Chapter 2 - Summarizing Data

Cameron Smith

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

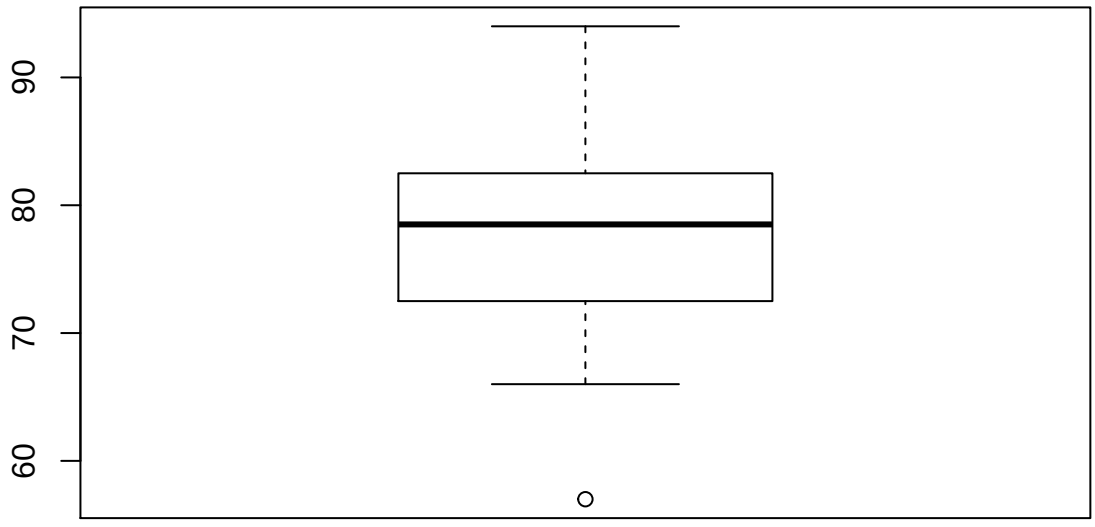
Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

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

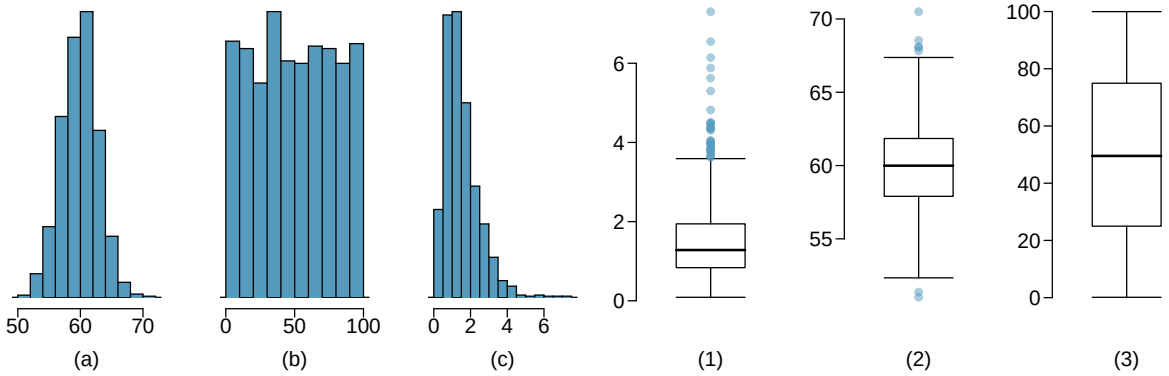
Answer

Following is the box plot:

```
boxplot(scores)
```



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



Answer

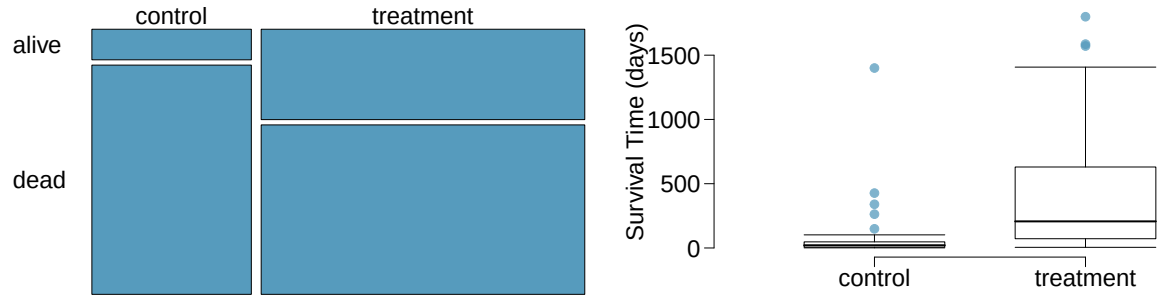
(a) This is a a bimodal distribution, and is symmetric. Matches with boxplot 2.
(b) This is a multimodal distribution. The values are fairly uniform. Matches with boxplot 3.
(c) This is a bimodal distribution, right skewed. Matches with boxplot 1.

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.
(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.

Answer
(a) Expected distribution = **right skewed**, Mean vs median = **median**
(b) Expected distribution = **symmetric**, Mean vs median = **mean**
(c) Expected distribution = **left skewed**, Mean vs median = **median**
(d) Expected distribution = **right skewed**, Mean vs median = **mean**

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.

Answers
(a) Based on these results we can reject independence. Although the sample size was relatively small, the significant difference in lifespans most likely overcomes the possibility of 'random noise'. Specifically, 88% of those in the control group died and only 65% of those in the treatment group died - a difference of 23%.
(b) The box plots make it pretty clear that those in the treatment group tended to live significantly longer - many by a year or two and several (the 4th quartile) up to several years.
(c) 65% in the treatment group died and 88% in the control group died.

i. What are the claims being tested?

Answer
The claim being tested is that an experimental heart transplant program increased lifespan. The testing is being done to confirm whether the result was purely due to chance.

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.

Answer is within text below

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 **due to chance**. 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?

Answer
The low percentages of random chance in the simulation in comparison to the actual difference (23%) at the outcome of the study suggests an alternative model, which is that the heart transplant program was in fact effective in inreasing lifespan for the patients who participated in the stufy. It should be noted however that without additional studies it cannot be expected that these results can be replicated to the entire population, only two patients who were similar to those in the study (i.e. gravely ill and would most likely benefit from a new heart).