Chapter 2 - Summarizing Data

Gabriel Campos

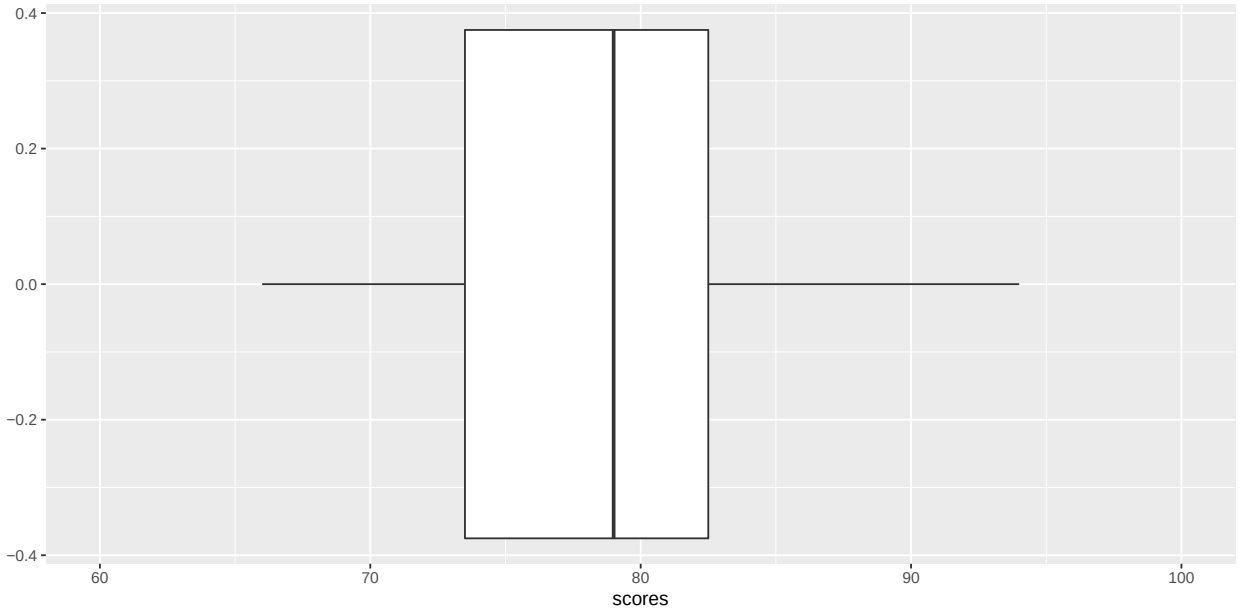
Stats scores. (2.33, p. 78) Below are the final exam scores of twenty introductory statistics students.

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

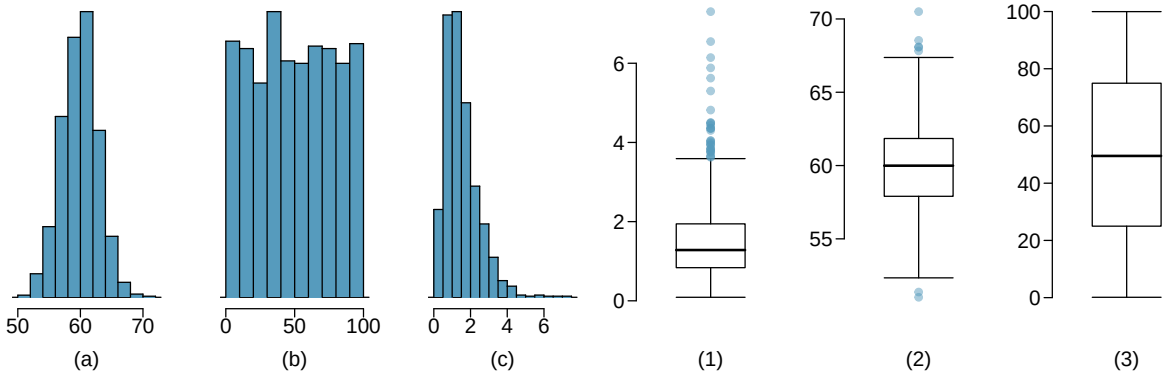
Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

