Solutions to Homework 1

Statistics 302 Professor Larget

Textbook Exercises

2.13 Rock-Paper-Scissors Rock-Paper-Scissors, also called Roshambo, is a popular two-player game often used to quickly determine a winner and loser. In the game, each player puts out a fist (rock), a flat hand (paper), or a hand with two fingers extended (scissors). In the game, rock beats scissors which beats paper which beats rock. The question is: Are the three options selected equally often by players? Knowing the relative frequencies with which the options are selected would give a player a significant advantage. A study observed 119 people playing Rock-Paper-Scissors. Their choices are shown the table.

(a) What is the sample in this case? What is the population? What does the variable measure?

(b) Construct a relative frequency table of the results.

(c) If we assume that the sample relative frequencies from part (b) are similar for the entire population, which option should you play if you want the odds in your favor?

(d) The same study determined that, in repeated plays, a player is more likely to repeat the option just picked than to switch to a different option. If your opponent just played paper, which option should you pick for the next round?

Option Selected	Frequency
Rock	66
Paper	39
Scissors	14
Total	119

Solution

(a) The sample is the 119 players who were observed. The population is all people who play rock-paper-scissors. The variable records which of the three options each player plays. This is a categorical variable.

(b) A relative frequency table is shown below. We see that rock is selected much more frequently than the others, and then paper, with scissors selected least often.

Option Selected	Relative Frequency
Rock	0.555
Paper	0.328
Scissors	0.118
Total	1.0

(c) Since rock is selected most often, your best bet is to play paper.

(d) Your opponent is likely to play paper again, so you should play scissors.

2.14 Home Field Advantage in Soccer In the book *Scorecasting*, we learn that "Across 43 professional soccer leagues in 24 different countries spanning Europe, South America, Asia, Africa, Australia, and the United States (covering more than 66,000 games), the home field advantage [percent of games won by the home team] in soccer worldwide is 62.4%." Is this a population or a sample? What are the cases and approximately how many are there? What is the variable and is it categorical or quantitative? What is the relevant statistics, including correct notation?

Solution

The dataset includes all professional soccer games, so this is a population. The cases are the soccer games, and there are approximately 66,000. The variable is whether or not the home team won the game, and it is categorical. The relevant statistic is $p = 0.624$.

2.25 Smoking and Pregnancy Rate Studies have concluded that smoking while pregnant can have negative consequences, but could smoking also negatively one's ability to become pregnant? A study collected data on 678 women who had gone off birth control with the intention of becoming pregnant. Smokers were defined as those who smoked at least one cigarette a day prior to pregnancy. We are interested in the pregnancy rate during the first cycle off birth control. The results are summarized in the table below.

	Smoker	Non-Smoker	Total
Pregnant	38	206	244
Not Pregnant	97	337	424
Total	135	543	678

(a) Is this an experiment or an observational study? Can we use these data to determine whether smoking influences one's ability to get pregnant? Why or why not?
(b) What is the population of interest?
(c) What is the proportion of women successfully pregnant after their first cycle ($\hat{p}$)? Proportion of smokers successful ($\hat{p}_s$)? Proportion of nonsmokers successful ($\hat{p}_{ns}$)?
(d) Find and interpret ($\hat{p}_{ns} - \hat{p}_s$) the difference in proportion of success between non-smokers and smokers.

Solution

(a) Since no one assigned smoking or not to the participants, this is an observational study. Because this is an observational study, we can not use this data to determine whether smoking influences one's ability to get pregnant. We can only determine whether there is an association between smoking and ability to get pregnant.

(b) The sample collected is on women who went off birth control in order to become pregnant, so the population of interest is women who have gone off birth control in an attempt to become pregnant.

(c) We look in the total section of our two way table to find that out of the 678 women attempting to become pregnant, 244 succeeded in their first cycle, so $\hat{p} = 244/678 = 0.36$. For smokers we look only in the Smoker column of the two way table and observe 38 of 135 succeeded, so $\hat{p}_s = 38/135 = 0.28$. For non-smokers we look only in the Non-smoker column of the two way table and observe 206 of 543 succeeded, so $\hat{p}_{ns} = 206/543 = 0.38$.

(d) For the difference in proportions, we have $\hat{p}_{ns} - \hat{p}_s = 0.38 - 0.28 = 0.10$. This means that in this sample, the percent of non-smoking women successfully getting pregnant in the first cycle is 10 percentage points higher than the percent of smokers.

