Assignment 1—Assumptions and T-Tests.

Yudhajit Ain

Student ID: 30182745

University of Calgary

PSYC 615—Lab

TAs: Benjamin Moon & Christopher Davie

Assignment 1—Assumptions and T-Tests.

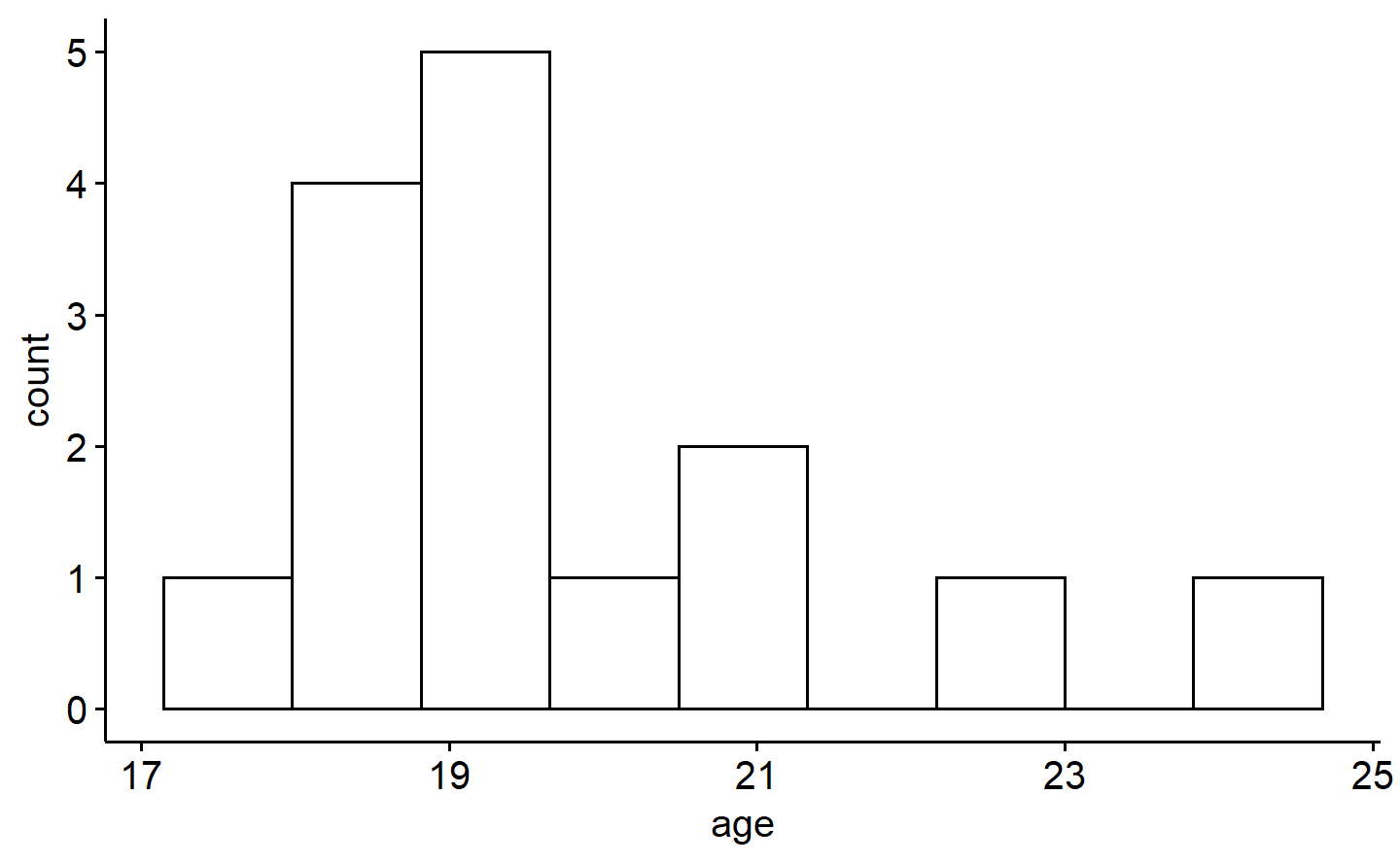
# Question 1

*Because it’s fun (or rather, educational), you will be experimenting with making your own datasets. Please include your final data points under each question; no need for an APA formatted table, listing them is fine!*

## Question 1a

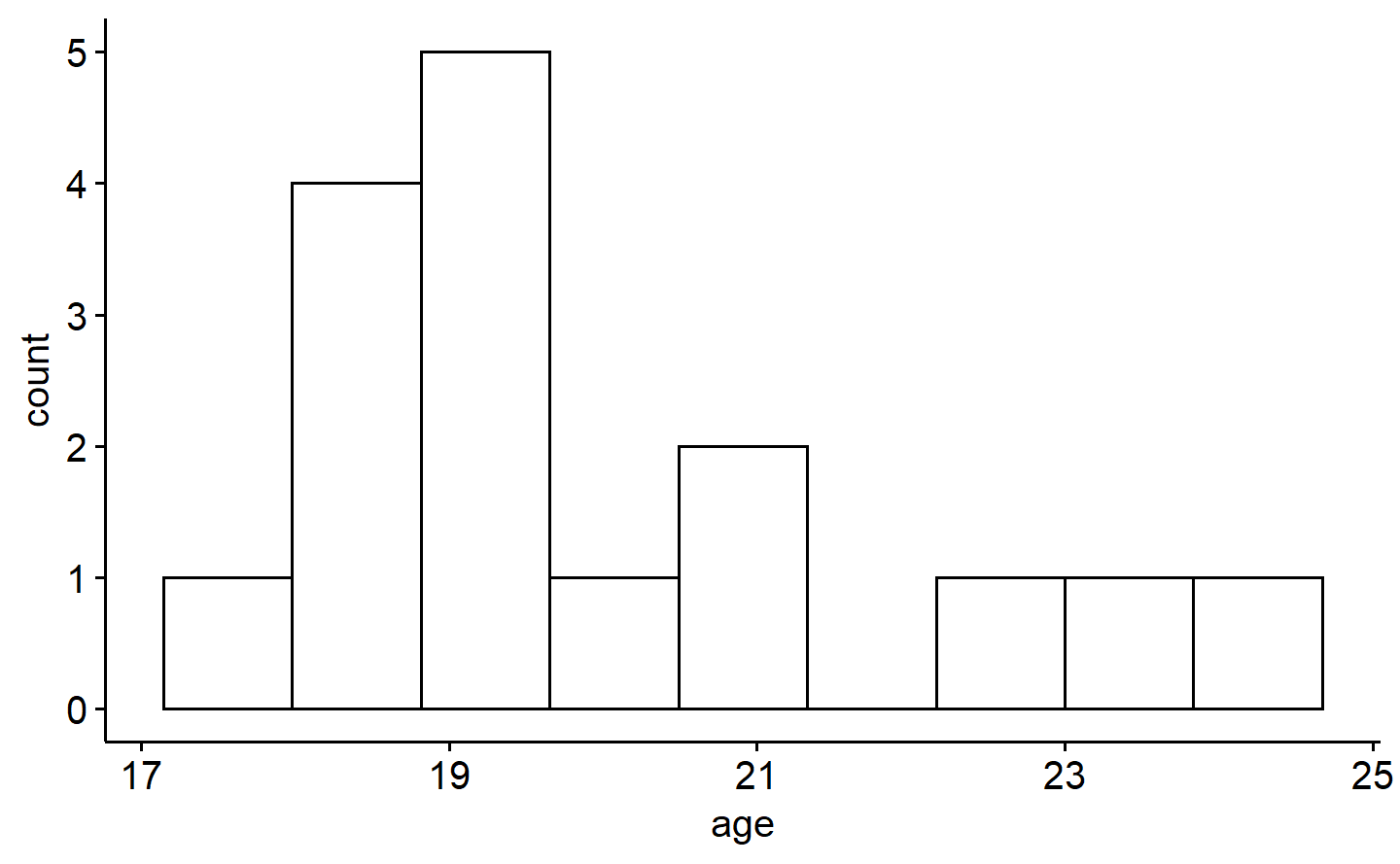
*Make a positively skewed distributed (i.e., NOT normal) set of 18 participants on a fictitious variable (think of a variable that makes sense with the numbers). Report your variable’s mean and SD. Report and interpret the skewness numerical measure and an appropriate statistical test of normality. (3 marks)*

#> [1] "mean=19.7998920811782"  
#> [1] "median=19.4246248250152"  
#> [1] "SD=1.72122202885716"  
#> Skewness | SE  
#> ----------------  
#> 1.222 | 0.520  
#>   
#> Shapiro-Wilk normality test  
#>   
#> data: data\_1a$age  
#> W = 0.90644, p-value = 0.1194



