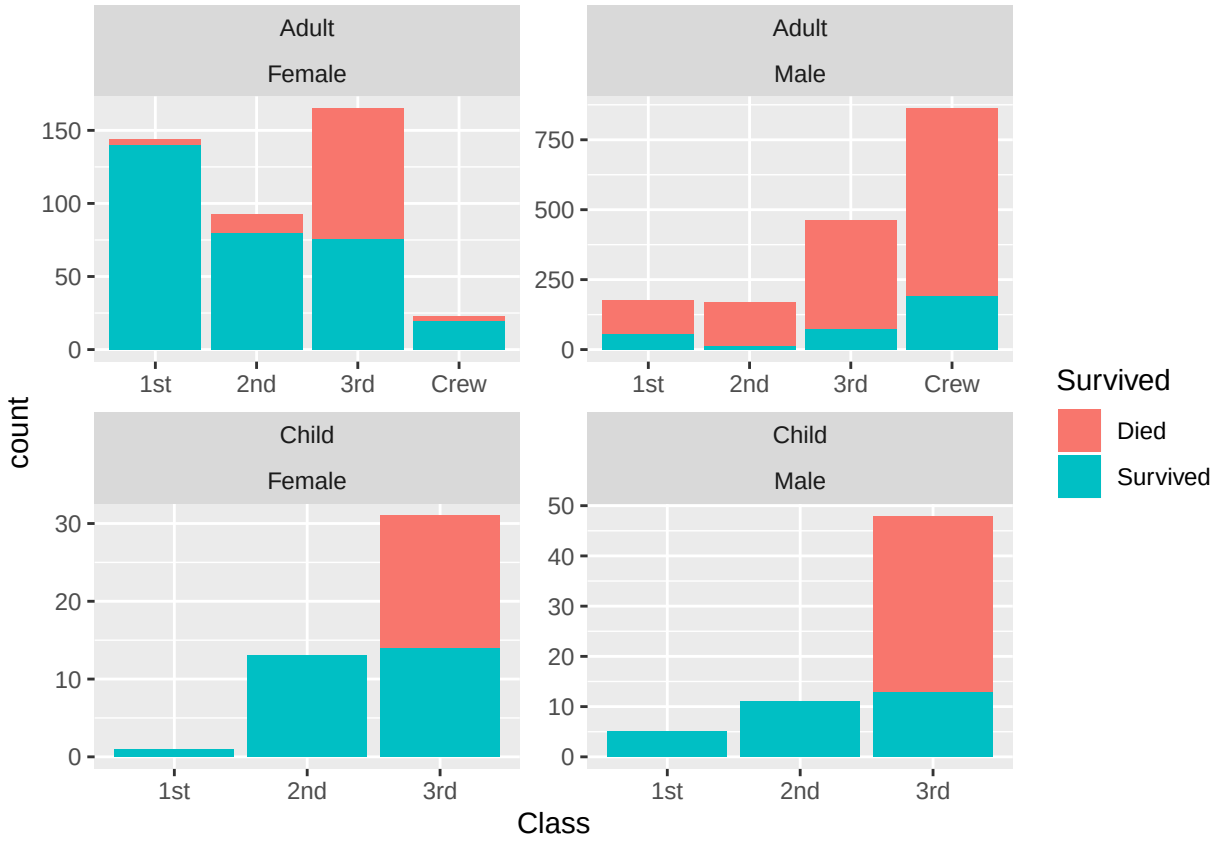
```r
ggplot(titanic, aes(x = Class)) +
  geom_bar(position = "stack", aes(fill = Survived))+
  xlab("Class")
```



```
ggplot(titanic, aes(x = Class)) +
  geom_bar(position = "stack", aes(fill = Survived)) +
  facet_wrap(c("Age", "Sex")) +
  xlab("Class")
```




```
ggplot(titanic, aes(x = Class)) +
  geom_bar(position = "stack", aes(fill = Survived)) +
  facet_wrap(c("Age", "Sex"), scales = "free") +
  xlab("Class")
```


Adult
Female
Adult
Male
Child
Female
Child
Male
count
Class
1st
2nd
3rd
Crew
Survived
Died
Survived

Group Bar Plot Exercise - we will use the barplot() to work the following problem