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

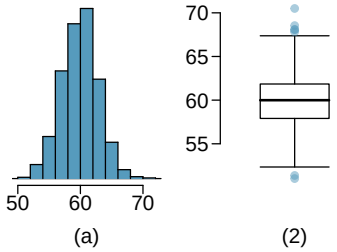
```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



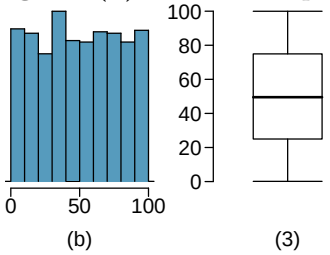
Mix-and-match. (2.10, p. 57) Describe the distribution in the histograms below and match them to the box plots.



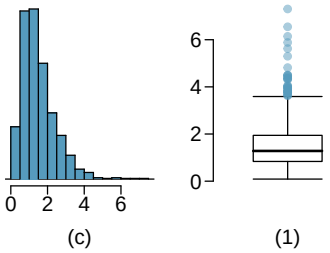
Histogram (a) is symmetrical so boxplot (2) is it's best match



Histogram (b) is has multiple modals or is multimodal and so boxplot (3) is it's best match



Histogram (c) is left skewed therefore it is matched with boxplot (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.

25% = Q1 = \$350,000 | 50% = Q2 = \$450,000 | 75% = Q3 = \$1,000,000

Q2-Q1 < Q3-Q2 AND *"... a meaningful number of houses that cost more than \$6,000,000"* indicates the distribution is right skewed. Median and IQR therefore are best used for representing a typical observation and variability, since the mean and standard deviation would shift based on the prices of the more expensive homes.

(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.

The distribution is symmetric with outliers that are not in excess nor drastically more than the rest of the distribution. In this scenario, mean and standard deviation would be best.

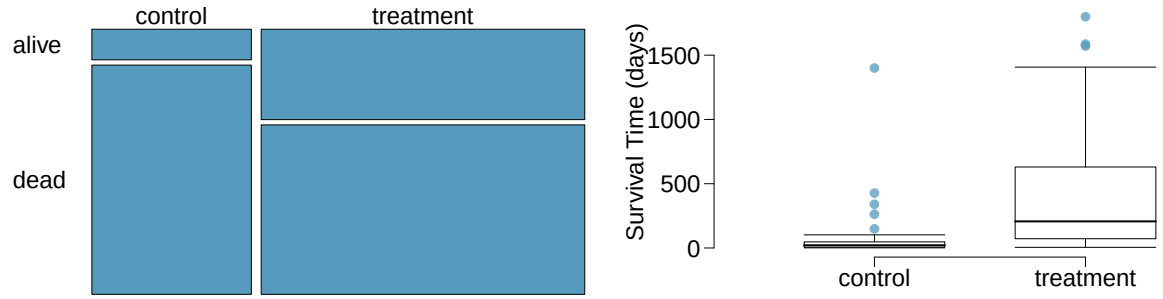
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

The distribution is based on a population that is at or close to 0 alcoholic drinks, with only a few outliers that drink in excess. Therefore the distribution should be left-skewed, median and IQR should be used to reduce the affect of the non or excessive drinkers, since standard deviation would be sensitive to the two extremes.

(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.

The distribution is symmetric, but because the few higher salaries are *much higher* the median and IQR should be used as in scenario (a)

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.
Noting the larger "dead" quantity for the control group, survival is dependent on treatment.

(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.
When comparing the median of the control group

```
# mean of control survival time
#mean(heartTr$survtime[heartTr$transplant == "control"])
# median of control suvival time
median(heartTr$survtime[heartTr$transplant == "control"])
```

```
## [1] 21
```

with that of the treatment

```
# mean of control survival time
#mean(heartTr$survtime[heartTr$transplant == "treatment"])
#median of control suvival time
median(heartTr$survtime[heartTr$transplant == "treatment"])
```

```
## [1] 207
```

We can conclude the survival rate is significantly higher for the treatment group considering the median for the control group is 21 vs 207 for the control group. The larger median and larger value outliers means there is a longer survival time for those who have a transplant compared to those that do not.
(c) What proportion of patients in the treatment group and what proportion of patients in the control group died?

```r
# Get count of control and treatment
count(heartTr,transplant)
```

```
##   transplant  n
## 1    control 34
## 2  treatment 69
```

```r
# proportion for treatment
sum(heartTr$survived == "dead" & heartTr$transplant == "treatment")/sum(heartTr$transplant == "treatmen
```

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

```r
# proportion for treatment
sum(heartTr$survived == "dead" & heartTr$transplant == "control")/sum(heartTr$transplant == "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?
Whether survival rate is dependent on receipt of a 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.

```r
# Number patients who are survivors
sum(heartTr$survived == "alive")
```

```
## [1] 28
```

```r
# patients who did not survive
sum(heartTr$survived == "dead")
```

```
## [1] 75
```

```r
# Number of patients in treatment group
sum(heartTr$transplant == "treatment")
```

```
## [1] 69
```

```r
# Number of patients in control group
sum(heartTr$transplant == "control")
```

```
## [1] 34
```

```r
# Calculation for ration
(1-sum(heartTr$survived == "dead" & heartTr$transplant == "treatment")/sum(heartTr$transplant == "treatr
```

```
## [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** [0.230179] ___. 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 chart shows that no observation had a difference of at least 23%, making the chances of such an observation 0%. Because of this we can reject the independence or null hypothesis and conclude having a heart transplant increases survival rate 100% of the time.