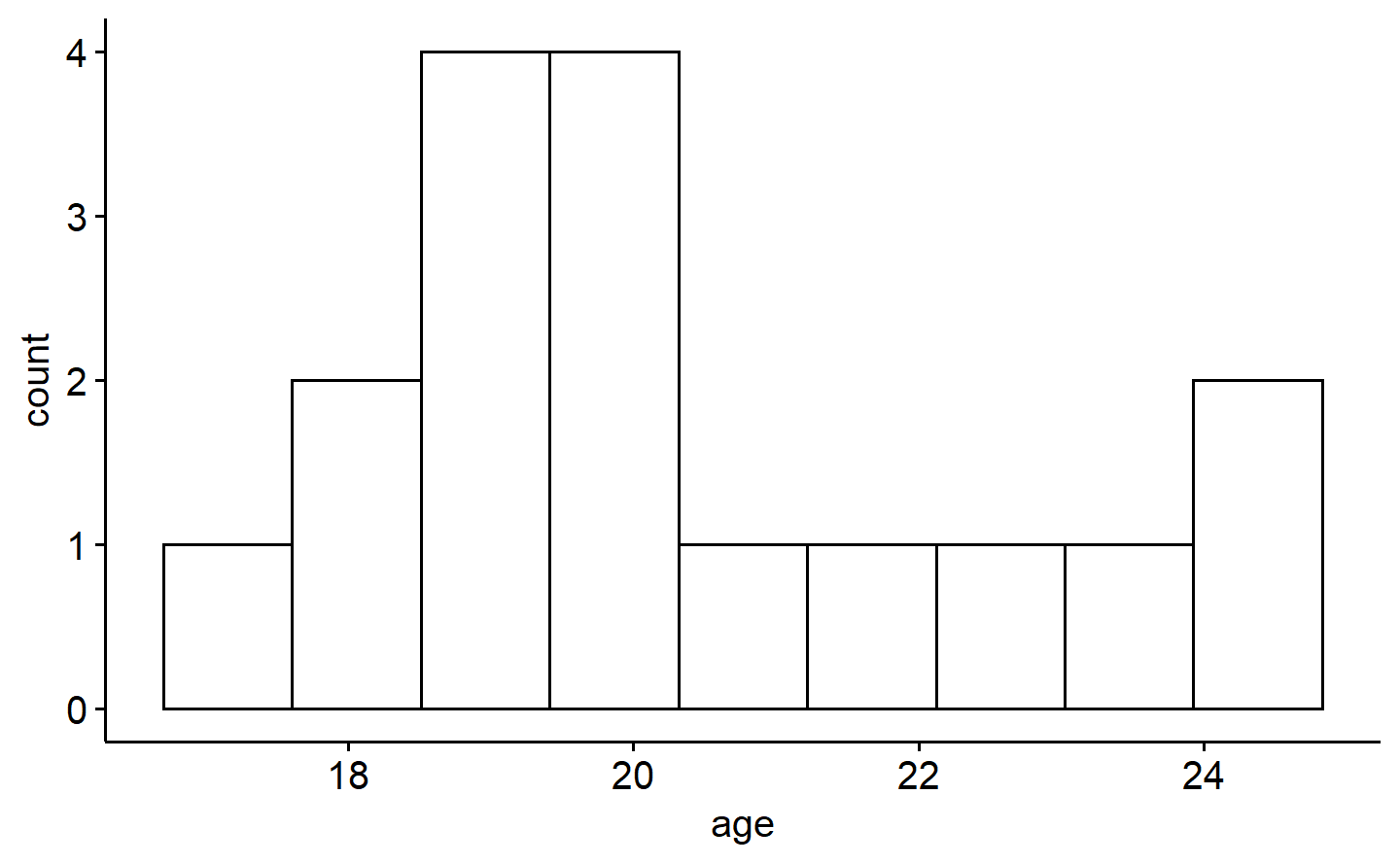
*Figure* *1.*  Distribution of 15 samples

#> Skewness | SE  
#> ----------------  
#> 0.977 | 0.510  
#>   
#> Shapiro-Wilk normality test  
#>   
#> data: data\_1a\_added$age  
#> W = 0.90365, p-value = 0.09196



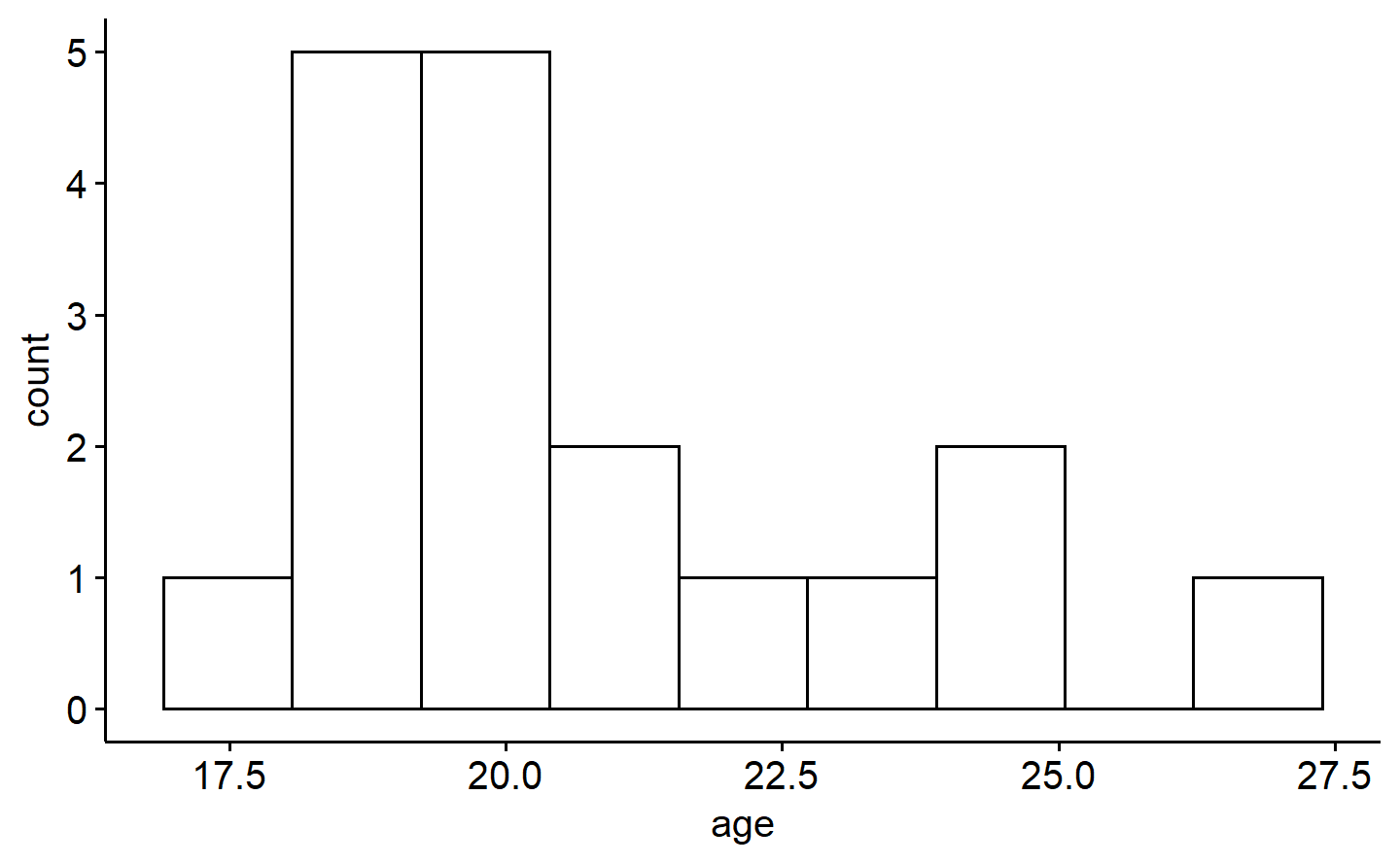
*Figure* *2.*  Distribution of 16 samples

#> Skewness | SE  
#> ----------------  
#> 0.882 | 0.500  
#>   
#> Shapiro-Wilk normality test  
#>   
#> data: data\_1a\_added$age  
#> W = 0.89481, p-value = 0.05565



*Figure* *3.*  Distribution of 17 samples

#> Skewness | SE  
#> ----------------  
#> 1.028 | 0.491  
#>   
#> Shapiro-Wilk normality test  
#>   
#> data: data\_1a\_added$age  
#> W = 0.89087, p-value = 0.03994  
#> participant\_number age  
#> 1 1 18.74922  
#> 2 2 22.38850  
#> 3 3 20.06175  
#> 4 4 19.58283  
#> 5 5 19.26064  
#> 6 6 20.83675  
#> 7 7 19.42462  
#> 8 8 18.74892  
#> 9 9 21.22067  
#> 10 10 18.19683  
#> 11 11 17.37652  
#> 12 12 18.50157  
#> 13 13 19.48071  
#> 14 14 24.06706  
#> 15 15 19.10177  
#> 16 16 23.50000  
#> 17 17 24.60000  
#> 18 18 26.70000  
#> [1] "mean=20.655465623204"  
#> [1] "median=19.5317735539798"  
#> [1] "SD=2.5741210745179"



*Figure* *4.*  Distribution of 18 samples

## ANSWER

We first make a normal distribution of 15 samples repeatedly, till we get a distribution with positive skewness, but not significantly non-normal according to the Shapiro-Wilk’s test ((15)=0.91, =.12). This is saved as “Q\_1a\_data\_15.csv”. Then, we add three large values closer to the right tail of the distribution to give it a significantly non-normal positive skew. This is saved as Q\_1a\_data.csv, and displayed on the console. For this final simulated sample of 18 graduate students (participants), mean age was =20.65 years, and the standard deviation () was 2.57 years.

The skewness value for age is 1.03, implying positive skewness, with mean > median. Shapiro-Wilk’s test ((18)=0.89, =.04) confirms the distribution is non-normal by rejecting the null hypothesis at <.05.