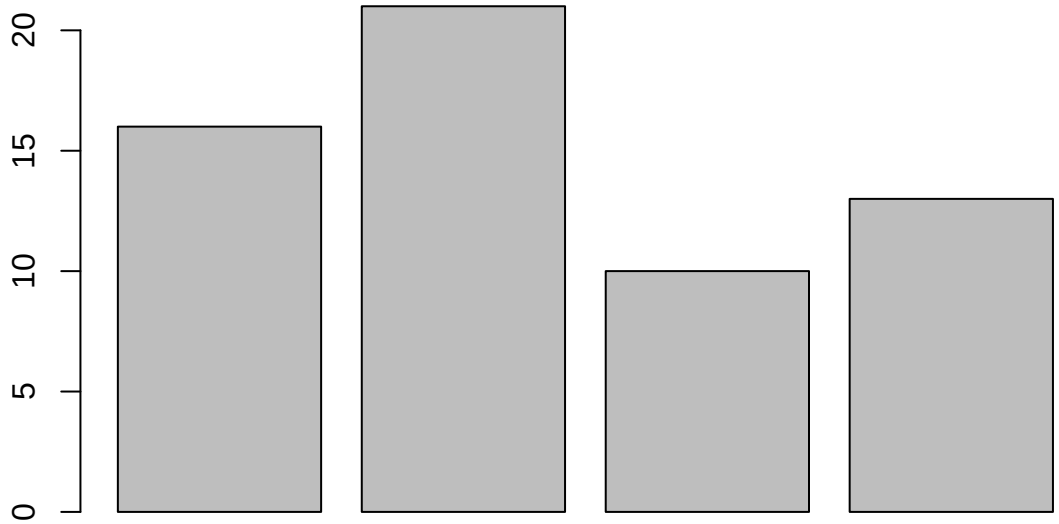
We are interested in the amount of people who eat at Iowa City restaurants downtown. On the night of this study, there were 16 people at Shorts, 21 at Donnelly's Pub, 10 at Blue Moose, and 13 at Joe's Place.

Step One

Create a vector named "people" of the counts.

Step Two

Create a basic bar plot of "people".



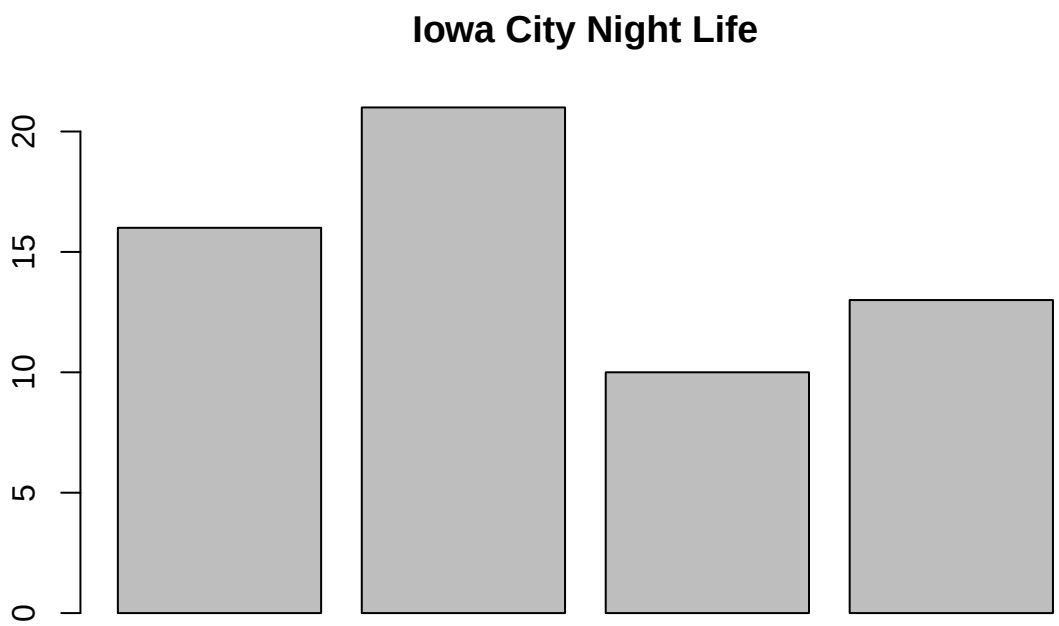
What improvements could be made?

