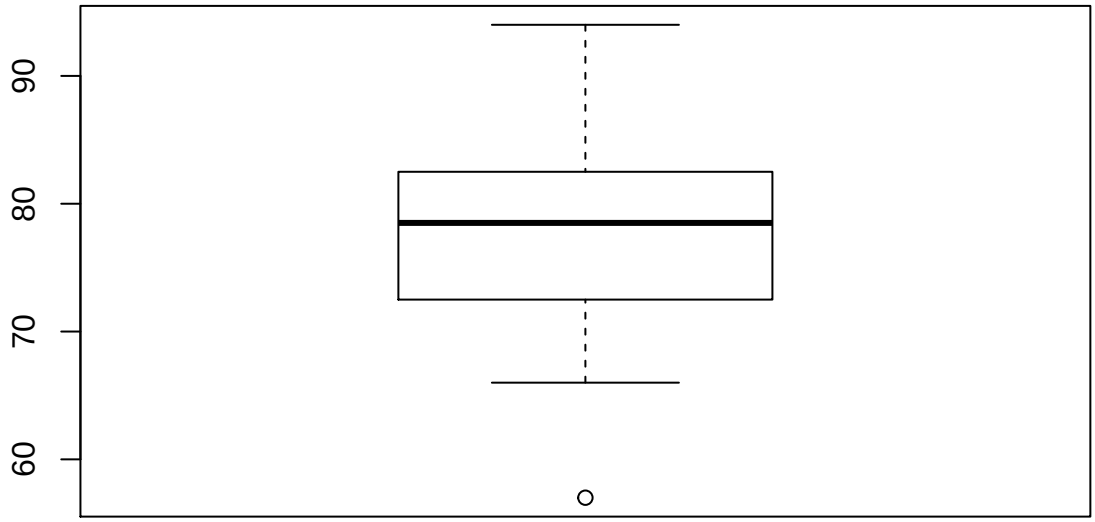
Chapter 2 - Summarizing Data

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

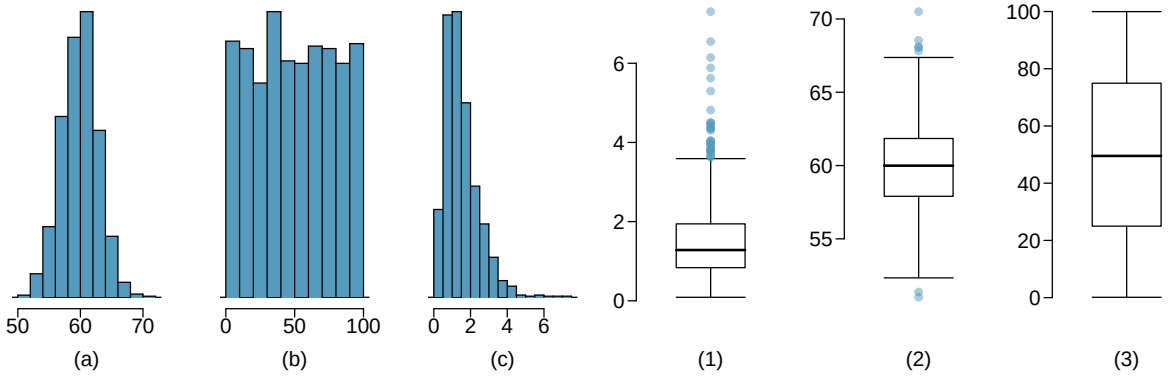
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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



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



- a: A unimodal histogram that is approximately normally distributed with values ranging from 50-70
- b: A unimodal histogram that is approximately uniformly distributed with values ranging from 0-100
- c: A multimodal histogram that is strongly right skewed with values ranging from 0-6
- Histogram a matches box plot 2
- Histogram b matches box plot 3
- Histogram c 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.

With 75% of the houses costing less than 1M while a "meaningful" cost 6x more would cause a right skewed histogram. The median/IQR would best represent a typical observation from this data set since the mean/std will be inflated due to the more expensive houses.

(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.

With very few houses costing more than 1.2M which is 33% more that 900K would imply a symmetric histogram. Prices are within a narrow band so a mean/std will be a good representation since the spread of data is small, although a median/IQR would be bad as well

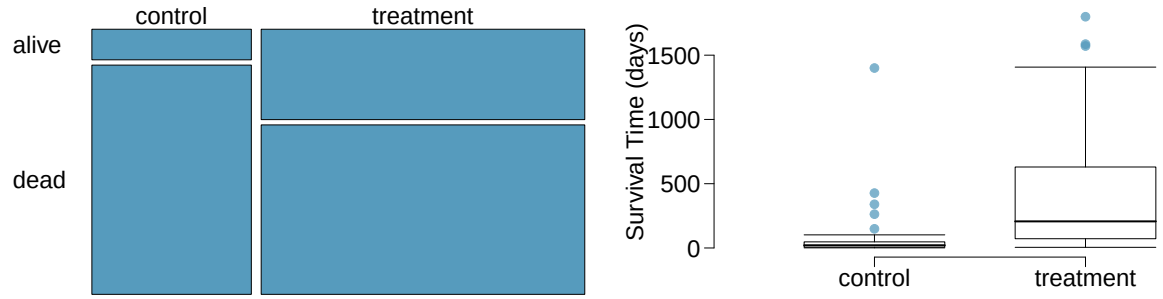
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

With most students not drinking, the histogram would be empty for smaller bins counts and very tall for larger bin counts thus producing a left skewed distribution. The median/IQR would best represent a typical observation from this data set since the mean/std will be suppressed due to the many 0s in the data set.

(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.

With only a few high level executives the histogram will be unimodal on the right size causing a left tail. The median/IQR would best represent a typical observation from this dataset since the mean/std will be inflated due to only a few high level executives earn much higher salaries.

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.

From the mosaic plot those that got the treatment are ~3x more likely to survive than those who didn't, thus making survival dependent on those who got the treatment.

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

The control has very low efficacy with a few outliers, the 4 that survived plus one more. The treatment has high efficacy only a few outliers and most values are above the median

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

	dead	alive	Total
treatment	45	24	69
control	30	4	34
Total	75	28	103

Survived	Died
24/69~=35%	45/69~=65%
4/34~=12%	30/34~=88%

(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?

The researchers are trying to determine whether an experimental heart transplant program will increased lifespan of an already ill heart patient. Control: A patient who stays with their current heart Treatment: A patient given a new heart

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 _____0.05_____. 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 results below show the transplant was effective