2.31 Which of These Things Is Not Like the Other? Four students were working together on a project and one of the parts involved making a graph to display the relationship in a two-way table of data with two categorical variables: college accept/reject decision and type of high school (public, private, parochial). The graphs submitted by each student are shown in the book. Three are from the same data, but one is inconsistent with the other three. Which is the bogus graph? Explain.

Solution
Graph (b) is the impostor. It shows more parochial students than private school students. The other three graphs have more private school students than parochial.

2.57 Fiber in the Diet The number of grams of fiber eaten in one day for a sample of ten people are

10 11 11 14 15 17 21 24 28 115

(a) Find the mean and the median for these data.
(b) The value of 115 appears to be an obvious outlier. Compute the mean and the median for the nine numbers with the outlier excluded.
(c) Comment on the effect of the outlier on the mean and the median.

Solution
(a) The mean is $\bar{x} = \frac{10+11+11+14+15+17+21+24+28+115}{10} = \frac{266}{10} = 26.6$
The median is $\frac{15+17}{2} = 16$.

(b) Without the outlier, we have $\bar{x} = 16.78$. Since n = 9, the median is the middle number. We have m = 15.

(c) The outlier has a very significant effect on the mean and very little effect on the median.

2.58 Beta-Carotene Levels in the Blood The plasma beta-carotene level (concentration of beta-carotene in the blood), in ng/ml, was measured for a sample of $n = 315$ individuals, and the results are shown in the histogram in the book.

(a) Describe the shape of the distribution. Is it symmetric or skewed? Are there any obvious outliers?
(b) Estimate the median of this sample.
(c) Estimate the mean of this sample.

Solution
(a) The distribution has a right skew. There are a number of apparent outliers on the right side.

(b) The actual median is 140 ng/ml. Estimates between 120 and 160 are reasonable.

(c)The actual mean is 189.9 ng/ml. Estimates between 160 and 220 are reasonable. Note that the outliers and right skew should make the mean larger than the median.

2.62 Does Sexual Frustration Increase the Desire for Alcohol? Apparently, sexual frustration increases the desire for alcohol, at least in fruit flies. Scientists randomly put 24 fruit flies into one of two situation. The 12 fruit flies in the ”mating” group were allowed to mate freely with many available females eager to mate. The 12 in the “rejected” group were put with females that had already mated and thus rejected any courtship advances. After four days of either freely mating or constant rejection, the fruit flies spent three days with unlimited access to both normal fruit fly food and the same food soaked in alcohol. The percent of time each fly chose the alcoholic food was measured. The fruit flies that had freely mated chose the two types of food about equally often, choosing the alcohol variety on average 47% of the time The rejected males, however, showed a strong preference for the food soaked in alcohol, selecting it on average 73% of the time. (The study was designed to study a chemical in the brain called neuropeptide that might play a role in addiction.)

(a) Is this an experiment or an observational study?
(b) What are the cases in this study? What are the variables? Which is the explanatory variable and which is the response variable?
(c) We are interested in the difference in means, where the means measure the average percent preference for alcohol (0.47 and 0.73 in this case). Find the difference in means and give the correct notation for your answer, using the correct notation for a mean, subscripts to identify groups, and a minus sign.
(d) Can we conclude that rejection increases a male fruit fly's desire for alcohol? Explain.

Solution
(a) This is an experiment since the treatment was randomly assigned and imposed.

(b) The cases are the 24 fruit flies. There are two variables. The explanatory variable is which of the two groups the fly is in. The response variable is percent of time the alcoholic mixture is selected.

(c) Using $\bar{x}_R$ for the mean of the rejected group and $\bar{x}_M$ for the mean for the mated group, we have $\bar{x}_R - \bar{x}_M = 0.73 - 0.47 = 0.26$.

(d) Yes, since this was a randomized experiment.

2.66 Number of Children The first table below shows the number of women (per 1000) between 15 and 44 years of age who have been married grouped by the number of children they have had. The second table below gives the same information for women who have never been married.

Table 1: Women who have been married

Number of Children	Women per 1000
0	162
1	190
2	290
3	289
4	48
5+	21

(a) Without doing any calculations, which of the two samples appears to have the highest mean number of children? Which of the distributions appears to have the mean most different from the median? Why?
(b) Find the median for each dataset.