- Main title
- Label the bars
- Label the x axis
- Color
- Make the y axis longer than the tallest bar

Step Three

Give the bar plot a main title.

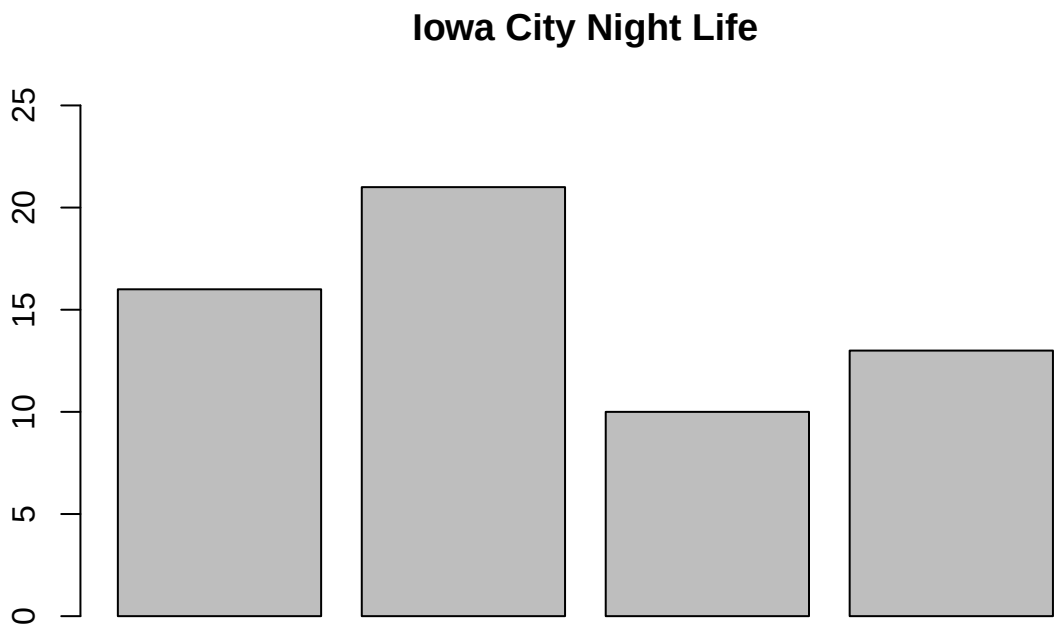
Hint: use the "main" parameter.



Step Four

Adjust the y axis to be high enough.