## Question 1b

*Now, create a normally distributed dataset (n = 18), with the same mean and standard deviation from 1a. (as close as possible, and cannot be more than 10% - 20% difference from either the mean or the SD). Report your mean and SD. Report and interpret the skewness numerical measure and an appropriate statistical test of normality. (3 marks)*

#> participant\_number age  
#> 1 1 22.10024  
#> 2 2 21.22476  
#> 3 3 18.62203  
#> 4 4 24.16933  
#> 5 5 21.67254  
#> 6 6 12.68929  
#> 7 7 21.45602  
#> 8 8 20.74622  
#> 9 9 20.32710  
#> 10 10 15.31316  
#> 11 11 22.58002  
#> 12 12 19.80146  
#> 13 13 18.87518  
#> 14 14 17.31327  
#> 15 15 23.59796  
#> 16 16 22.82637  
#> 17 17 21.12358  
#> 18 18 22.81051  
#> [1] "mean= 20.4027236298585"  
#> [1] "median= 21.1741681356096"  
#> [1] "SD=2.95597733460294"  
#> Skewness | SE  
#> ----------------  
#> -1.271 | 0.491  
#>   
#> Shapiro-Wilk normality test  
#>   
#> data: data\_1b$age  
#> W = 0.90167, p-value = 0.06145

## ANSWER

For this new sample of 18 graduate students (participants), mean age was =20.40 years, and the standard deviation () was 2.95 years. The dataset has been saved as Q\_1b\_data.csv, and displayed on the console

The skewness value for age is -1.27, implying negative skewness (more data points towards the left tail of the distribution than normal, with mean > median). Shapiro-Wilk’s test ((18)=0.90, =.061), however, fails to reject the null-hypothesis of normal distribution at <0.05.

# Question 2

*From the data directory, please import assignment-1\_q2.csv into R.*

*Background: Imagine you are a TA for an undergrad statistics course (or maybe you are now!) and your students just finished their first statistics exam of the year. You, a clever statistical student yourself, decided to record the following variables from the students: age, gender (1 = female, 2 = male), how many hours they study per week, their love of statistics (1 to 10 scale), and their heart rate both before the exam and after the exam (normal range is 60-100 for adults). You are now ready to investigate!*

## Question 2a

*Briefly describe why independence is important for analyses that assume it (1 mark). Give one potential limitation of why this sample (i.e., the students) might not be independent? (1 mark)*

Answers here.

Independence of samples is important for most statistical analyses since it preserves “randomness” of the variable. Independent sampling is also key to making sure our distribution falls in line with the central limit theorem, which is itself the basis of most mathematical formulae employed in statistical hypothesis testing

The observations in this current sample may not be independent since students who choose to take up an undergraduate degree (and a statistics course) may already be biased towards higher study-hours or love-for-statistics measures.

## Question 2b

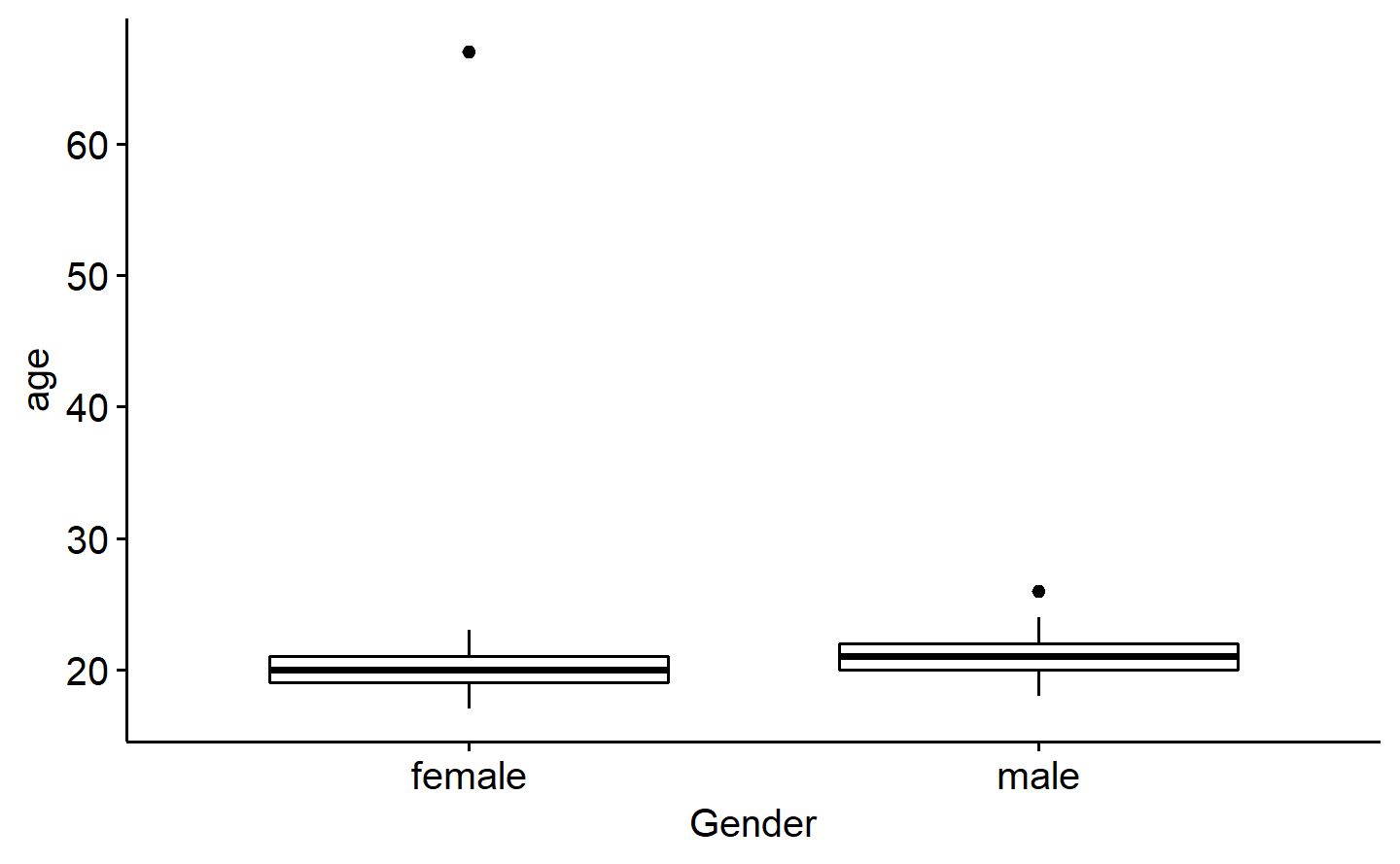
*First things first: explore your data set as a whole. Do you see any “weirdness” in any of the continuous variables (i.e., age, study, love, heart rates)? Explain what the weirdness is (including participant #s) and justify why you think it is weird for each instance (i.e., with visuals, statistics, etc.). With this information, would you remove any participants/data points from the future analyses? Why or why not? (5 marks)*

*Hint: Read the variable descriptions above.*

*Note 1: Please only include the relevant information here (we don’t need a novel). Just point out the glaring issues (you’ll see them we hope). Please don’t transform the data or anything drastic like that if there is skewness or something like that!*

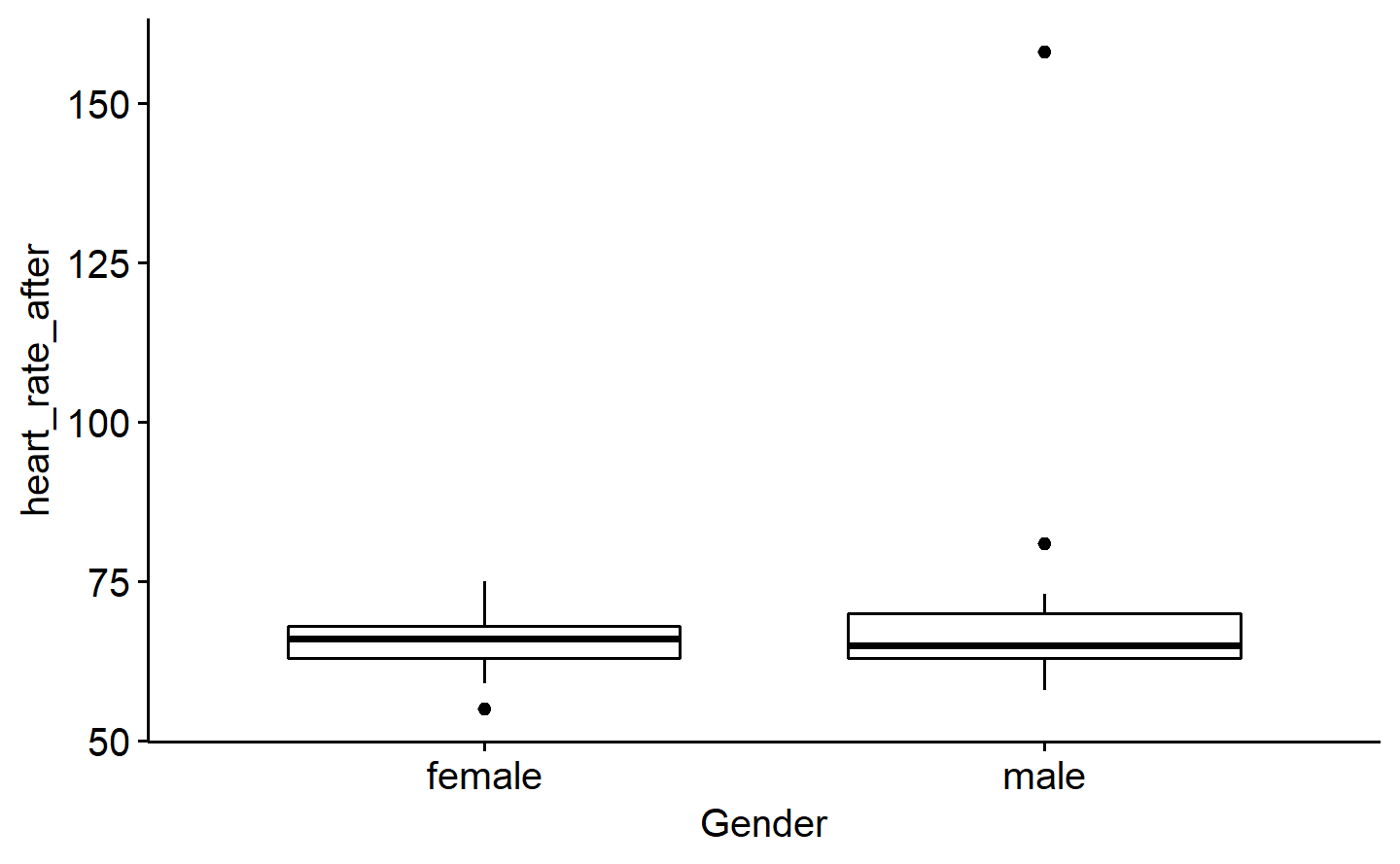
*Note 2. If you say that certain participants/data should be removed from the analyses, please do that for the remainder of the questions!*