Table 2: Women who have never been married

Number of Children	Women per 1000
0	791
1	108
2	53
3	29
4	12
5+	7

Solution

(a) It appears that the mean of the married women is higher than the mean of the never married women. We expect that the mean and the median will be the most different for the never married women, since that data is quite skewed while the married data is more symmetric.

(b) We have n = 1000 in each case. For the married women, we see that 162 women had 0 children, 190 had 1 child, and 290 had 2 children, so 162 + 190 + 290 = 642 had 0, 1, or 2 children. Less than half the women had 0 or 1 child and more than half the women had 0, 1, or 2 children so the median is 2. For the never married women, more than half the women had 0 children, so the median is 0.

2.101 Laptop Computers and Sperm Count Studies have shown that heating the scrotum by just 1 degree Celsius can reduce sperm count and sperm quality, so men concerned about fertility are cautioned to avoid too much time in the hot tub or sauna. A new study suggests that men also keep their laptop computers off their laps. The study measured scrotal temperature in 29 healthy male volunteers as they sat with legs together and a laptop computer on the lap. Temperature increase in the left scrotum over a 60-minute session is given as 2.31 ± 0.96 and a note tells us that "Temperatures are given as degrees Celsius; value are shown as mean ± SD." The abbreviation SD stands for standard deviation. (Men who sit with their legs together without a laptop computer do not show an increase in temperature.)

(a) If we assume that the distribution of the temperature increases for the 29 men is symmetric and bell-shaped, find an interval that we expect to contain about 95% of the temperature increases.

(b) Find and interpret the z-score for one of the men, who had a temperature increase of 4.9 degrees.

Solution

(a) We expect that 95% of the data will lie between $\bar{x} \pm 2s$. In this case, the mean is $\bar{x} = 2.31$ and the standard deviation is $s = 0.96$, so 95% of the data lies between $2.31 \pm 2(0.96)$. Since $2.31 - 2(0.96) = 0.39$ and $2.31 + 2(0.96) = 4.23$, we estimate that about 95% of the temperature increases will lie between 0.39° and 4.23°.

(b) Since $\bar{x} = 2.31$ and $s = 0.96$, the z-score for a temperature increase of 4.9° is

$$\text{z-score} = \frac{x - \bar{x}}{s} = \frac{4.9 - 2.31}{0.96} = 2.70.$$

The temperature increase for this man is 2.7 standard deviations above the mean.

2.112 Daily Calorie Consumption The five number summary for daily calorie consumption for the $n = 315$ participants in the **NutritionStudy** is (445, 1334, 1667, 2106, 6662).

(a) Give the range and the IQR.

(b) Which of the following numbers is most likely to be the mean of this dataset? Explain.

1550, 1667, 1796, 3605

(c) Which of the following numbers is most likely to be the standard deviation of this dataset? Explain.

5.72, 158, 680, 1897, 5315

Solution

(a) The range is $6662 - 445 = 6217$ and the interquartile range is $IQR = 2106 - 1334 = 772$.

(b) The maximum of 6662 is clearly an outlier and we expect it to pull the mean above the median. Since the median is 1667, the mean should be larger than 1667, but not too much larger. The mean of this data set is 1796.

(c) The best estimate of the standard deviation is 680. We see from the five number summary that about 50% of the data is within roughly 400 of the median, so the standard deviation is definitely bigger than 200. The two values above 680 would be way too large to give an estimated distance of the data values from the mean, so the only reasonable answer is 680. The actual standard deviation is 680.3.

2.136 Hits in Baseball Major League Baseball is split into two leagues, the National League (NL) and the American League (AL). The main difference between the two leagues is that pictures take at bats in the National League but not in the American League. Are total team hits different between the two leagues? The figure in the book shows side-by-side box-plots for the two leagues. The data are stored in **Baseball Hits**.

(a) Estimate the median number of hits for each league, and estimate the difference in median hits between the two leagues. Which league appears to get more hits?

(b) What is the other obvious difference between the two leagues.

Solution

(a) The median corresponds to the middle line in each box. An estimated median for the AL is about 1455 hits, and an estimated median for the NL is about 1410 hits (other answers are possible, but should be similar). So an estimated difference in median number of hits is $1455 - 1410 = 45$ hits. The American League has more hits.

(b) The other obvious difference is that the variability is greater for teams in the American League. Other correct ways to state this are that the standard deviation is greater, the IQR is greater, or the range is greater.