Hint: Use the "ylim" parameter

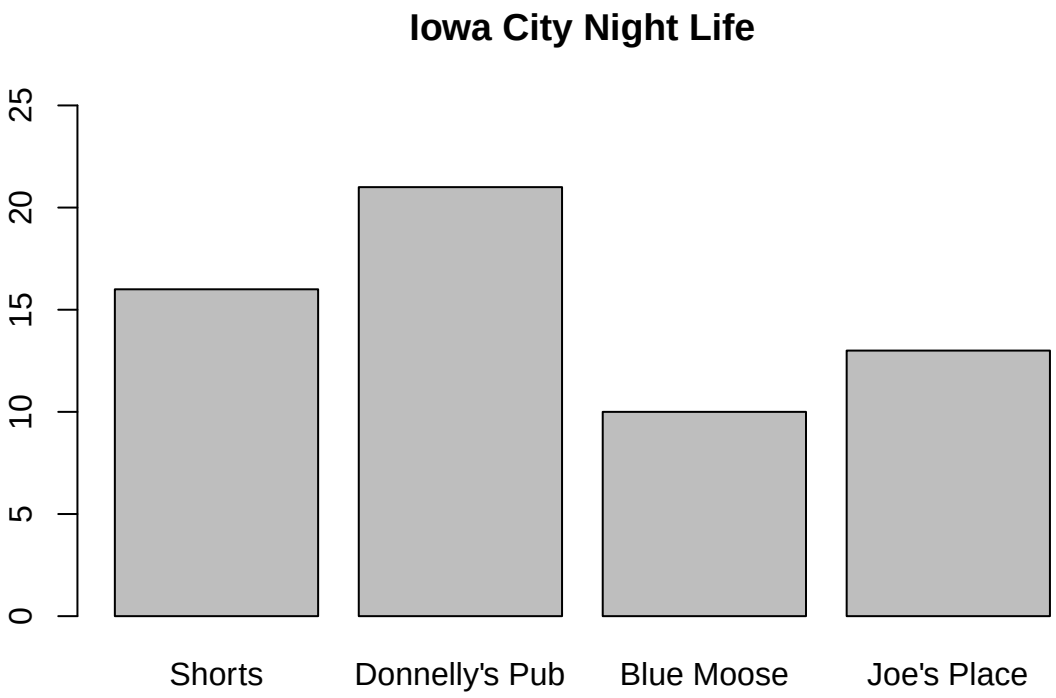


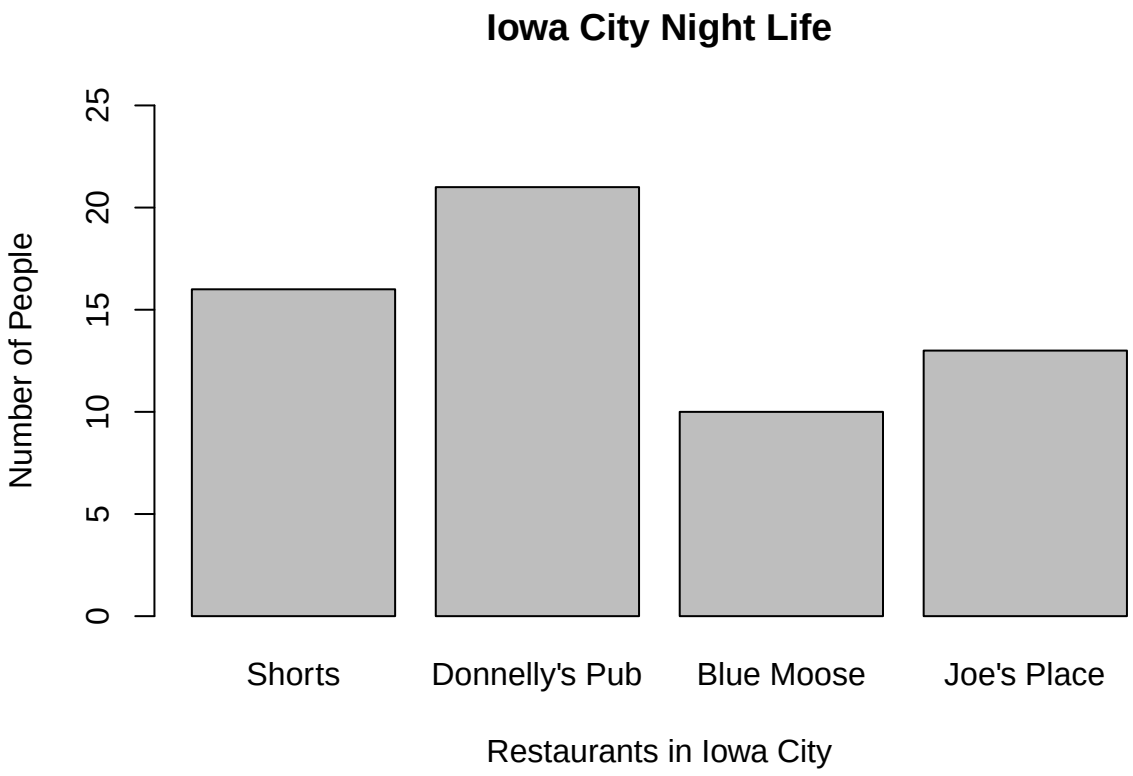
Step Five

Give the axes appropriate names and labels.

To do this first create a vector named "restaurants" that contains the names of the restaurants.

Hint: Use the "names.arg", "xlab", and "ylab" parameters.


