Lab-4-Solutions

Linder Wendt

2/11/2020

Objectives

In today's lab we will:

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

Summary Statistics

Today we will be using the tailgating dataset. 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.

Is this an observational or controlled study?

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

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. The 'by' function allows us to run a function over data set into groups.

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

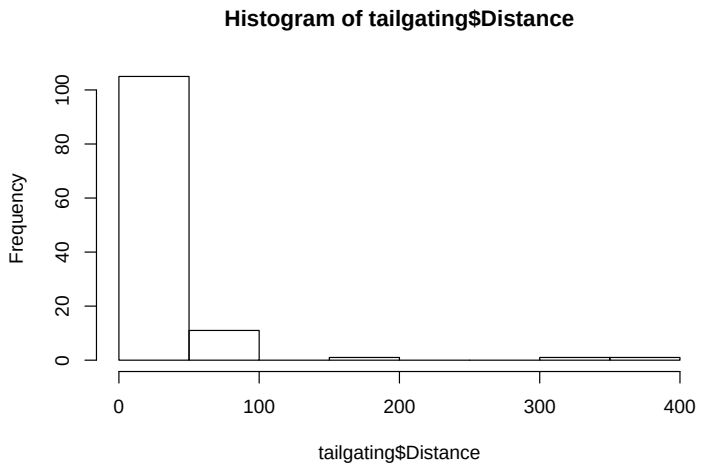
```
## tailgating$Group: ALC
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.89   28.83   35.42   36.83   40.21   68.34
## ------------------------------------------------------------
## tailgating$Group: MDMA
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.01   22.32   26.83   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.

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)
```

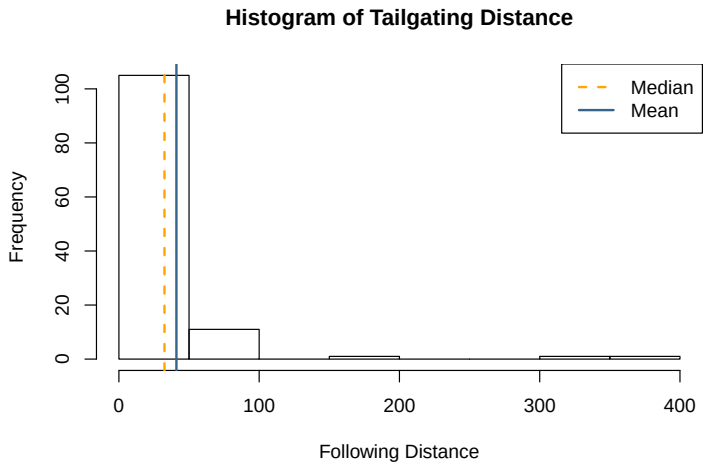


Is the distribution of distance normal (bell-shaped pattern)? If it is skewed, is it left- or right- skewed?

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$Distance, main = "Histogram of Tailgating Distance", xlab = "Following Distance")
abline(v=mean(tailgating$Distance), col="steelblue4", lwd=2) #lwd makes the line thicker (line width)
abline(v=median(tailgating$Distance), col="orange", lwd=2, lty = 2)#lty makes the line dashed (line typ
legend(x = "topright",
       legend = c("Median", "Mean"),
       col = c("orange", "steelblue4"),
       lwd = 2, lty = c(2,1))
```



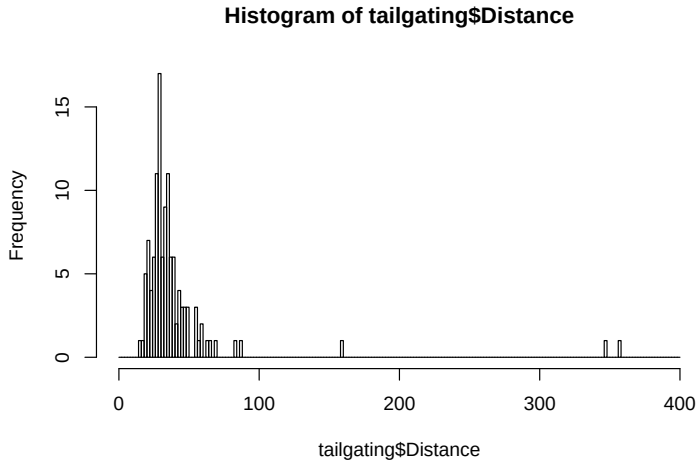

