BIOS: 4120 Lab 4

February 6-7, 2018

Objectives

In today's lab we will:

1. Learn how to visualize continuous data using figures
2. Compute and compare summary statistics
3. Review for Quiz 1

Summary Statistics

If you go to the class website you will find a data set named tailgating which we will be using today.

This study used driving simulation to evaluate the potential link between recreational drug use and risky driving behavior, as measured by average following distance during a car-following task. In the task, drivers were instructed to follow a lead vehicle that was programmed to randomly vary its speed. As it does so, more cautious drivers respond by following a safer distance, while riskier drivers respond by tailgating. The dataset variables include drug use status: ALC (alcohol), MDMA (ecstasy), THC (marijuana), and NODRUG (no drugs), distance, and binary drug use.

The data can be uploaded to R using the following code.

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

Which variables are continuous and which are categorical?

We have already learned how to compute some summary statistics in R, but today we will learn how to visual the distribution of continuous data. First, let's take a look at the summary of distance.

```
summary(tailgating$Distance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.82   27.57   32.49   41.01   39.52  356.96
```

Standard deviation

Summary tells us most of the information we would like to know. How about the standard deviation? Use the function 'sd'

```
sd(tailgating$Distance)
```

```
## [1] 44.16035
```

How would you generally interpret this? Do you think the data have a small or large spread?

Data by drug group

Now let's look at distance by drug group status.

```
by(tailgating$Distance, tailgating$Group, summary)
```

```
## tailgating$Group: ALC
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.89   28.83   35.42   36.83   40.21   68.34
## ------------------------------------------------------------
## tailgating$Group: MDMA
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.01   22.32   26.83   27.61   28.46   56.61
## ------------------------------------------------------------
## tailgating$Group: NODRUG
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.70   28.80   33.37   47.33   43.57  356.96
## ------------------------------------------------------------
## tailgating$Group: THC
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.82   27.75   31.90   42.61   39.52  346.72
```

Summarize the differences between the groups.

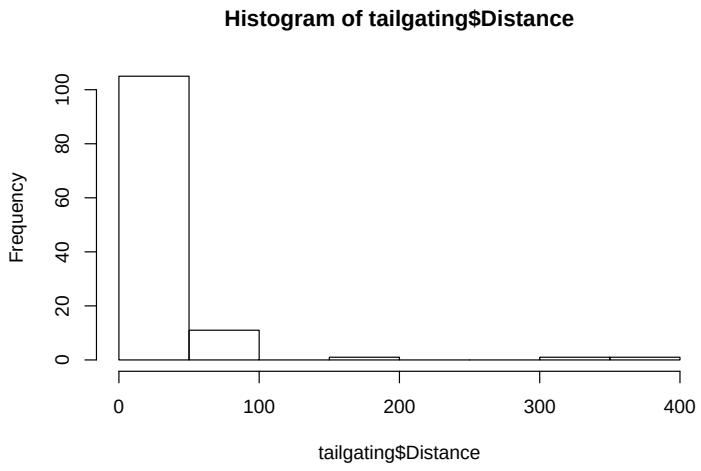
Is this an observational or controlled study?

Could this study be influenced by confounding factors? Why? And what might be some examples?

Histograms

As discussed in class, we can visualize the distribution of distance using a histogram. Here is how to do this in R:

```
hist(tailgating$Distance)
```

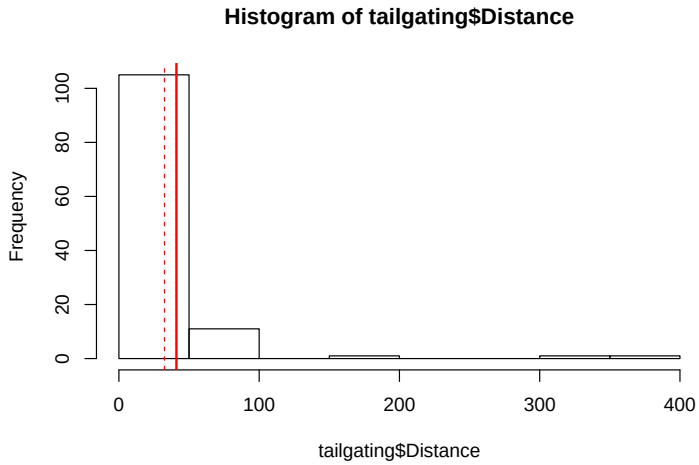


Is the distribution of distance normal (bell-shaped pattern)? If it is skewed, is it left- or right- skewed?