Iowa City Night Life
0
5
10
15
20
25
Shorts
Donnelly's Pub
Blue Moose
Joe's Place

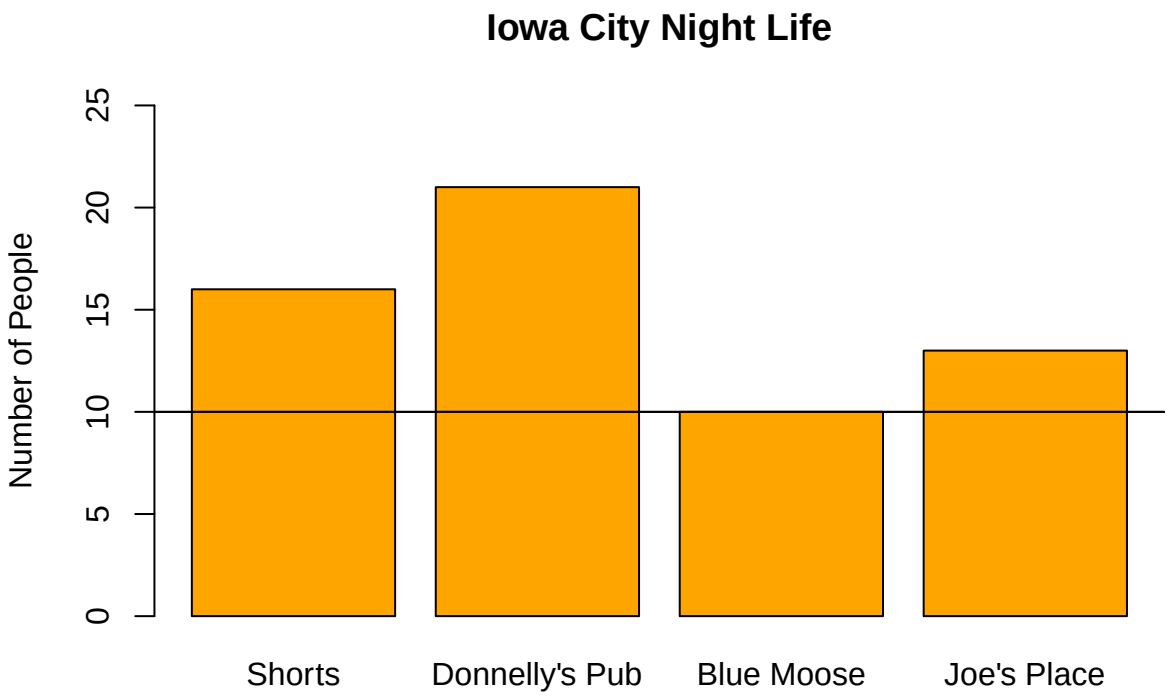


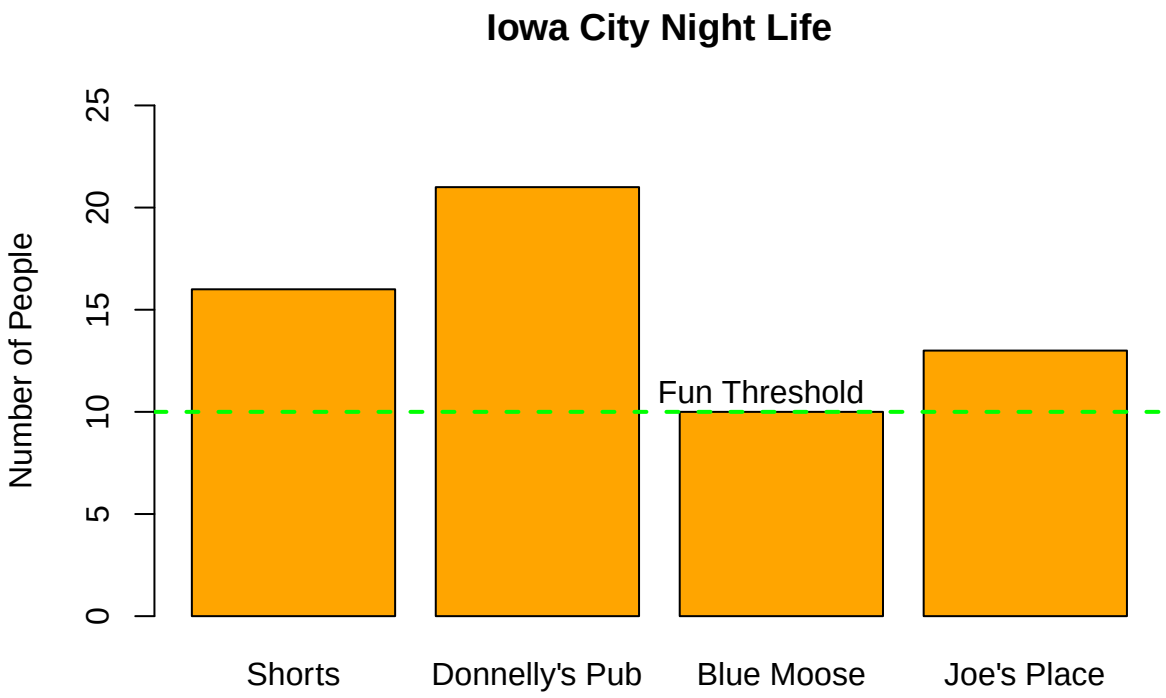
Step Six

Let's say that the threshold of having fun is 10 people.

Add a line at this threshold and label it "Fun Threshold"

