Chapter 2 - Summarizing Data

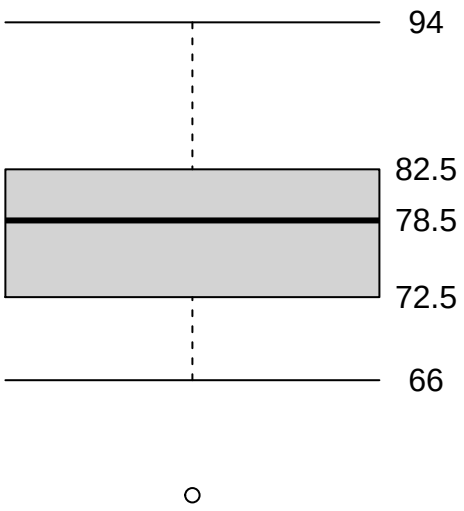
Stats scores. (2.33, p. 78) Below are the final exam scores of twenty introductory statistics students.

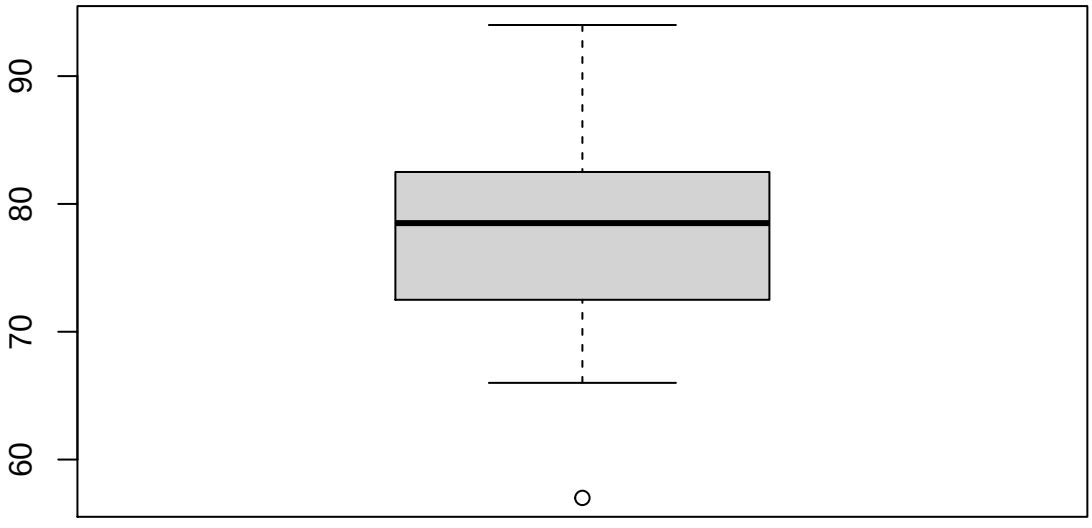
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

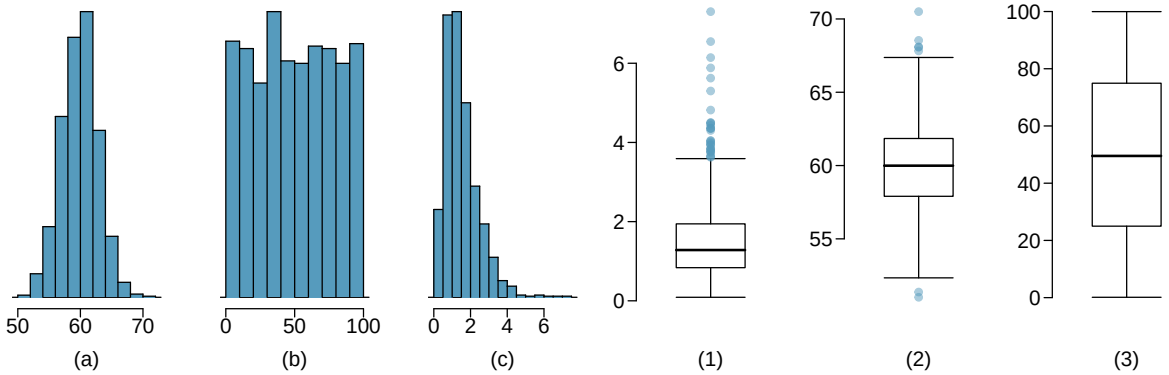
Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

Answer 1 **********


94
82.5
78.5
72.5
66


90
80
70
60

Mix-and-match. (2.10, p. 57) Describe the distribution in the histograms below and match them to the box plots.