Compare the mean and median

We could add the mean and median to the plot using the function 'abline' to add lines. Let's see how the two compare.

```
hist(tailgating$Distance)
abline(v=mean(tailgating$Distance), col="red", lwd=2) #lwd makes the line thicker (line width)
abline(v=median(tailgating$Distance), col="red", lty=2) #lty makes the line dashed (line type)
```



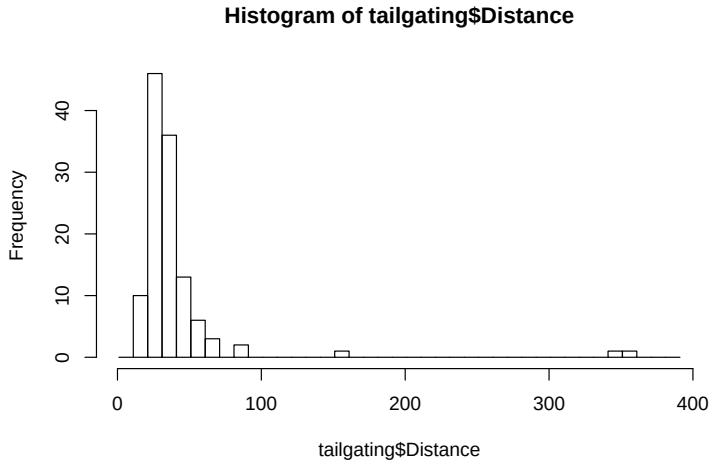

```
#the mean is the solid line and the median is the dashed line
```

Change the bin size

The bin size (increments seen on the x-axis) can impact how our data look. If these bins are large, we might not be able to see our data in detail. Above, our bin size is pretty large because the range is vast. Let's see how the data look if we change the bin size using the argument 'breaks'

As a refresher, the seq() function will give you a list of numbers where the minimum is the first argument the maximum is the second argument and the increment (OR for this purpose the bin size) is specified by the third argument.

```
hist(tailgating$Distance, breaks = seq(1, 400, 10)) #bin size of 10
```

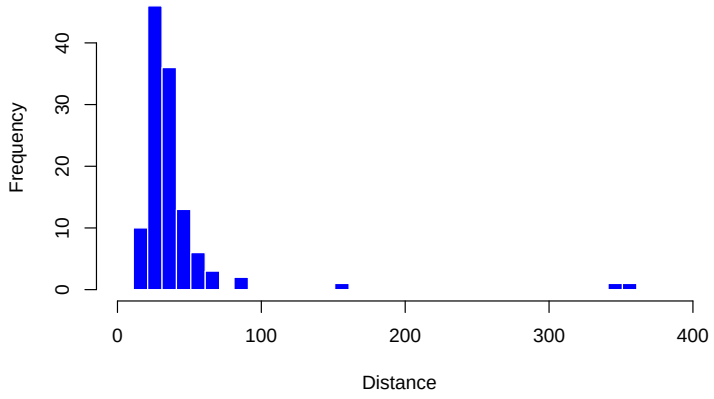


That's better! You can now see that most of the data follows a bell-shaped, normal distribution, but there are some crazy outliers that cause the data to be positively skewed.

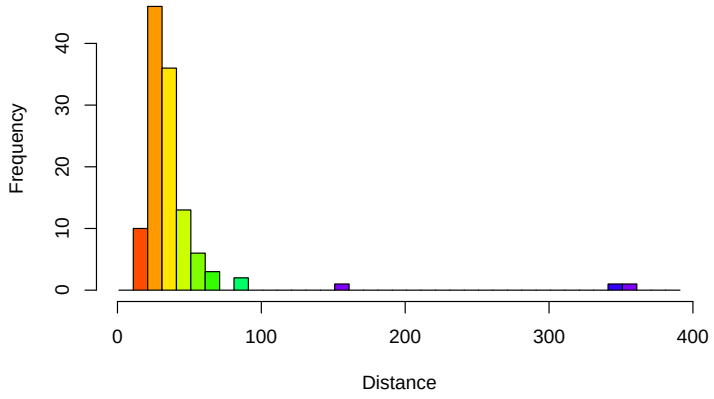
Customize your histogram

The great thing about R is there are some many options to customize your figures. Below is code for the same figure but I have added arguments to customize the x and y label (xlab & ylab, respectively). The main title function (main) allows you to create a title or you can choose to omit the default by using the argument "". There are also many color options in R. The col function allows you to color the bars. The rainbow palette is fun to use if you want an array of colors, but you need to specify how many different colors you want to use. In this case we used 20 colors from the rainbow.

```
#customized labels, solid color & white border
hist(tailgating$Distance, col= "blue", border="white", breaks = seq(1, 400, 10),
     xlab="Distance",
     ylab="Frequency",
     main = "")
```




```
#array of colors
hist(tailgating$Distance, col= rainbow(20), breaks = seq(1, 400, 10),
     xlab="Distance",
     ylab="Frequency",
     main = "")
```