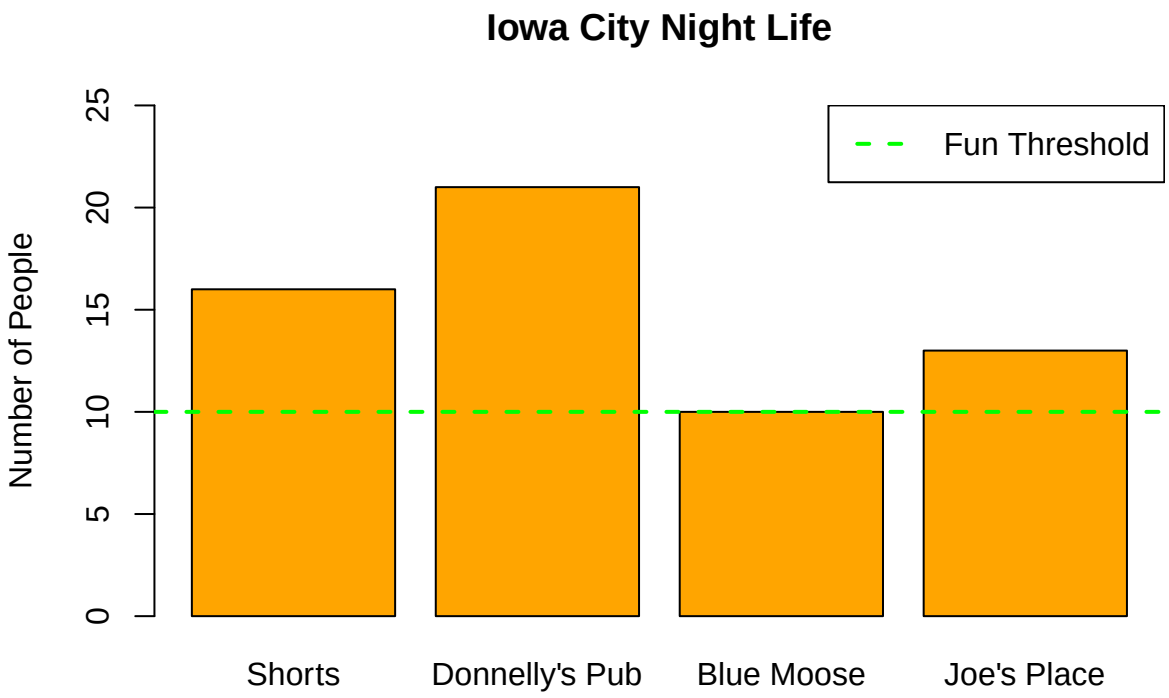
Hint: Use "abline", "text", and "legend" functions


Iowa City Night Life
Number of People
0
5
10
15
20
25
Shorts
Donnelly's Pub
Blue Moose
Joe's Place



Step Seven

Create a legend describing the line.



Step Eight

Save the plot.

