Chapter 2 - Summarizing Data

SOLUTIONS ARE FOR PERSONAL USE ONLY. NOT NOT DISTRIBUTE OR SHARE.

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

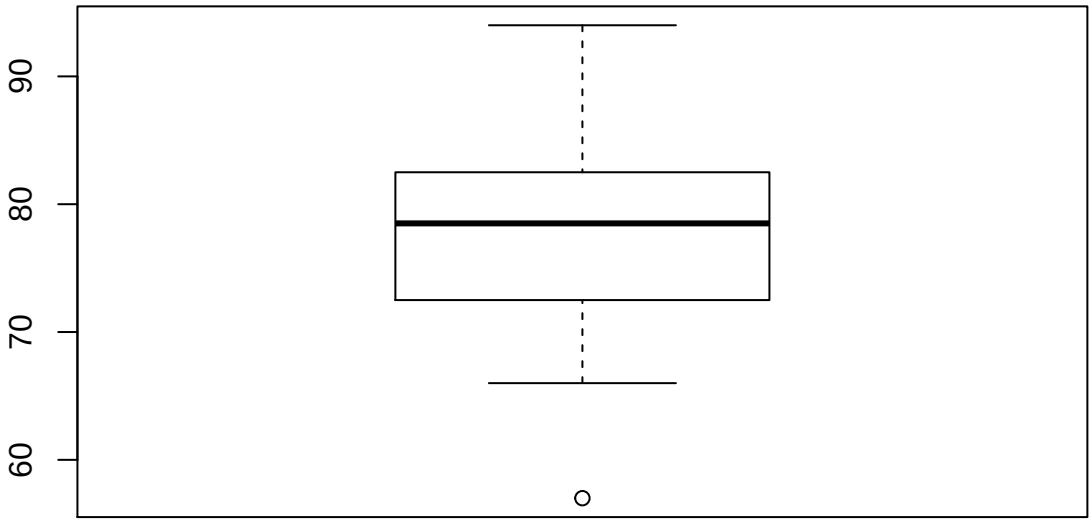
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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

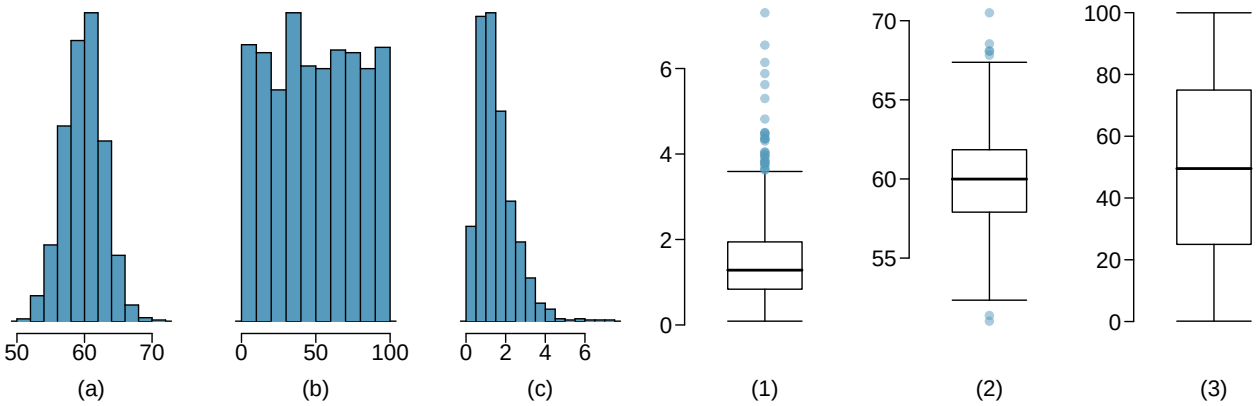
Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

Solution

```
boxplot(scores)
```



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



Solution

(a) The distribution is unimodal and symmetric, and about 95% of the data falls within about 7 units of the center, so the standard deviation will be about 3 or 4. This matches box plot (2).

(b) The distribution is uniform and values range from 0 to 100. This matches box plot (3) which shows a symmetric distribution in this range. Also, each 25% chunk of the box plot has about the same width and there are no suspected outliers.

(c) The distribution is unimodal and right skewed with a median between 1 and 2. 25th and 75th percentile are near 1 and 2, so the IQR is roughly 1. This matches box plot (1).