#> [1] "SD= 6.20360747974208"  
#> [1] "mean= 21.3"



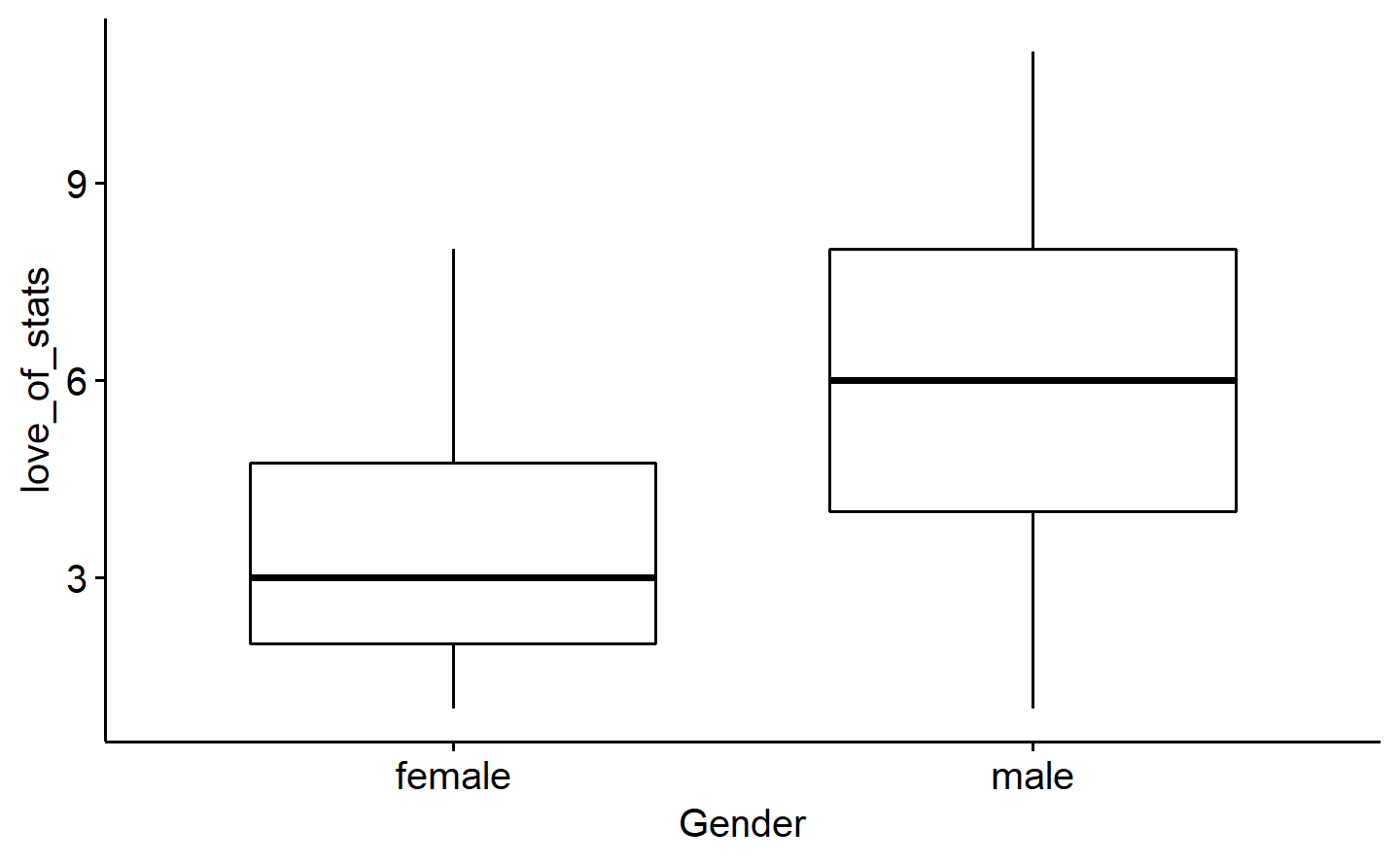
*Figure* *5.*  Boxplot of age grouped by gender

#> [1] "SD= 13.0196542520895"  
#> [1] "mean= 67.728813559322"



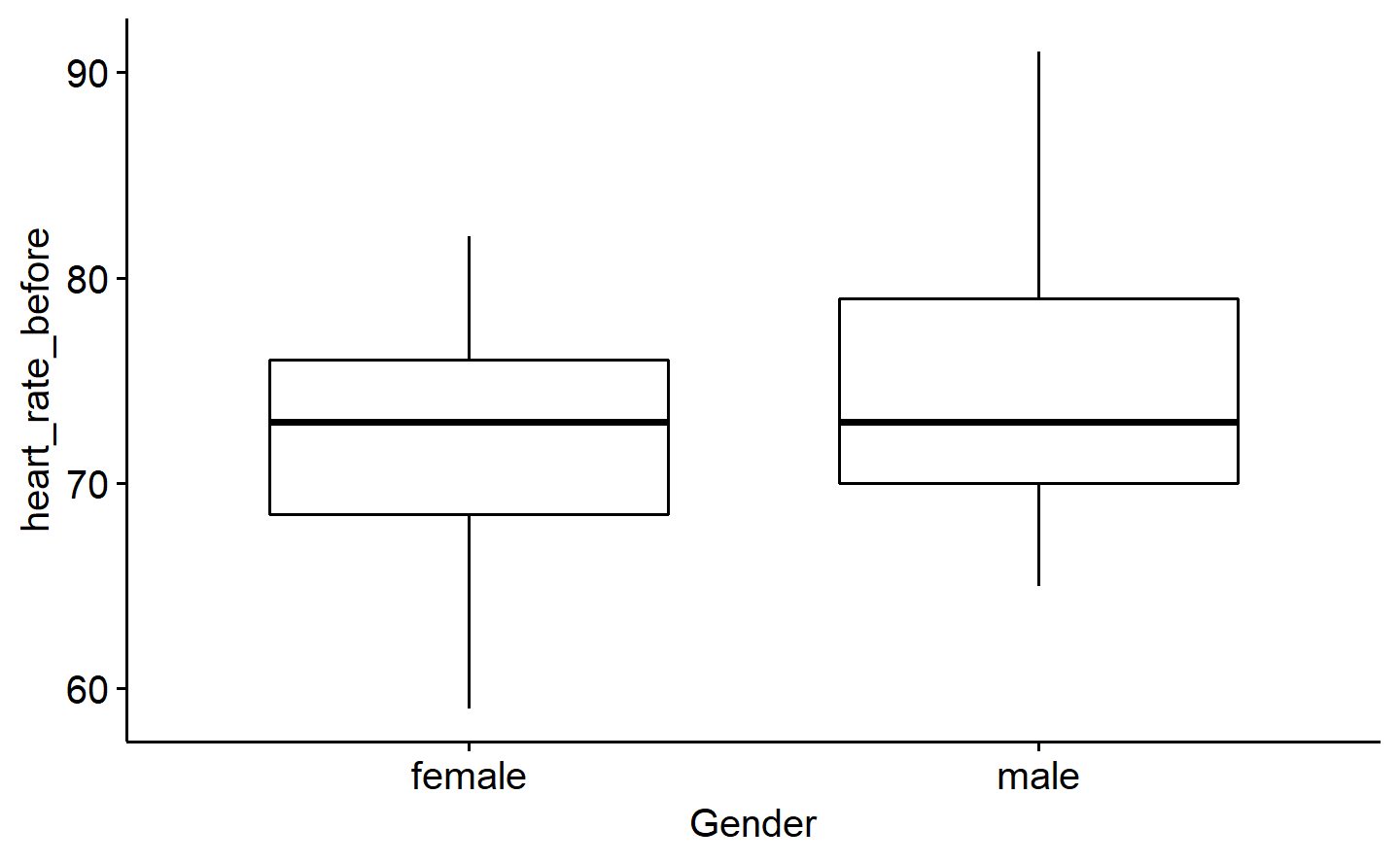
*Figure* *6.*  Boxplot of heart\_rate\_after grouped by gender

#> [1] "SD= 2.76417317791989"  
#> [1] "mean= 4.79310344827586"