2.138 Daily Calories by Gender The figure in the book shows side-by-side box plots comparing calorie consumption by gender.

(a) Which gender has the largest median daily calorie consumption? Which gender has the largest outliers? Which gender has the most outliers?

(b) Does there seem to be an association between gender and calorie consumption? Explain.

Solution

(a) The median appears to be about 500 calories higher for the males than for the females. The largest outlier of 6662 calories in one day is a male, but the females have many more outliers.

(b) Yes, there does appear to be an association. Females appear to have significantly lower calorie consumption than males. We see that every number in the five number summary is higher for males than it is for females. The median for females is even lower than the first quartile for males.

2.173 Vegetables and Obesity The **USStates** dataset includes information on the 50 US states, including the percent of the population of each state that eats at least five servings of fruits and vegetables a day and the percent of the population of each state that is obese. The figure in the book shows a scatterplot of these two variables.

(a) Does the scatterplot show a positive or negative association? Explain why your answer makes sense for these two variables.
(b) Where would a very healthy state be located on the scatterplot: top left, top right, bottom left, bottom right, or middle? What about a very unhealthy state?
(c) Pick a point in a very healthy location in the scatterplot, and use the dataset **USStates** to find the state it represents. Pick a point in a very unhealthy location and find the state it represents.
(d) Is the data from a sample or a population? What is the correct notation for the correlation?
(e) Which of the following is most likely to be the correlation between these two variables?

-1, -0.941, -0.605, -0.083
0.172, 0.445, 0.955, 1

(f) Would a positive correlation imply that eating more vegetables will cause you to gain weight?
(g) Would a negative correlation imply that eating more vegetables will cause you to lose weight?
(h) One state stands out for eating an average number of vegetables but having a particularly low obesity rate. What state is this?

Solution

(a) There appears to be a negative association, which means in this context that states with a larger proportion of the population eating lots of vegetables tend to have a lower obesity rate. This makes sense since people who eat lots of vegetables are less likely to be obese.

(b) A healthy state would include a large percentage of people eating lots of vegetables and a small percentage of people who are obese, which corresponds to the bottom right corner. An unhealthy state would include a small percentage of people eating lots of vegetables and a high obesity rate, which corresponds to the top left corner.

(c) There is one point furthest in the bottom right corner and it corresponds to the state of Vermont. There are three dots in the unhealthy top left corner, and they correspond to the states of Mississippi, Kentucky, and Oklahoma.

(d) Since all 50 US States are included, this is a population. The correct notation is ρ.

(e) The two variables appear to be negatively correlated with a moderately strong linear relationship. The correct correlation is $\rho = -0.605$.

(f) No, correlation does not imply cause and effect relationships.

(g) No, correlation does not imply cause and effect relationships.

(h) That point corresponds to the state of Colorado.

2.176 SAT Scores: Math vs Verbal The **StudentSurvey** dataset includes scores on the Math and Verbal portions of the SAT exam.

(a) What would a positive relationship between these two variables imply about SAT scores? What would a negative relationship imply?

(b) The figure in the book shows a scatterplot of these two variables. For each corner of the scatterplot (top left, top right, bottom left, bottom right), describe a student whose SAT scores place him or her in that corner.

(c) Does there appear to be a strong linear relationship between these two variables? What does that tell you about SAT scores?

(d) Which of the following is most likely to be the correlation between these two variables?

-0.941, -0.605, -0.235, 0.445, 0.751, 0.955

Solution

(a) A positive relationship would imply that a student who is good at one of the tests is also likely to be good at the other - that students are generally either good at both or bad at both. A negative relationship implies that students tend to be good at either math or verbal but not both.

(b) A student in the top left is good at verbal and bad at math. A student in the top right is good at both. A student in the bottom left is bad at both, and a student in the bottom right is good at math and bad at verbal.

(c) There is not a strong linear relationship as the dots appear to be all over the place. This tells you that the scores students get on the math and verbal SAT exams are not very closely related.

(d) Since the linear relationship is not very strong, the correlation is likely to be one of the values closest to zero -either -0.235 or 0.445. Since there is more white space in the top left and bottom right corners than in the other two corners, the weak relationship appears to be a positive one. The correct correlation is 0.445.

Computer Exercises

Before completing these exercises, you will need to install the ggplot2 library. Open R and do this one time only.

```
> install.packages(”ggplot2”)
```

To actually load the this package into your active session, type the following command. You need to do this each session.

```
> library(”ggplot2”)
```

Load the data set SleepStudy from the textbook into R.