********** ANSWER2 - a) is symmetric distribution and this matches with 2. b is roughly uniform distribution and it matches with with 3 c) is right skewed distribution and it matches with 1 **********

Distributions and appropriate statistics, Part II. (2.16, p. 59) For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(a) Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000. ********** ***Answer:The distribution is expected to be right skewed. There are more expensive house which clearly indicate this.Median and variability will be better represented by IQR is better in scenarios because unlike mean they would not give the false impression of the average prices in the location.*** **********

(b) Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000. ********** ***Answer: we can see the uniformity/symmetric distribution here and hence we cannot decide the left or right skewness. For uniformity we can use mean and standard deviation*** **********

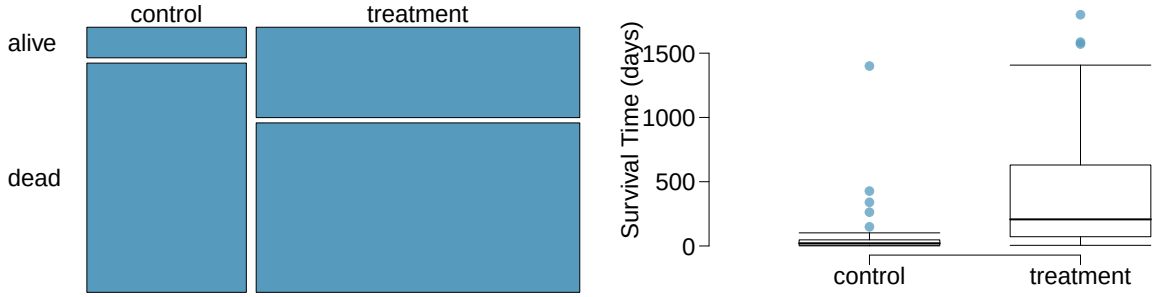
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively. ********** ***Answer: This would be represented by right skewness.Most of the data will show the increase as they cross 21 but then it will show the decrease with age increasing. This can be best represented by median and IQR .*** **********

(d) Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than the all other employees.

***Answer:

###HEART TRANSPLANT EXERCISE###

Heart transplants. (2.26, p. 76) The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable *transplant* indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Of the 34 patients in the control group, 30 died. Of the 69 people in the treatment group, 45 died. Another variable called *survived* was used to indicate whether or not the patient was alive at the end of the study.



(a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning. ********** ***Answer: No it is not as clearly indicated by the study findings that people who got transplant treatment survived.

(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment. ********** ***Answer: we can see from the data that the efficacy is higher in treatment group as can be seen from the data in q3 group.

(c) What proportion of patients in the treatment group and what proportion of patients in the control group died? ********** ***Answer:

```
heartTr
control_level<-subset(heartTr,heartTr$transplant=='control')
control_deadpatients<-subset(heartTr,heartTr$survived=='dead' &  heartTr$transplant=='control')
controlpercentdied=100*nrow(control_deadpatients)/nrow(control_level)
print(paste0("Total percent died in control group is :",controlpercentdied))

treatment_level<-subset(heartTr,heartTr$transplant=='treatment')
treatment_deadpatients<-subset(heartTr,heartTr$survived=='dead' &  heartTr$transplant=='treatment')
treatmentpercentdied = 100*nrow(treatment_deadpatients)/nrow(treatment_level)

print(paste0("Total percent died in treatment group is :",treatmentpercentdied))
```

(d) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

i. What are the claims being tested? **********
Answer: relationship of importance of treatment leading to survival of patients. **********

ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

We write *alive* on **24+4**_____ cards representing patients who were alive at the end of the study, and *dead* on **30+4**_____ cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size **69**_____ representing treatment, and another group of size _____34_____ representing control. We calculate the difference between the proportion of *dead* cards in the treatment and control groups (treatment - control) and record this value. We repeat this 100 times to build a distribution centered at **0**_____. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are _____30/34-45/69=.23____. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.

iii. What do the simulation results shown below suggest about the effectiveness of the transplant program?

Answer:Analysis show that the treatment is effective in saving lives. **********