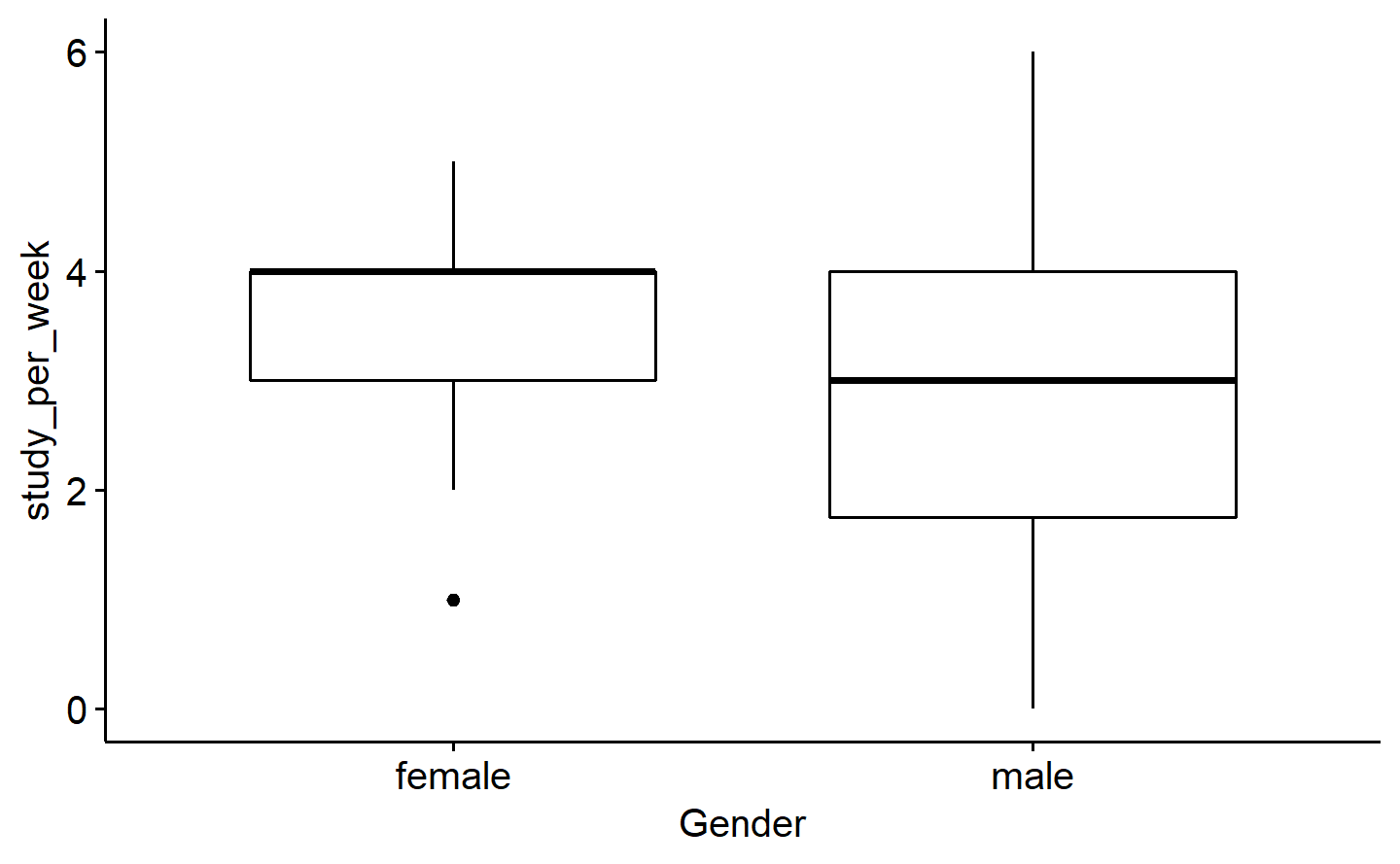
*Figure* *7.*  Boxplot of love\_Of\_stats grouped by gender

#> [1] "SD= 6.14999545049729"  
#> [1] "mean= 73.6034482758621"



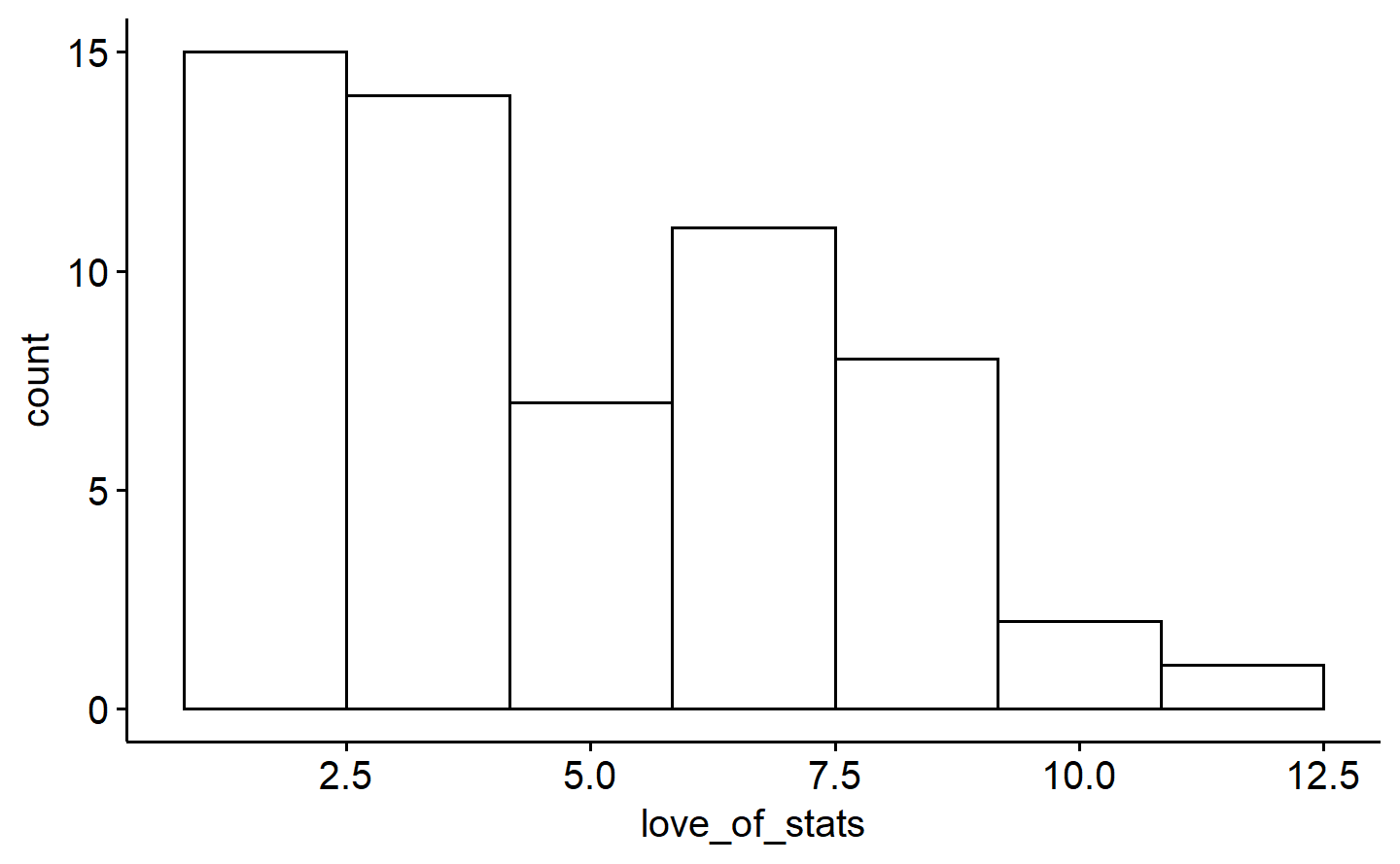
*Figure* *8.*  Boxplot of heart\_rate\_before grouped by gender

#> [1] "SD= 1.40703014086475"  
#> [1] "mean= 3.05172413793103"



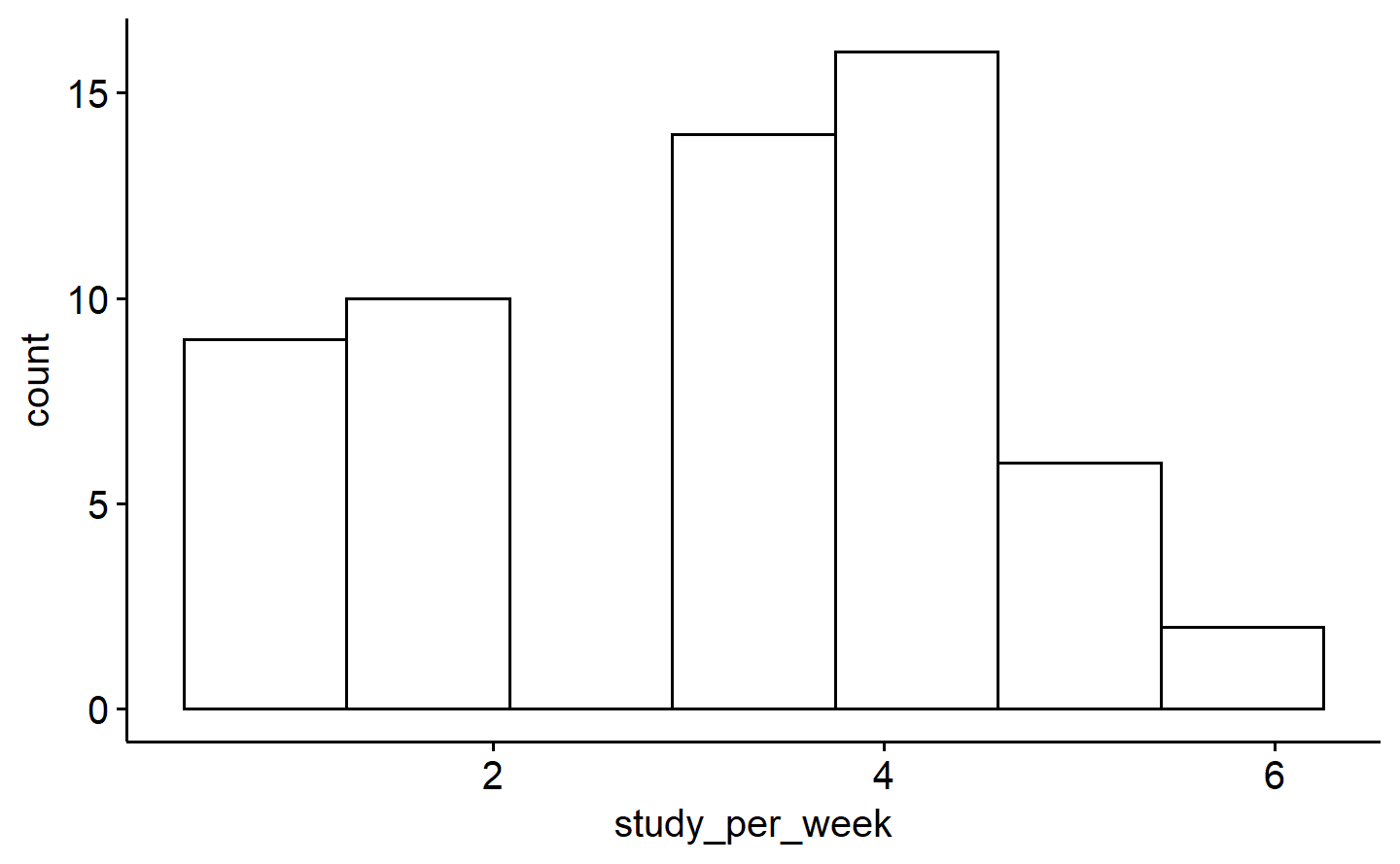
*Figure* *9.*  Boxplot of study\_per\_week grouped by gender