R problem 1 Make a table of the variables Stress and AlcoholUse.

R Code:

```
library(ggplot2)
library(Lock5Data)
data(SleepStudy)
with(SleepStudy,table(Stress,AlcoholUse))
```

R Output:

```
> with(SleepStudy,table(Stress,AlcoholUse))
        AlcoholUse
Stress   Abstain Heavy Light Moderate
  high         8     5    20       23
  normal      26    11    63       97
```

Using the output from R, we are table to create a table which displays the number of students which fit into the categories of Stress and AlcoholUse from the SleepStudy data.

	Abstain	Heavy	Light	Moderate
high	8	5	20	23
normal	26	11	63	97

1. What proportion of students are in each alcohol use category?

R Code:

```
prop = with(SleepStudy,table(AlcoholUse))
round(prop/sum(prop), 3)
```

R Output:

```
> round(prop/sum(prop), 3)
AlcoholUse
 Abstain    Heavy    Light Moderate
   0.134    0.063    0.328    0.474
```

From the R output, we are able to create a table which shows the proportion of students which are in each category based on level of alcohol use.

	AlcoholUse
Abstain	0.13
Heavy	0.06
Light	0.33
Moderate	0.47

2. What proportion of students in the high stress group report high alcohol use?

There are 5 students who reported high stress level and heavy alcohol use. There were a total of 56 students who were in the high stress group. Thus, the proportion of students in the high

	Abstain	Heavy	Light	Moderate
high	0.143	0.089	0.357	0.411

stress group who reported high alcohol use is $\frac{5}{56} = 0.089$.

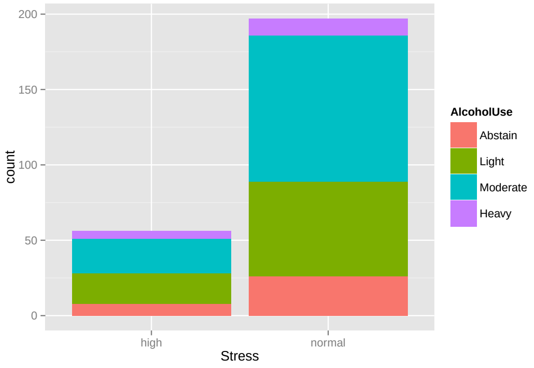
We could have also created a table in r to show us this proportion.
Here is the r code that I used to create this table.

```
newdata <- subset(SleepStudy, SleepStudy$Stress=='high')
highprop <- with(newdata,table(Stress,AlcoholUse))
```

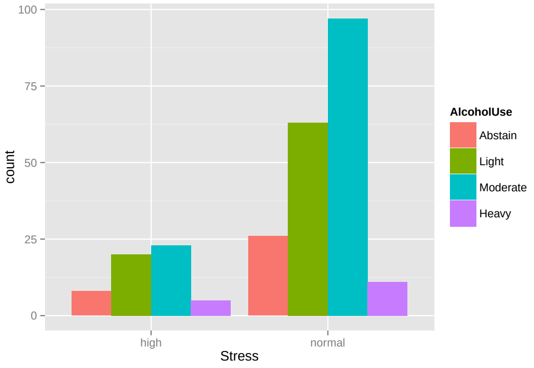
3. Display the data from this table in a bar graph that effectively compares the distribution of alcohol use for each stress group. Write the R code you used to create this graph.

Below are two bar graphs which show how the distribution of alcohol use compares for each stress group.

Stacked Bar Graph of Students Stress Levels
Showing Distribution of Alcohol Use



Side by Side Bar Graph of Students Stress Levels
Showing Distribution of Alcohol Use



This is the code that I used to create these graphs.

```
foo = rep(0, nrow(SleepStudy))
foo[with(SleepStudy, AlcoholUse == "Abstain")] = 1
foo[with(SleepStudy, AlcoholUse == "Light")] = 2
foo[with(SleepStudy, AlcoholUse == "Moderate")] = 3
foo[with(SleepStudy, AlcoholUse == "Heavy")] = 4
foo
SleepStudy$AlcoholUse = with(SleepStudy, reorder(AlcoholUse, foo))
rm(foo)

library(ggplot2)
ggplot(SleepStudy,aes(x=Stress,fill=AlcoholUse))+geom_bar()
ggplot(SleepStudy, aes(x=Stress, fill=AlcoholUse)) + geom_bar(position ="dodge")
```

4. Describe the patterns you see in the data.

From the graphs, we are able to see that there are a much larger number of students reporting normal stress levels than high stress levels. However, it appears that the distribution of alcohol use is the same for both groups. For both students with normal and high stress levels, the alcohol use levels ranked from lowest number of counts to highest number of counts are heavy, abstain, light, and moderate.