```
#the mean is the solid line and the median is the dashed line
```

Change the bin size

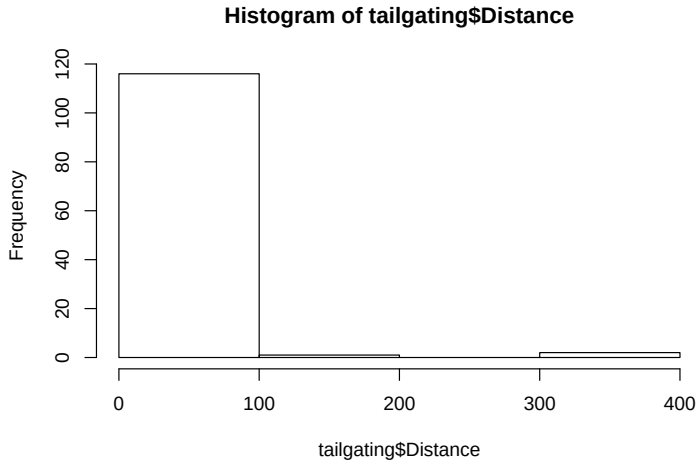
The bin size (increments seen on the x-axis) can impact how our data look. If these bins are large, we might not be able to see our data in detail. Above, our bin size is pretty large because the range is vast. Let's see how the data look if we change the bin size using the argument 'breaks'

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.

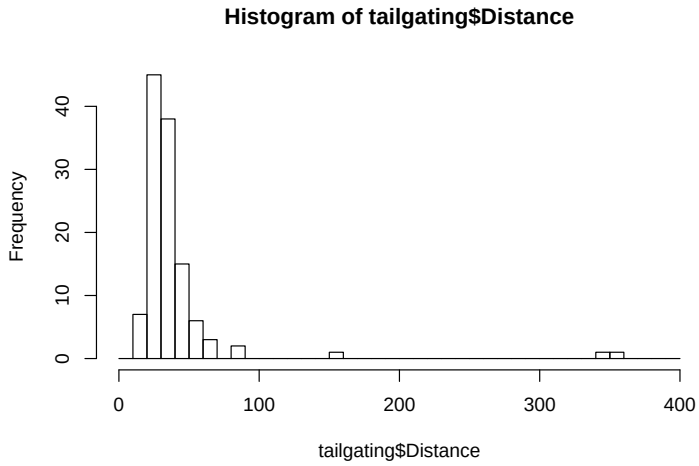
```
hist(tailgating$Distance, breaks = seq(0, 400, 2)) #bin size of 2
```




