Chapter 2 - Summarizing Data

Leticia Salazar

2021-09-08

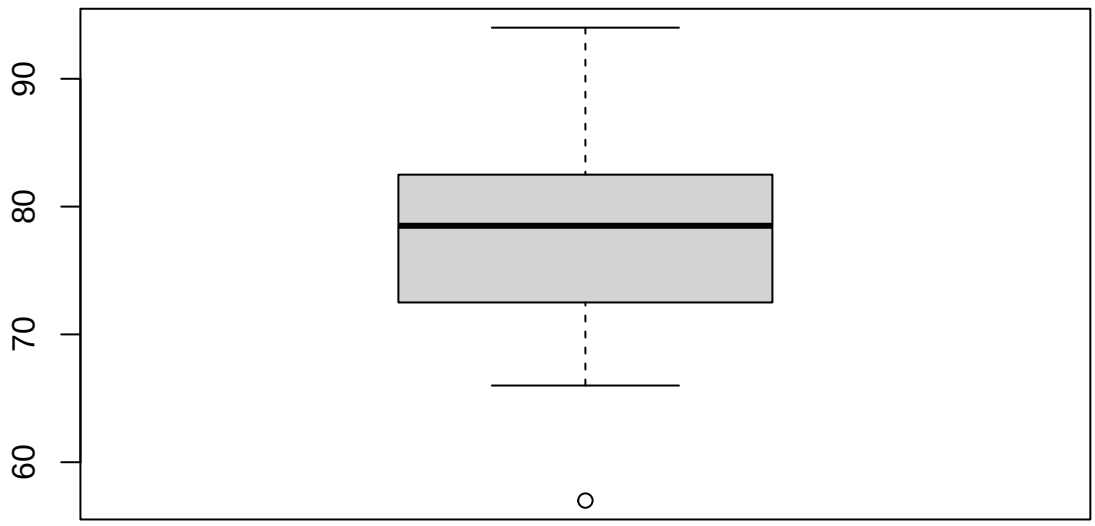
Stats scores. (2.33, p. 78) Below are the final exam scores of twenty introductory statistics students.

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

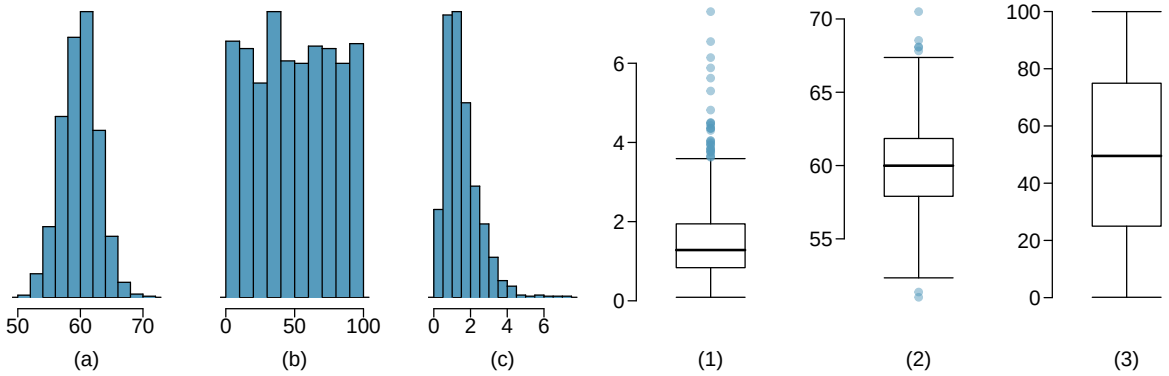
Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

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

```
# Create Box plot for Scores data
boxplot(scores)
```



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



- ***(a) Histogram a is a symmetric uni-modal distribution and matches box plot 2.***
- ***(b) Histogram b is multi-modal distribution and matches box plot 3.***
- ***(c) Histogram c is right skewed uni-modal distribution and 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.

- ***This data is right skewed since most of the data is towards the left leaving a long tail to the right for those houses that cost more than $6,000,000. Since the data is skewed to the right Median and IQR would be the best representation for typical observation and variability. The median would represent a more accurate look into the data since there are some outliers present in the cost of housing. The IQR would show those houses that are out of range.***

(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.

- ***This data is symmetric meaning there's an even distribution of the data. Mean and Standard deviation would be more accurate representation for typical observation and variability. Since the data is symmetric there shouldn't be any outliers that would prevent a none accurate representation.***

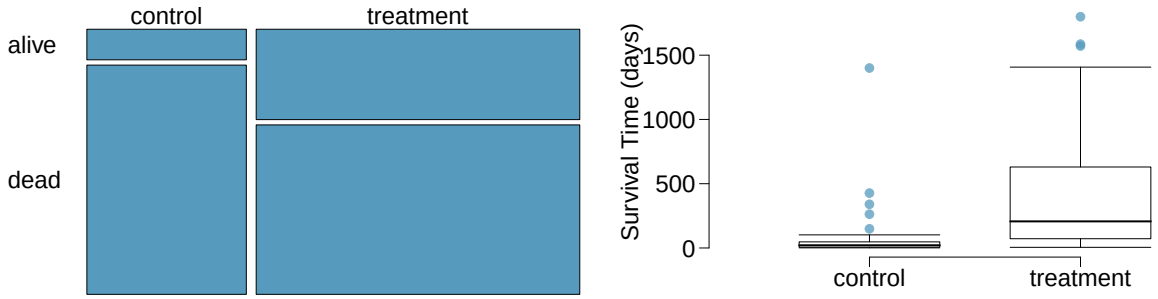
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

- ***This data is right skewed because plenty of students will be under 21 in a college setting. This gives the data concentration to be mainly on the left, leaving a right tail to those student who are 21 and drink. The median and IQR would also be a good representation of the data. Since there are a couple of outliers (those student's who are 21 and drink excessively) the median would be able to show the middle number from the data set and IQR will should the outliers.***

(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.

- ***This data is a symmetric distribution since only a few are high level executives and the rest earn average salaries. Mean and standard deviation will be useful to show the typical observation and variability because there are no outliers and the data is evenly distributed.***

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.

- ***Based on the mosaic plot, survival is not independent of whether or not the patient got a transplant. You see that within the control group there are less people who survived as opposed to those who received the treatment.***

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

- ***Based on the box plots, the control group had less chance of survival therefore, the effectiveness of the heart transplant treatment is greater.***

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

- ***For the control group, a total of 34 patients; 30 died and 4 lived (88.24%) while in the treatment group a total of 69 patients; 45 died and 24 lived (65.22%).***

(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?

- ***Null Hypothesis: Patient's survival is independent on whether they get the transplant or not.***
- ***Actual Hypothesis: Patient's survival is dependent on receiving the transplant.***

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 **28** cards representing patients who were alive at the end of the study, and *dead* on **75** 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 **24/69 - 4/34 = 0.23**. 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?

- ***This simulation graph shows that the further away from 0 the data gets we can reject our null hypothesis and indeed, the patient's survival is dependent on the transplant.***