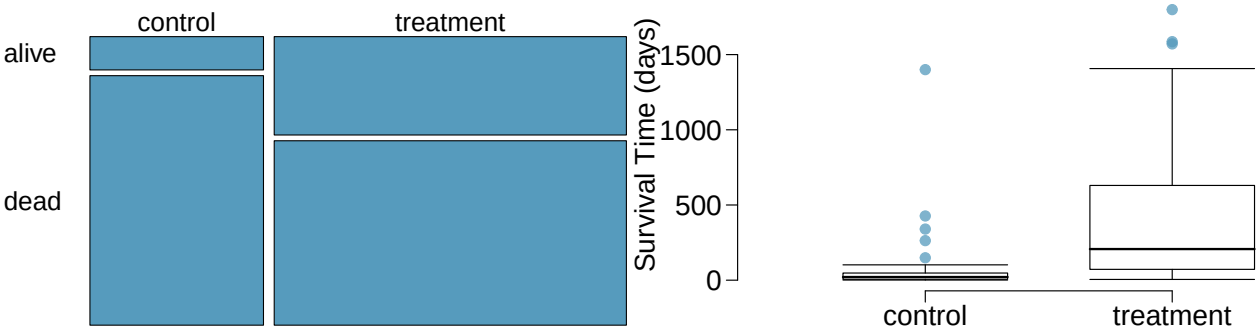
Distributions and appropriate statistics, Part II. (2.16, p. 59) For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability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

Solution

(a) The distribution is right skewed with potential outliers on the positive end, therefore the median and the IQR are preferable measures of center and spread.
(b) The distribution is somewhat symmetric and has few, if any, extreme observations, therefore the mean and the standard deviation are preferable measures of center and spread.
(c) The distribution would be right skewed. There would be some students who did not consume any alcohol, but this is the minimum since students cannot consume fewer than 0 drinks. There would be a few students who consume many more drinks than their peers, giving the distribution a long right tail. Due to the skew, the median and IQR would be preferable measures of center and spread.
(d) The distribution would be right skewed. Most employees would make something on the order of the median salary, but we would anticipate upper management makes much more. The distribution would have a long right tail, and the median and the IQR would be preferable measures of center or spread.

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.

(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.

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

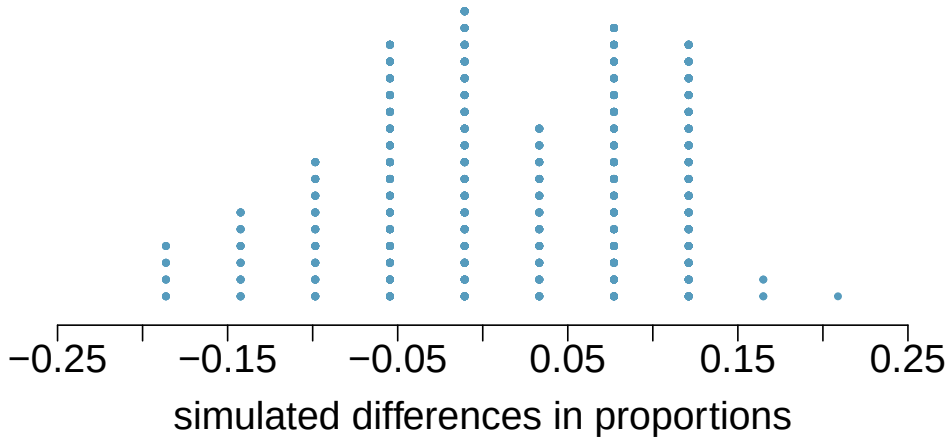
(d) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

i. What are the claims being tested?

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 _____ cards representing patients who were alive at the end of the study, and *dead* on _____ cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size _____ representing treatment, and another group of size _____ 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 _____. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are _____. 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?



Solution

(a) Proportion of patients who are alive at the end of the study is higher in the treatment group than in the control group. These data suggest that survival is not independent of whether or not the patient got a transplant.

(b) The shape of the distribution of survival times in both groups is right skewed with one very clear outlier for the control group and other possible outliers in both groups on the high end. The median survival time for the control group is much lower than the median survival time for the treatment group; patients who got a transplant typically lived longer. Tying this together with the much lower variability in the control group, evident by a much smaller IQR than the treatment group (about 50 days versus 500 days), and we can see that patients who did not get a heart transplant tended to consistently die quite early relative to those who did have a transplant. Overall, very few patients without transplants made it beyond a year while nearly half of the transplant patients survived at least one year. It should also be noted that while the first and third quartiles of the treatment group is higher than those for the control group, the IQR for the treatment group is much bigger, indicating that there is more variability in survival times in the treatment group.

(c) Proportion of patients who in the treatment group that died: $45 = 0.652$ 69 Proportion of patients who in the control group that died: $30 = 0.882$ 34

(d) i. H0: The variables group and outcome are independent. They have no relationship, and the difference in survival rates between the control and treatment groups was due to chance. In other words, heart transplant is not effective.
HA: The variables group and outcome are not independent. The difference in survival rates between the control and treatment groups was not due to chance and the heart transplant is effective.

ii. 28, 75, 69, 34, 0, -0.23 or lower.

iii. Under the independence model, only 2 out of 100 times (2%) did we get a difference of -0.23 or lower between the proportions of patients that died in the treatment and control groups. Since this is a low probability, we can reject the claim of independence in favor of the alternate model. There is convincing evidence to suggest that the transplant program is effective.