```
hist(tailgating$Distance, breaks = seq(0, 400, 100)) #bin size of 100
```




```
hist(tailgating$Distance, breaks = seq(0, 400, 10)) #bin size of 10
```

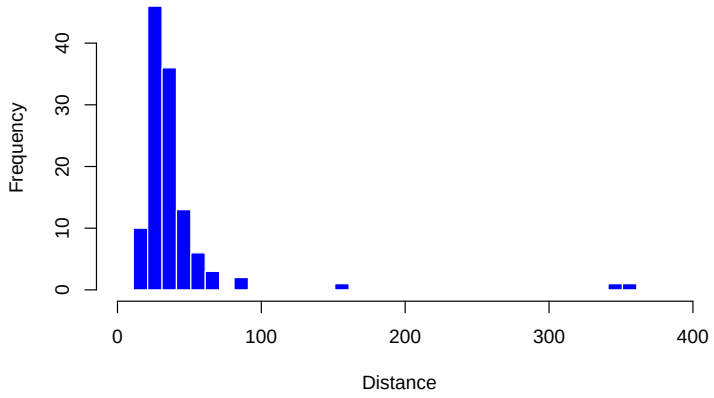


That's better! You can now see that most of the data follows a bell-shaped, normal distribution, but there are some crazy outliers that cause the data to be positively skewed. The group with a bin size of 10 allows us to see an apporopriate amount of detail in the plot.

Customize your histogram

The great thing about R is there are 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.

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

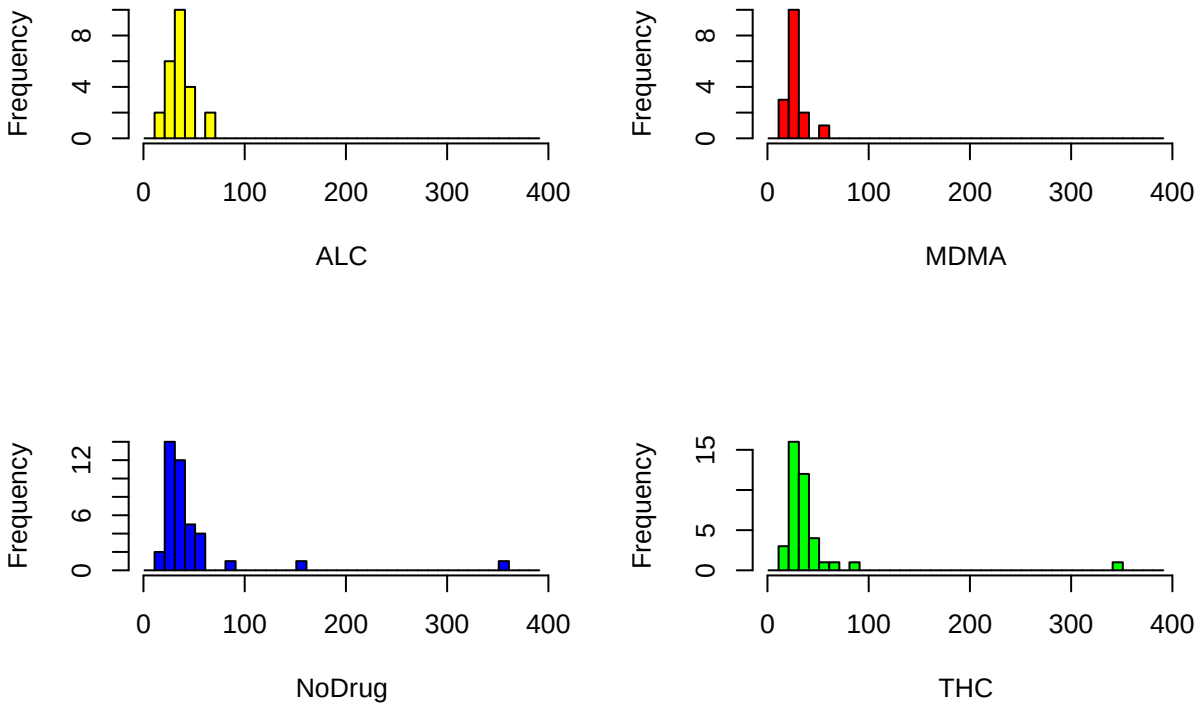


Specify histogram by drug group

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")
```

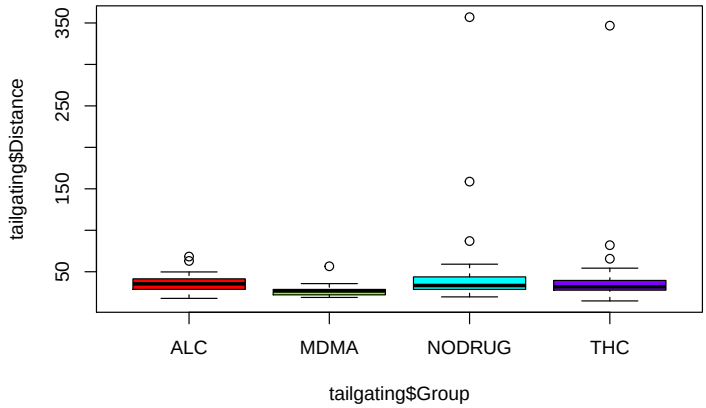


What groups are the main culprits of outliers?