#> Skewness | SE  
#> ----------------  
#> 0.282 | 0.308  
#>   
#> Shapiro-Wilk normality test  
#>   
#> data: exams\_filtered$love\_of\_stats  
#> W = 0.94162, p-value = 0.008338



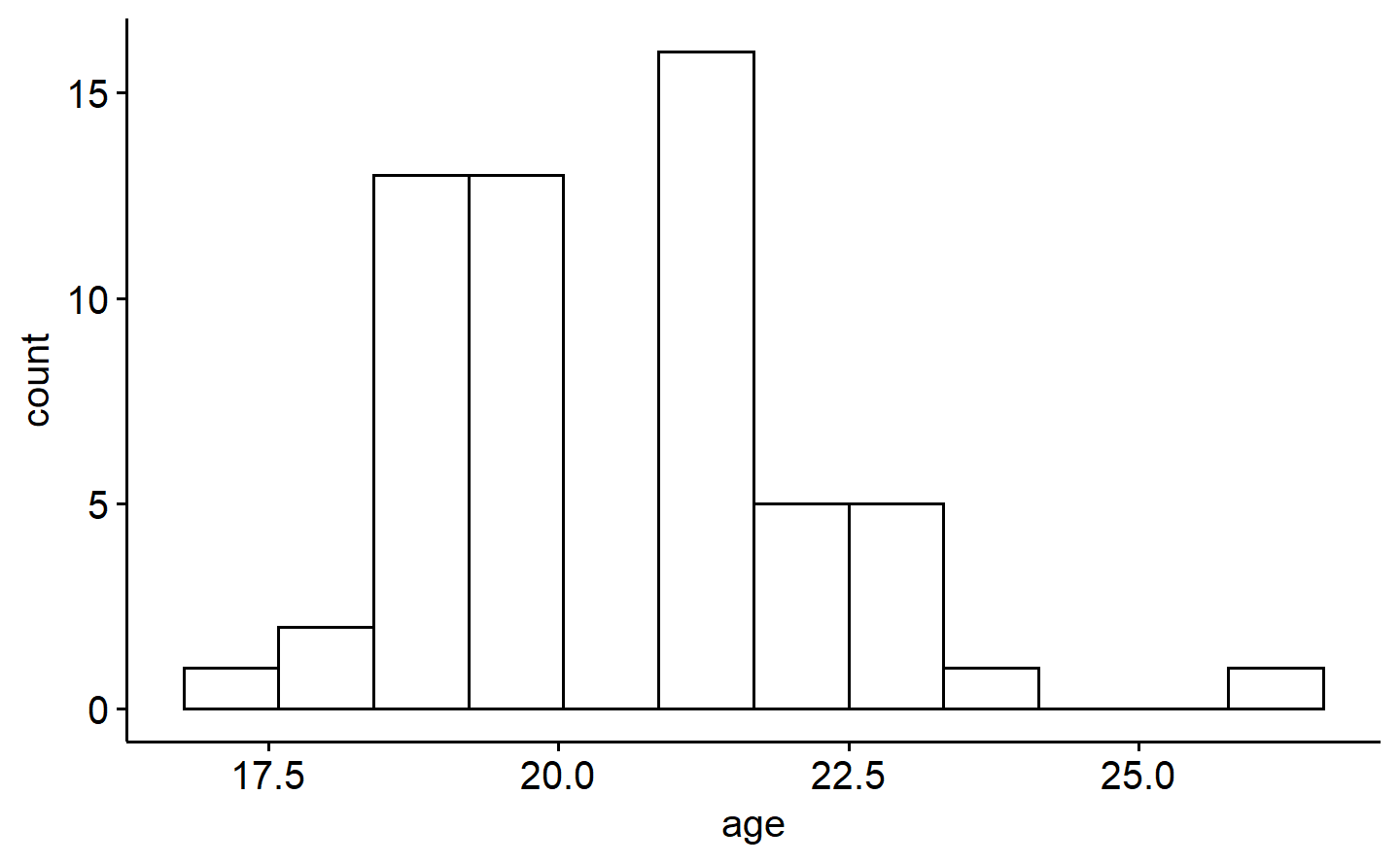
*Figure* *10.*  Distribution of love\_of\_stats

#> Skewness | SE  
#> ----------------  
#> 0.024 | 0.308  
#>   
#> Shapiro-Wilk normality test  
#>   
#> data: exams\_filtered$study\_per\_week  
#> W = 0.92846, p-value = 0.002324



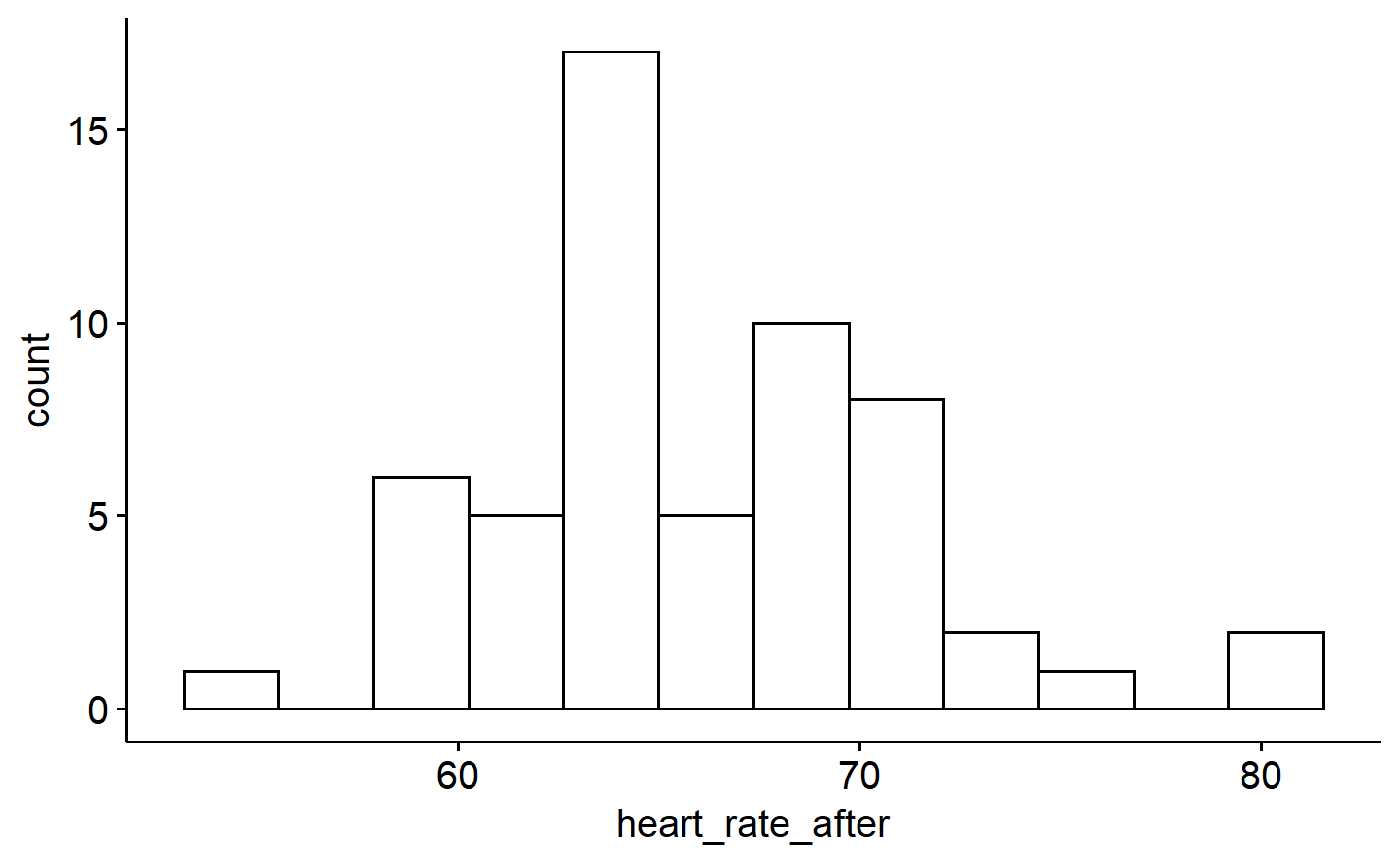
*Figure* *11.*  Distribution of study\_per\_week

