Chapter 2 - Summarizing Data

Irene Jacob

2020-09-10

Stats scores

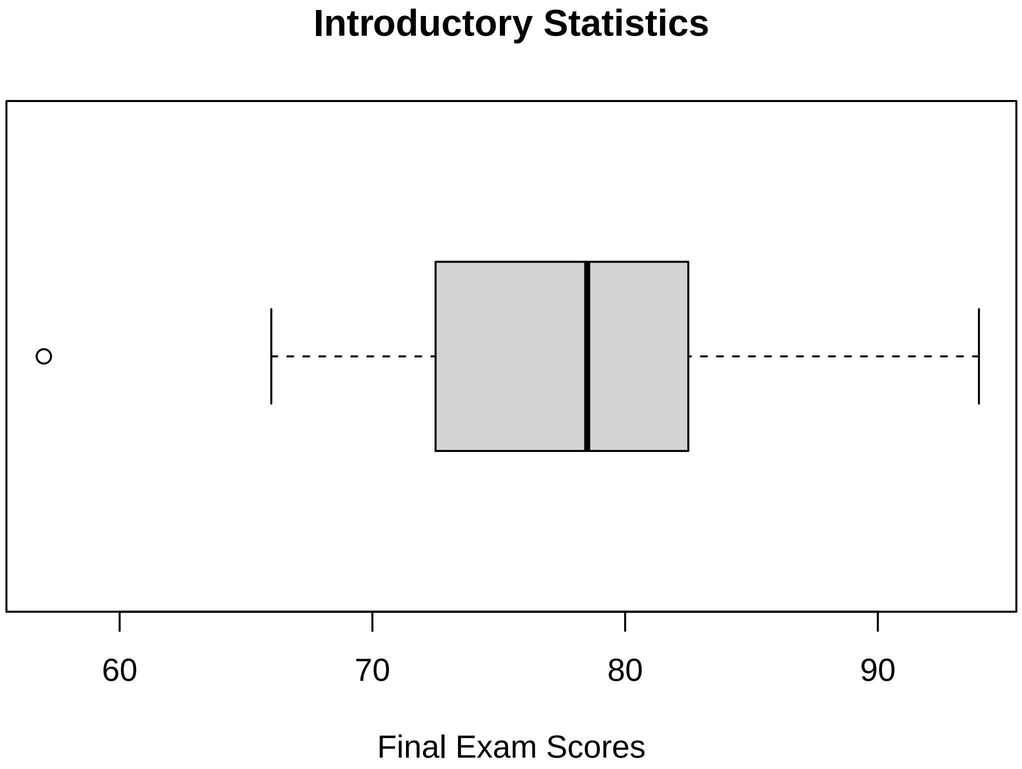
Below are the final exam scores of twenty introductory statistics students.

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.

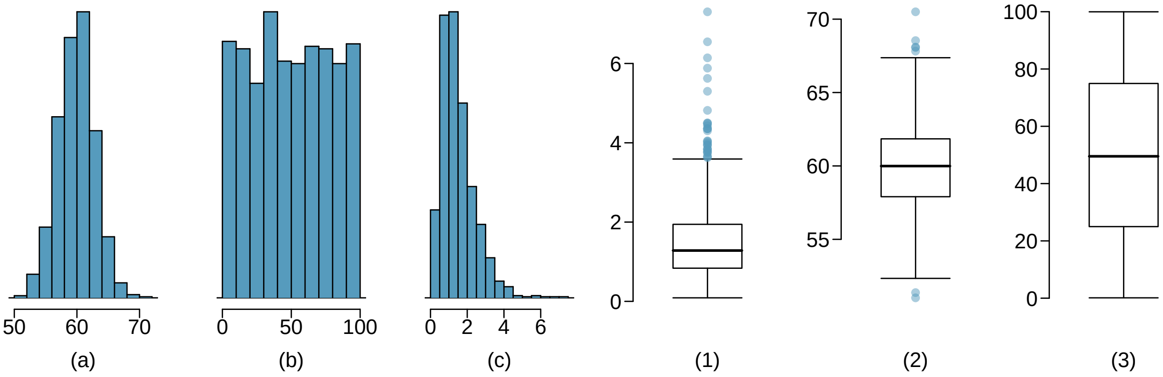
Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

```r
scores <- c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94)
boxplot(scores, xlab = "Final Exam Scores", main = "Introductory Statistics", horizontal = TRUE)
```



Mix-and-match

Describe the distribution in the histograms below and match them to the box plots.



The description is as follows:

Histogram	Match
(a) symmetrical distribution	(2)
(b) multimodal distribution	(3)
(c) right-skewed distribution	(1)

Distributions and appropriate statistics, Part II

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 has right skewed distribution. The outliers do not have much effect on the data so Median would best represent this data. Due to this IQR would better represent the variability 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.

This data has symmetrical distribution. There are no outliers and due to the symmetry Mean would better represent this data. Due to this Standard Deviation would better represent the variability of the data.

(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.

Since majority of the students are below 21 the number of drinks consumed will keep decreasing as the age decreases. This makes it a right skewed distribution. The outliers do not have much effect on the data so Median would best represent this data. Due to this IQR would better represent the variability of the data.

(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 has right skewed distribution. The outliers do not have much effect on the data so Median would best represent this data. Due to this IQR would better represent the variability of the data.

Heart transplants

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.

Survival is not independent of whether or not the patient got a transplant. From the mosaic plot and the box plot it is clear that the paitients who got the treatment have a higher survival rate.

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

The Q1 of treatment group is almost same as the Q3 of control group so the heart transplant treatment has roughly 50% more survival rate.

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

From the below calculations we can see that **65.22%** of patients in the treatment group and **88.24%** of patients in the control group died.

```
n_treat <- subset(heart_transplant, heart_transplant$transplant == 'treatment')
dead_treat <- subset(heart_transplant, heart_transplant$survived == 'dead' & heart_transplant$transplan
nrow(dead_treat)/nrow(n_treat)
```

```
## [1] 0.6521739
```

```
n_control <- subset(heart_transplant, heart_transplant$transplant == 'control')
dead_control <- subset(heart_transplant, heart_transplant$survived == 'dead' & heart_transplant$transpla
nrow(dead_control)/nrow(n_control)
```

```
## [1] 0.8823529
```

(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?

"Heart transplants decrease the number of deaths" is the claim being tested here.

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.

```
nrow(dead_control)/nrow(n_control) - nrow(dead_treat)/nrow(n_treat)
```

```
## [1] 0.230179
```

> 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 **23.02%**. 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?

The simulated differences in proportions is small as it is **23.02%**. The null hypothesis should be rejected.