Box Plots

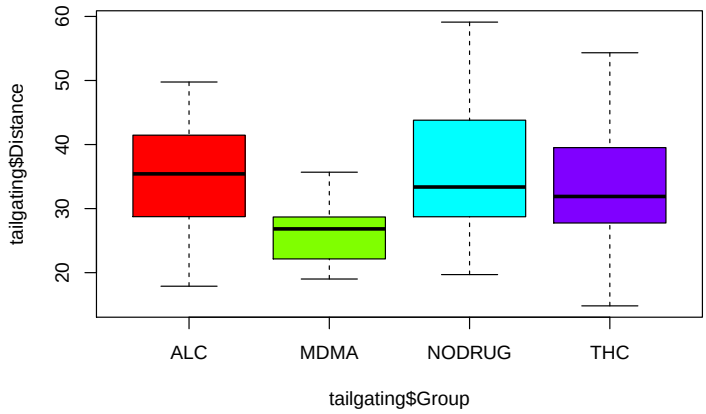
We can also plot this data using a box plot

```
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

Observational vs Experimental

State whether each of the following scenarios are an observation or an experimental study design

** Assume that we are interested in assessing the effects of mothers' smoking habits during pregnancy on the weight of their babies. To assess this question, we collected data including the mothers' smoking habits during pregnancy and the birth weight of their babies from medical records.

Observational

** Assume that we are interested in assessing whether students perform better on exams if they drink a caffenated beverage the morning of the exam. We randomize students into two groups - one group that drinks caffenated coffee and one that drinks decaf coffee on the day of the exam. We then assess exam performance for each group using a predetermined metric.

Experimental

Errors

Test	NullT	NullF	Total
RejY			
RejN			
Total			

Type I Error

A Type I error is committed when a true null hypothesis is rejected. In other words, a type I error is the probability of rejecting the null hypothesis when the null is in fact true. In terms of disease detection (where the null hypothesis is no disease), this is a false positive.

Type I Error Rate (α)

The Type I error rate is the proportion of true hypotheses that were rejected.

Type II Error

A Type II error is committed when a false null hypothesis is not rejected. In other words, a type II error is the probability of failing to reject the null hypothesis when the null is in fact false.

Type II Error Rate (β)

The Type II error rate is the proportion of false null hypotheses that failed to be rejected.

False Discovery Rate

The false discovery rate is the fraction of null hypothesis rejections that were incorrect.

Practice Questions:

1. Fill in the table above using the following information: Suppose that an investigator conducts 800 experiments with the null hypothesis being true 700 times. The investigator rejected the null hypothesis when the null was true 10% of the time and failed to reject the null when the null was false 20% of the time.

Test	NullT	NullF	Total
RejY	70	80	150
RejN	630	20	650
Total	700	100	800

2. Suppose instead that a funding agency allocates resources such that 400 studies can take place. Of those studies, 140 did not have statistically significant results. Of the studies that were found to be statistically significant, 100 were actualy due to random chance and of those without statistically significant results 60 were not due to chance.

Test	NullT	NullF	Total
RejY	100	160	260
RejN	80	60	140
Total	180	220	400

Note: For the tables from the previous two problems, the positioning of the two rows and columns can vary without changing the overall meaning of the table.

3. Consider a study in which researchers were interested in whether students think better when they were standing versus when they were seated, and were also interested in whether drinking/eating may effect academic performance. To test this researchers asked 7th grade students across several different middle schools around the country a series of questions and recorded whether or not they answered all questions correctly. Students were assigned at random to be either seated or standing when given the questions, and they were also randomized to be either drinking or eating during the process. Their data are presented below:

	Drinking		Eating	
Group	Perfect.Responses	Total.Responses	Perfect_Responses	Total_Responses
Sitting	71	134	88	104
Standing	94	142	57	90

(a) Is this a controlled experiment or observational study?

Controlled Experiment

(b) Which group had a higher success rate (standing or sitting)?

Sitting Success Rate: $\frac{71+88}{134+104} = .668$ Standing Success Rate: $\frac{94+57}{142+90} = .651$

(c) Which type of consumption (eating or drinking) had a higher success rate?

Eating Success Rate: $\frac{88+57}{104+90} = .747$ Drinking Success Rate: $\frac{71+94}{134+143} = .598$

Note: This is not subject to confounding, as it is a randomized experiment.

Vocab recap

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

Generalizing from a sample to a different population

Anytime the study violates the principle of generalizing to the population that the sample was drawn from.

In each of the following examples, determine which bias(es), if any, may be present.

** Doctors want to investigate whether Tylenol performs better than Ibuprofen in curing head-aches. They design an experiment in which participants are randomly assigned to one of the two treatments. The patients are blind to which treatment they receive.

No Bias

** A parent-teacher association at an affluent school in Chicago, Illinois wanted to study how pervasive the drug culture was among high school students in the United States. To answer this question, they handed out a survey to their high school students at a school assembly.

Selection Bias

** A recent investigator conducted a survey to study how long New Year's resolutions last. The individuals who were still on track with their goals were more likely to respond.

Non-Response Bias