Hint: Use the "pdf" and "dev.off" functions.

```
## pdf
##   2
```

Hypothesis Testing and Confidence Intervals

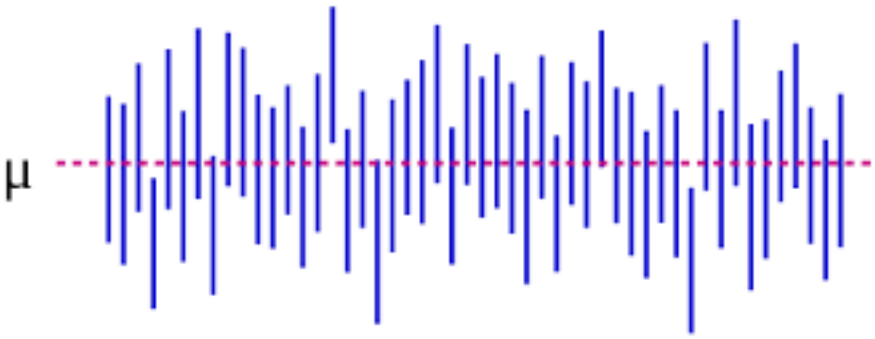
There is close relationship between confidence intervals and hypothesis testing. All values within a constructed 95% interval are considered "plausible" values for the parameter that we are estimating. Values outside the interval are rejected as unlikely and improbable.

Confidence Intervals Interpretation:

If you were to repeat the process of creating a confidence interval an infinite number of times, 95% of the interval estimates for μ will contain the true parameter value, μ. We treat the population mean μ as being fixed. Any particular interval may or may not contain the true population mean μ.

- We say that we are "95% confident" that the interval contains the true population μ because the procedure used to construct this interval produces a correct interval estimate 95% of the time.
- We **DO NOT** say there is a 95% probability that μ lies between these two values. (μ is fixed)

Visualization



Hypothesis Testing - "Null Until Proven Alternative"

In class, you learned that there are a lot of wrong ways to think about they hypothesis testing process. The courtroom is a helpful example that illustrates the correct usage of p-values and hypothesis tests. Let's look at it in terms of "innocent until proven guilty": As the person analyzing data, you are the judge. The hypothesis test is the trial, and the null hypothesis is the defendant. The alternative hypothesis is like the prosecution, which needs to make its case beyond a reasonable doubt (say, with 95% certainty).

If the evidence presented doesn't prove the defendant is guilty beyond a reasonable doubt, you still have not proved that the defendant is innocent. (We never say that we accept the null hypothesis)

So how would that verdict be announced? It enters the court record as "Not guilty." That phrase is perfect: "Not guilty" doesn't mean the defendant is innocent, because that has not been proven. It just means the prosecution couldn't prove its case to the necessary, "beyond a reasonable doubt" standard. It failed to convince the judge to abandon the assumption of innocence.

If you follow that rationale, then you can see that "failure to reject the null" is just the statistical equivalent of "not guilty." In a trial, the burden of proof falls to the prosecution. When analyzing data, the entire burden of proof falls to the sample data you've collected. This is why our sampling procedure is so important. Just as "not guilty" is not the same thing as "innocent," neither is "failing to reject" the same as "accepting" the null hypothesis.

This method of thinking about hypothesis tests will come in handy when we start formally testing our own hypotheses.

Source: http://blog.minitab.com/blog/understanding-statistics/things-statisticians-say-failure-to-reject-the-null-hypothesis

Relationship between Confidence Intervals & Hypothesis Testing

If the value of the parameter specified by the null hypothesis (for instance Ho = 0) is contained within the 95% interval, then the null hypothesis cannot be rejected at the 0.05 level. If the value specified by the null hypothesis is not in the interval, then the null hypothesis can be rejected at the 0.05 level. Likewise, for a 99% confidence interval, if the value specified by the null hypothesis is in the interval, then the null hypothesis cannot be rejected at the 0.01 level.

Practice Problems

In lab last week we worked with the titanic dataset. Today we are wanting to know whether sex played a significant role in the survival rates of the passengers on-board. Therefore, we want to compare survival rates between males and females.

1. Define the null hypothesis for this study on the 'titanic' dataset?
2. Say for example that we have the following null hypothesis $H_o : \mu_{female} = 0.5$. We obtain a 95% confidence interval (0.415, 0.481). Remember that interpretation of this confidence interval states that we are 95% confident that the true population μ lies within this interval. Would we reject or retain the null hypothesis?

True or False

3. If the null hypothesis can be rejected at the 0.05 level of significance, the confidence interval contains the specified null hypothesis?
4. If the null hypothesis cannot be rejected at the 0.01 level of significance, the confidence interval contains the specified null hypothesis?

Answers

1.

Ho: average survival rate for females = average survival rate for males

2.

Reject

3.

False

4.

True