Chapter 2 - Summarizing Data

Coffy Andrews-Guo

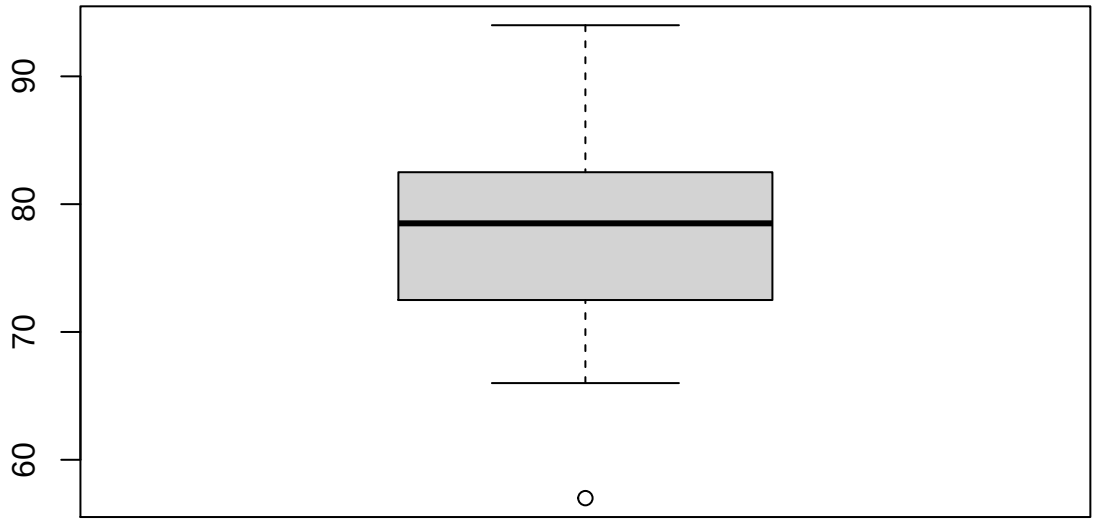
Stats scores. (2.33, p. 78) Below are the final exam scores of twenty introductory statistics students.