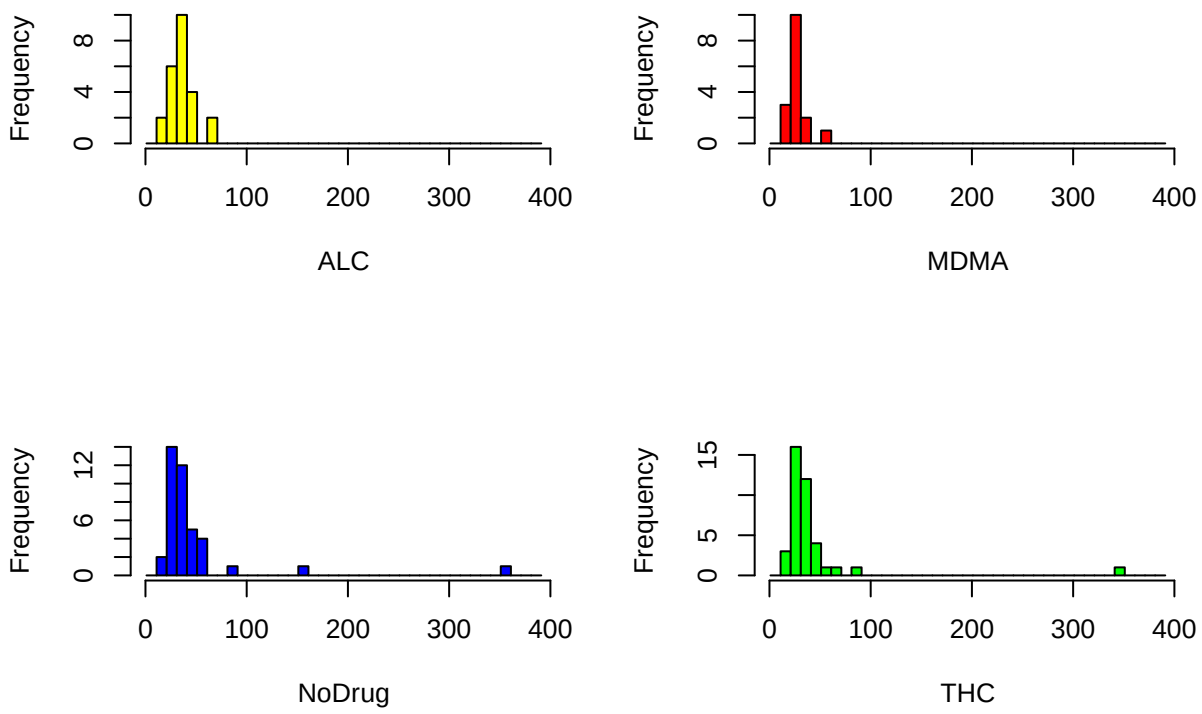


Specify histogram by drug group

Now let's visualize distance broken down by drug group

```
par(mfrow=c(2,2)) #view all four histograms in a 2 by 2 window

hist(tailgating$Distance[tailgating$Group=="ALC"], col= "yellow", breaks = seq(1, 400, 10),
     main = "", xlab = "ALC")
hist(tailgating$Distance[tailgating$Group=="MDMA"], col= "red", breaks = seq(1, 400, 10),
     main = "", xlab = "MDMA")
hist(tailgating$Distance[tailgating$Group=="NODRUG"], col= "blue", breaks = seq(1, 400, 10),
     main = "", xlab = "NoDrug")
hist(tailgating$Distance[tailgating$Group=="THC"], col= "green", breaks = seq(1, 400, 10),
     main = "", xlab = "THC")
```

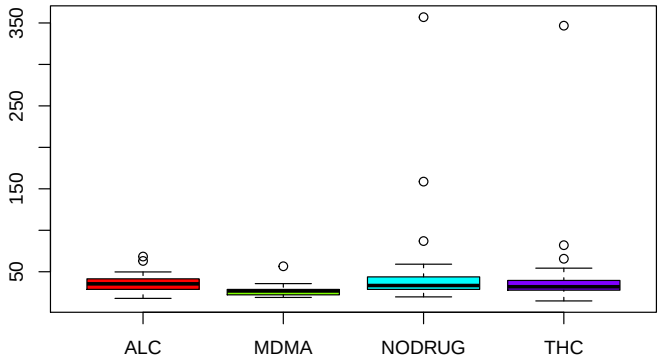


What groups are the main culprits of outliers?