R problem 2 Examine the variables Drinks and LarkOwl.

1. Find the mean, median, and standard deviation of the number of alcoholic drinks per week for the entire group of students.

Below are the mean, median, and standard deviation of the number of alcoholic drinks per week for the entire group of students.

statistic	value
mean	5.56917
median	5
standard deviation	4.095119

The following is the r code that I used to produce these results.

```
data(SleepStudy)
str(SleepStudy)
attach(SleepStudy)
mean(Drinks)
median(Drinks)
sd(Drinks)
```

2. Do the same as (1), but separately for each group of students that classify themselves as early risers (larks), night owls, or neither.

The table below contains the mean, median, and standard deviation of number of alcoholic drinks per week divided by whether or not the student classified themselves as a lark, an owl, or neither.

Statistics	Lark	Owl	Neither
mean	4.59	6.39	5.571
median	3	7	5
standard deviation	4.20	4.38	3.95

This is the r code that I used to solve the problem.

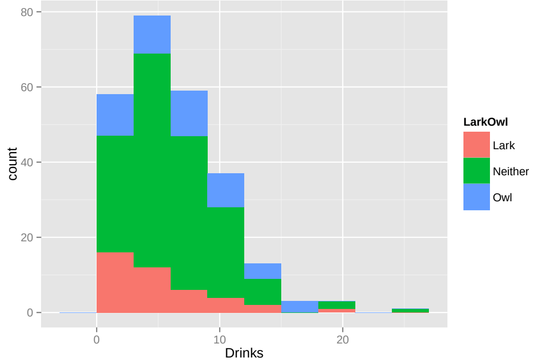
```
OwlData <- subset(SleepStudy, SleepStudy$LarkOwl==’Owl’)
mean(OwlData$Drinks)
median(OwlData$Drinks)
sd(OwlData$Drinks)
NeitherData <- subset(SleepStudy, SleepStudy$LarkOwl==’Neither’)
mean(NeitherData$Drinks)
median(NeitherData$Drinks)
```

```
sd(NeitherData$Drinks)
LarkData <- subset(SleepStudy, SleepStudy$LarkOwl=='Lark')
mean(LarkData$Drinks)
median(LarkData$Drinks)
sd(LarkData$Drinks)
```

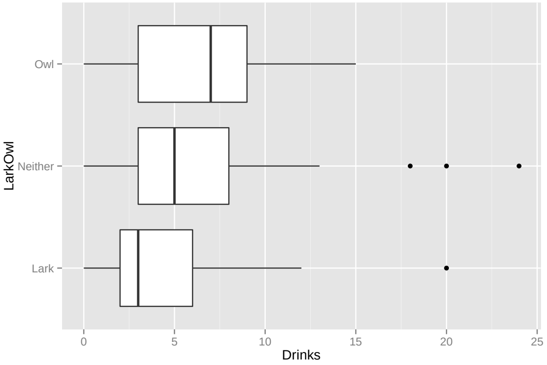
3. Create an effective display of the Drinks variable that shows how the distribution of number of drinks per week varies among larks, night owls, and neither.

Below are a stacked bar graph and box plots which show the how the distribution of number of drinks per week varies among larks, night owls, and neither.

Stacked Bar Graph of Number of Drinks Consumed
By Students Showing Distribution of LarkOwls



Box plots of Number of Drinks Consumed
Divided by NightOwl Categories



4. Describe the patterns you see in the data.

From the bar graph, we can see that a larger number of students classify themselves as neither. Additionally, for all groups, the number of students decreases as the number of drinks increase. From the box plots, we see that the median number of drinks per week is the highest for the owls and lowest for the larks. In general, it appears that the students who classify themselves as owls tend to drink more than the other two categories. For example, the first quantile of the owls is approximately same as the median of the larks. We also see that all three categories have people who do not drink at all, and the owls have the largest range. However, there are a few outliers on the end of high drinks for both the neither group and the larks.