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

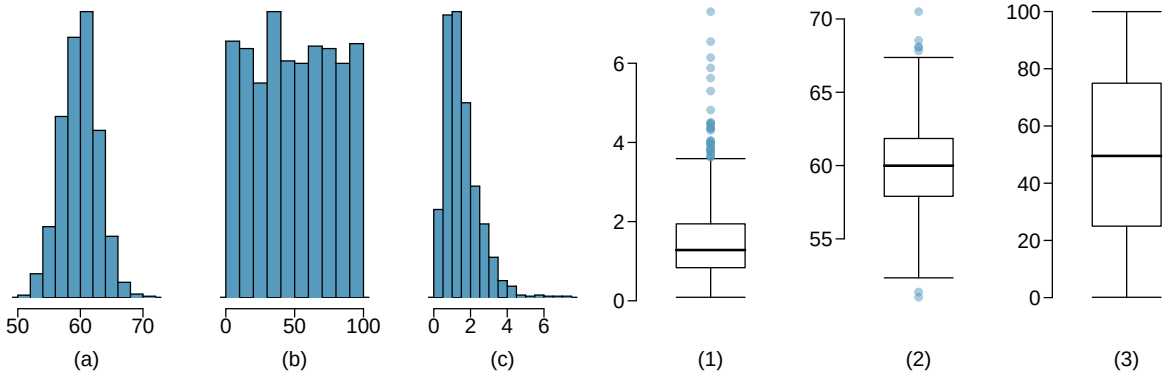
Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   57.00   72.75   78.50   77.70   82.25   94.00
```



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



The distribution in the histograms with matching box plots are as follows: Histogram A, is a unimodal distribution (has one prominent peak) and match box plot #2 Histogram B, is a multimodal distribution (has more than two prominent peaks) and match box plot #3 Histogram C, is a bimodal distribution (has two peaks) and match 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.

**Response (a)

The distribution of the data will be a strong left skewed histogram and the median house cost will be the best representation for this observational data. To show the variability of the observations the IQR calculation should be used because the number of houses that cost more than $6,000,000 extends beyond the 75th percentile and very distant from most of the data.**

(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.

**Response (b)

The distribution of the data will be left skewed and the mean house cost will be the best representative for this observational data. To show the variability of the observations the standard deviation should be used because houses that cost more than $1,200,000 is within range and inside the maximum observed values.**

(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.

**Response (c)

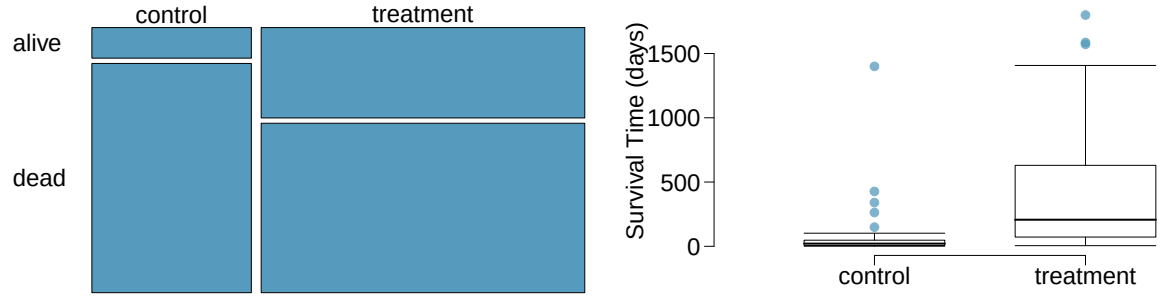
The distribution of the data will be symmetric and the mean will be the best representative for this observational data. To show the variability of the observations the standard deviation should be used because the consumption of drinks will peak after the initial and reduce at the due to implementing moderation or limitations to focus on school courses.**

(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.

**Response (d)

The distribution of the data will be a strong right skewed histogram and the median annual salaries will be the best representation for this observational data. To show the variability of the observations the IQR calculation should be used because the high level executives much higher salaries will negatively affect the range of the other employees.

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.

Response (a)

The mosaic plot shows that survival is dependent on whether or not the patient got a transplant.

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

Response (b)

The plot box suggest the treatment group had more alive cancer patients than the control group.

(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?

Response (d.i): The claims being tested are (1) whether a heart transplant will increase lifespan and (2) the study results will indicate a dependency model.

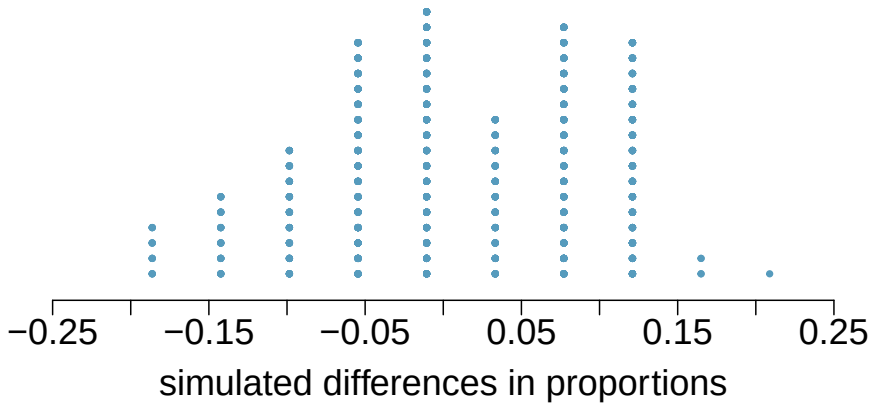
ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

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*** *_53_____* representing treatment, and another group of size _____50_____ 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 ***1.08_*.** **Lastly, we calculate the fraction of simulations where the simulated differences in proportions are** _-0.016_____. 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?

Response (d.iii)

The simulation results shows the cancer patients in treatment had a greater rate of living than the control group.



Document available on RPubs GitHub