Box Plots

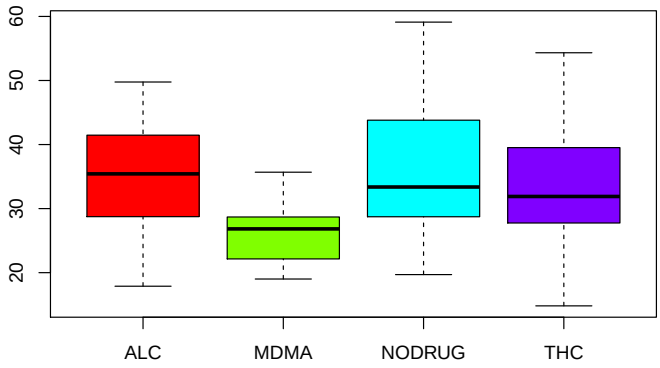
We can also plot this data using a box plot

```
boxplot(tailgating$Distance ~ tailgating$Group, col=rainbow(4))
```



Although it is good to know there are outliers, the large range can make it difficult to see the boxes. We can remove the outliers for a better look by using the argument outline= FALSE.

```
boxplot(tailgating$Distance ~ tailgating$Group, col= rainbow(4), outline=FALSE)
```



Calculate a quantile

As an FYI, you can find specific quantiles of interest using the quantile function (2nd argument asks for what quantile you would like)

```
quantile(tailgating$Distance, 0.30)
```

```
##      30%
## 28.15008
```

Quiz Review

Errors

```
##        H0 False H0 True
## Reject        A       B
## FtR           C       D
```

Type I Error

A Type I error is committed when a true null hypothesis is rejected.
In terms of disease detection (where the null hypothesis is no disease), this is a false positive.
In the table above, this is B.

Type I Error Rate (α)

The Type I error rate is the proportion of true hypotheses that were rejected.
In the table above, this is B/(B+D).

Type II Error

A Type II error is committed when a false null hypothesis is not rejected.
In the table above, this is C.

Type II Error Rate (β)

The Type II error rate is the proportion of false null hypotheses that failed to be rejected.
In the table above, this is C/(C+A).

False Discovery Rate

The false discovery rate is the fraction of null hypothesis rejections that were incorrect.
In the table above, this is B/(B+A).

Question: Fill out the table using the following information: 800 experiments were conducted and the null hypothesis was true 700 times. The type I error rate was 0.10 and the type II error was 0.20.

Vocab recap

Generalizability

We use hypothesis tests and confidence intervals to make inferences about the population using the information from our sample. However, it is important to consider whether the sample is representative of the population. If we are comparing the health outcomes of mice to the human population, we violate the principle that the sample is representative of the population. We can still learn about potential health outcomes in mice studies but we cannot conclude that the results will be exactly what we would observe in a true human population.

P-value

The probability of obtaining results as extreme or more extreme than the one observed in the sample, given that the null hypothesis is true.

Selection bias

Instead of random sampling, certain subgroups of the population were more likely to be included than others.

Nonresponse bias

Nonresponders can differ from responders in many important ways

Perception bias

The perception of benefit from a treatment (placebo effect)

Diagnostic bias

Doctors may change their diagnosis if they know whether a patient has been given treatment/placebo. In double blind studies, this is not an issue.

In each of the following examples, determine which bias(es) may be present. If possible, determine which direction the bias may skew the results. Then, state the null and alternative hypotheses.

A doctor wanted to investigate whether Tylenol is better than Ibuprofen in curing head-aches, so he designed an experiment in which he randomly selected which treatment he would give people and blinded them to which one they got. He then noted how much their condition improved in either case.

A parent-teacher association for schools in Austin, Minnesota were wondering how pervasive drug culture was among their high school students, compared to the national average. In order to gain a handle on the situation, they handed out a survey to the students at a school assembly during homecoming week.

In a randomized controlled double blind study, 12 people in the treatment group died before receiving the treatment, so the researchers decided to omit them from the data analysis.

Weighted mean & proportion set up

What study design is used to eliminate confounding between the treatment and control groups?

What technique did we learn in class to adjust for potential confounding?

What three components do you need to know to find a weighted mean?

What is the general notation and formula to find a weighted mean?

Know how to word the results of hypothesis tests and confidence intervals