#> Skewness | SE  
#> ----------------  
#> 0.735 | 0.308  
#>   
#> Shapiro-Wilk normality test  
#>   
#> data: exams\_filtered$age  
#> W = 0.93426, p-value = 0.004038



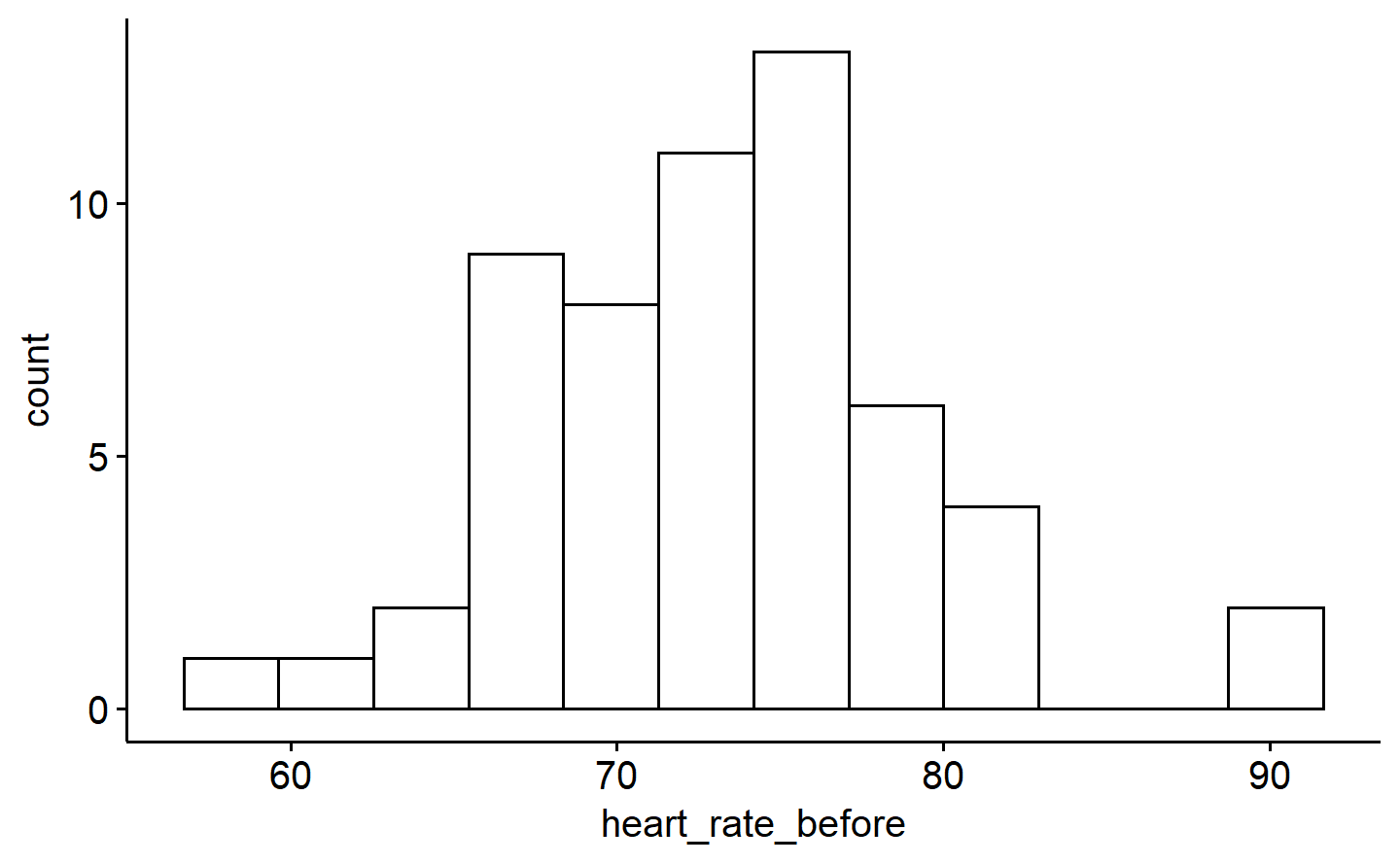
*Figure* *12.*  Distribution of age

#> Skewness | SE  
#> ----------------  
#> 0.565 | 0.308  
#>   
#> Shapiro-Wilk normality test  
#>   
#> data: exams\_filtered$heart\_rate\_after  
#> W = 0.96926, p-value = 0.155



*Figure* *13.*  Distribution of heart\_rate\_after

#> Skewness | SE  
#> ----------------  
#> 0.304 | 0.308  
#>   
#> Shapiro-Wilk normality test  
#>   
#> data: exams\_filtered$heart\_rate\_before  
#> W = 0.98089, p-value = 0.5026



*Figure* *14.*  Distribution of heart\_rate\_before

Answers here.

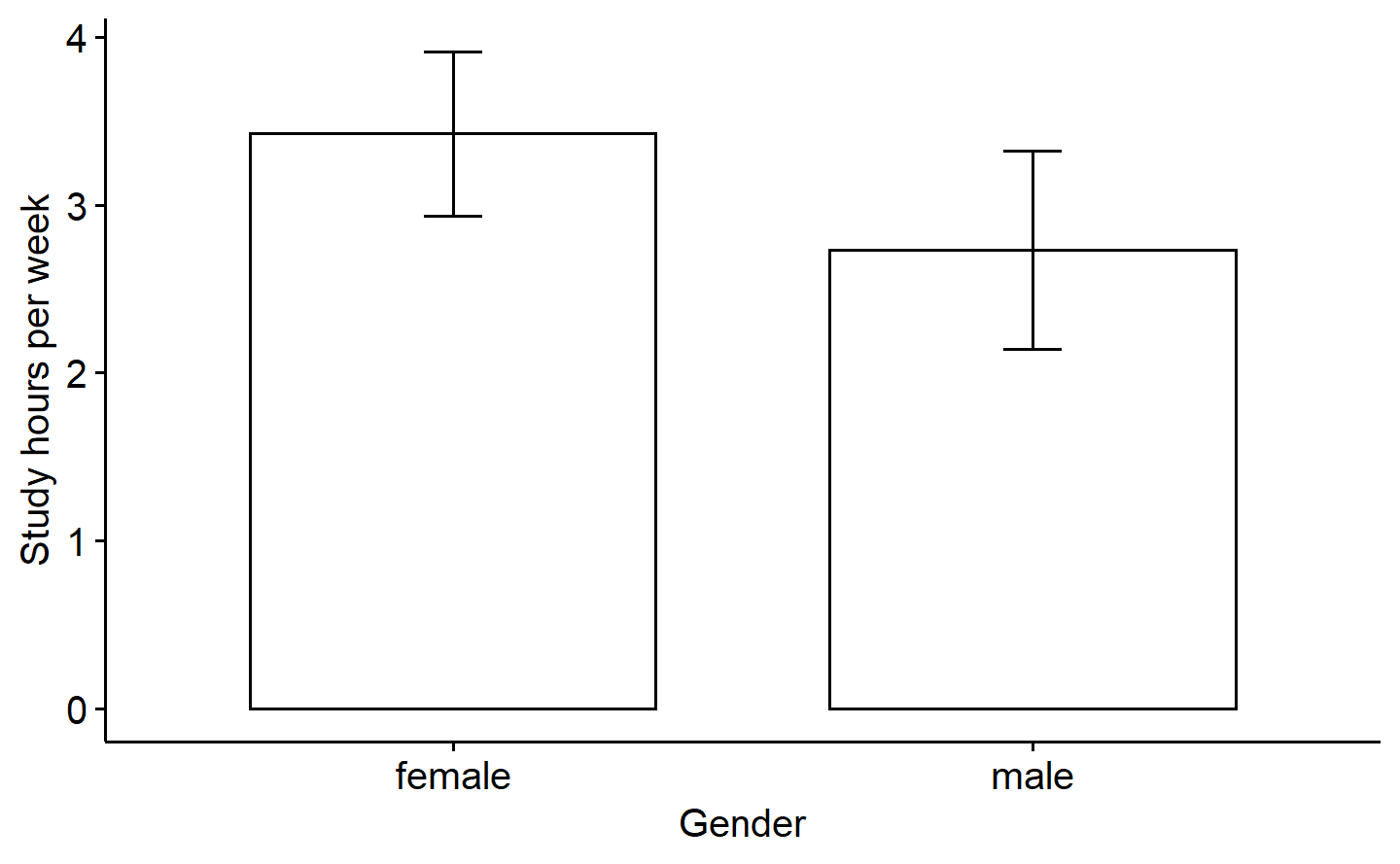
Outliers >3 s away from the mean were identified visually from box plots of age and heart\_rate\_after. Student 40 had an extremely high heart rate after the exam. Student 50 was much more older than the other students in the course. These two students were filtered out. From looking at the histograms of all the variables thereafter, Student 9 was also filtered out of the dataset because the love\_of\_stats variable had a value of 11 for them, and it cannot exceed 10.

Even after filtering, all three variables age ((57)=0.93, =.004),love\_of\_stats ((57)=0.94, =.008) and study\_per\_week((57)=0.93, =0.002) seem to be significantly non-normal, at the <0.05 level. Histograms of their distributions showed their skewness. heart\_rate\_before and heart\_rate\_after, however, were both normally distributed, and their Shapiro-Wilk’s tests did not reach significance.

## Question 2c

*You are interested in whether males are significantly lower in the number of hours they study each week compared to females (Hypothesis 1). Write an APA formatted results section exploring Hypothesis 1. (5 marks) This includes stating the appropriate test (no justification needed), quickly reporting on assumptions (we assume independence from here on so no need for that, and only report normality after removing outliers – no need to repeat normality tests from b.) and reporting the findings (provide and reference an APA formatted figure and reference whether the hypothesis was supported or not). Hint: When reporting findings, you need to include practical significance!*

#> [1] "mean\_hours 3.42307692307692"  
#> [1] "sd\_hours 1.20575542870275"  
#>   
#> Shapiro-Wilk normality test  
#>   
#> data: exams\_filtered\_female$study\_per\_week  
#> W = 0.90265, p-value = 0.01795  
#> [1] "mean\_hours 2.83870967741935"  
#> [1] "sd\_hours 1.43983272624631"  
#>   
#> Shapiro-Wilk normality test  
#>   
#> data: exams\_filtered\_male$study\_per\_week  
#> W = 0.91063, p-value = 0.01342  
#> [1] "mean\_hours 2.73076923076923"  
#> [1] "sd\_hours 1.45760604575289"  
#> Levene's Test for Homogeneity of Variance (center = mean)  
#> Df F value Pr(>F)  
#> group 1 0.5218 0.4734  
#> 50  
#>   
#> Two Sample t-test  
#>   
#> data: exams\_filtered\_female$study\_per\_week and exams\_filtered\_male\_equal$study\_per\_week  
#> t = 1.8661, df = 50, p-value = 0.0679  
#> alternative hypothesis: true difference in means is not equal to 0  
#> 95 percent confidence interval:  
#> -0.05284645 1.43746184  
#> sample estimates:  
#> mean of x mean of y   
#> 3.423077 2.730769  
#> Cohen's d | 95% CI  
#> -------------------------  
#> 0.52 | [-0.04, 1.07]  
#>   
#> - Estimated using pooled SD.



*Figure* *15.*  Bar-plot of study hours per week grouped by gender

Answers here.

To test the hypothesis, we need to perform an independent samples -test. First, we test for the normality assumption in females and males by running the Shapiro-Wilk’s test at a -critical of 0.05. Study hours per week were distributed non-normally in both females ((26)=0.90, =.018) and males ((31)=0.91, =.013).

We know that while the independent t-test is generally robust to violations of normality, unequal sample sizes, like here, can lead to an inflation of Type I errors. Therefore, we downsample the male dataset from =31 to =26 (same as females) using resampling without replacement, and save that dataset as Q\_2c\_male\_equal.csv.

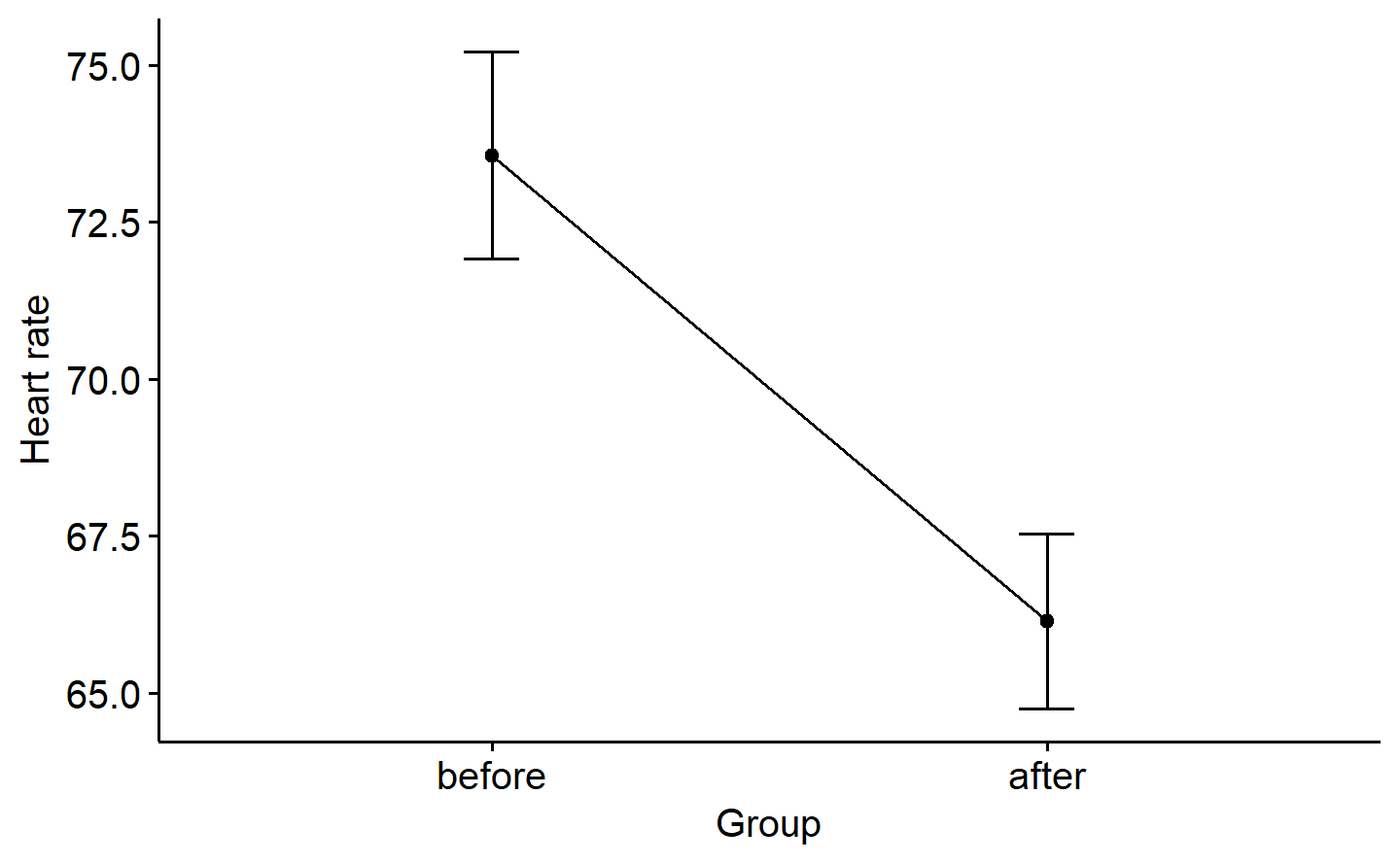
We then checked for homogeneity of variance between the equal-sized male and female groups using the Levene’s test. The variance across the groups can be assumed to be equal ((1,50)=0.52, =.473).

Finally, we run the independent samples -test. Results suggest that females (=3.42, =1.20) do not study significantly more number of hours per week than males (=2.73, =1.46), (50)=1.87, =.068, Cohen’s =0.52. The medium effect size, and a -value very close to the threshold however suggests that using a larger sample size may lend sufficient power to the -test to get a significant positive effect of being female on studying more hours per week.

## Question 2d

*You are also interested in whether there is a statistically significant increase in heart rates from before to after the exam for all students (Hypothesis 2). Write an APA formatted results section (with the same requirements from the previous question) for Hypothesis 2. (5 marks)*

#>   
#> Shapiro-Wilk normality test  
#>   
#> data: exams\_filtered$heart\_rate\_after  
#> W = 0.96926, p-value = 0.155  
#> [1] 66.14035  
#>   
#> Shapiro-Wilk normality test  
#>   
#> data: exams\_filtered$heart\_rate\_before  
#> W = 0.98089, p-value = 0.5026  
#> [1] 73.5614  
#>   
#> Paired t-test  
#>   
#> data: heart\_rate by group  
#> t = -16.706, df = 56, p-value < 2.2e-16  
#> alternative hypothesis: true mean difference is not equal to 0  
#> 95 percent confidence interval:  
#> -8.310941 -6.531164  
#> sample estimates:  
#> mean difference   
#> -7.421053  
#> Cohen's d | 95% CI  
#> --------------------------  
#> -2.21 | [-2.69, -1.73]



*Figure* *16.*  Line-graph of heart rates before and after the exam

Heart rates both before ((57)=0.98, =.503) and after ((57)=0.97, =.155) the exam were normally distributed. A paired samples -test reveals that there is a significant decrease in the heart rate of students from before the exam (=73.56, =6.20) to after the exam (=66.14, =5.24), (56)=-16.71, <.001. The effect size, suggested by a Cohen’s =-2.21, is very large.

## Question 2e

*Another TA brags to you that they collected similar data when they were teaching the class and their students’ love of statistics had a mean of 5.17. They claim that your class’ love was significantly lower (i.e., you aren’t as awesome as them). Does this TA have the right to brag, or should they spend more time studying statistics themselves? Choose an appropriate test and report the findings that justify your response to this TA (no APA results section needed, including no figure, assumptions, or effect sizes needed). (2 marks)*

#>   
#> One Sample t-test  
#>   
#> data: exams\_filtered$love\_of\_stats  
#> t = -1.3787, df = 56, p-value = 0.1735  
#> alternative hypothesis: true mean is not equal to 5.17  
#> 95 percent confidence interval:  
#> 3.978344 5.390077  
#> sample estimates:  
#> mean of x   
#> 4.684211

Answers here.

While the mean of my class’s love of statistics was lower than the other TA, the one-sample -test shows that the effect is not significant ((56)=-1.38, =.174). Moreover, we assume that the distribution of love\_of\_stats is normal, although we have seen before in Q2b that it is significantly skewed. Therefore, the -test is also not completely reliable.

As such, it seems the TA needs to brush up the basics of his own statistics.

# Bonus

*(.5 marks; not exceed 100% on assignment) When the two groups for an independent t-test do not have equal variance, what specific value is first adjusted to compensate for this in the analysis?*

Answers here.

When the two groups in an independent t-test do not have equal variance, the degrees of freedom () is adjusted according to the variances of the two groups. This leads to a consequent change in the critical -value for a particular level.