Lab 2b: More distributions and boxplots

Stat 131A, Fall 2018

Learning Objectives:

- Understand quantiles, percentiles, and quartiles.
- Learn how to read boxplots.
- Learn how to make basic boxplots in R.

General Instructions

- Write your solutions in an `Rmd` (R markdown) file.
- Name this file as `lab02b-first-last.Rmd`, where `first` and `last` are your first and last names (e.g. `lab02b-gaston-sanchez.Rmd`).
- Knit your `Rmd` file as an html document (default option).
- Submit your `Rmd` and `html` files to bCourses, in the corresponding lab assignment.

Problem 1

What percent of the observations in a distribution lie between the first quartile and the third quartile?

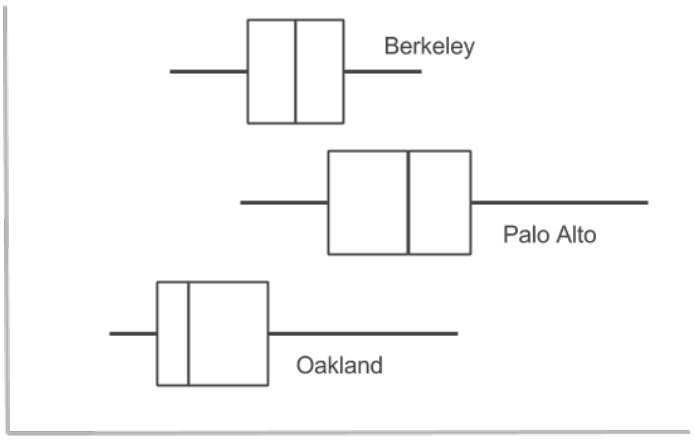
a. 25%

b. 50%

c. 75%

Problem 2

The figure below shows box plots for house prices in three cities around the bay area.



a. Which city has the lowest house price?

b. Which city has the highest house price?

c. Which city has the highest minimum house price?

d. Which city has the smallest median house price?

e. Which city has the highest median house price?

f. Which city has the smallest range of house prices?

g. Which city has the largest interquartile range?

h. Which city has the smallest interquartile range?

i. Which city seems to have a symmetric distribution?

j. Which city has the most right skewed distribution?

Problem 3

The following data (in an R vector) give the total number of fires in Ontario, Canada, in the months of 2002:

```
fires <- c(6, 13, 5, 7, 7, 3, 7, 2, 5, 6, 9, 8)
fires
```

```
##  [1]  6 13  5  7  7  3  7  2  5  6  9  8
```

Use R to calculate the following summaries:

- Minimum
- Maximum
- Range
- Q_1 (25th percentile)

- Q_2 (50th percentile)
- Q_3 (75th percentile)
- Mode
- Mean
- Variance
- Standard Deviation

Problem 4

With the results obtained in the previous question, use the function `boxplot()` to graph:

a. A default boxplot of the data.

b. A boxplot in which the whiskers comprise all the data (from min to max). Hint: look at the help documentation `?boxplot` and find which argument allows you to determine how far the whiskers extend out from the box.

Problem 5

Here is the 5-number summary for a group of 100 runners in a 5-kilometer race. The variable is the "time to complete the race":

- Minimum: 15 minutes
- Q1: 27 minutes
- Median: 31 minutes
- Q3: 32 minutes
- Maximum: 50 minutes

True or False:

a. Most people finished between 15 and 50 minutes, but some people took a little longer.

b. There were more runners in the 2nd quartile (Q1 = 27 min to Median = 31 min) than in the 3rd quartile (Median = 31 min to Q3 = 32 min).

c. At least 25 runners had times ranging from 31 minutes to 32 minutes.

Problem 6

Below is the five-number summary for 136 hikers who recently completed the John Muir Trail (JMT).

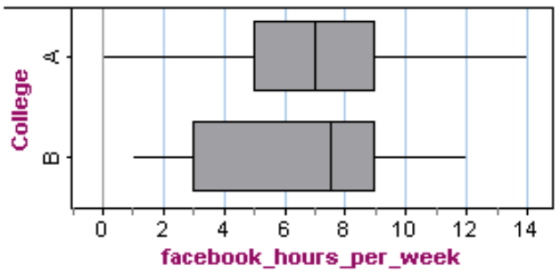
The variable is the amount of time to complete the 212-mile hike from Yosemite Valley across the high Sierras to the top of Mount Whitney.

- Minimum: 9 days
- Q1: 18 days
- Median: 21 days
- Q3: 28 days
- Maximum: 56 days

a) Use the 1.5 * IQR rule to determine which one of the known values is an outlier?

Problem 7

In a survey, students at two colleges estimated the hours they spend on Facebook each week.



Which of the following statements is a valid conclusion that can be drawn from the boxplots? Indicate all that apply.

a. The medians are close; they differ by about half an hour. This suggests that the typical student at each college spends about the same amount of time on Facebook each week.

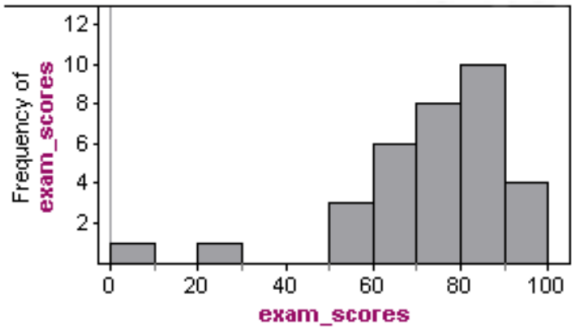
b. If we use Q1 and Q3 to define a typical range of values, the typical students at College B are on Facebook 7.5 to 9 hours a week.

c. If we use range as a measure of variability, then College A has more variability.

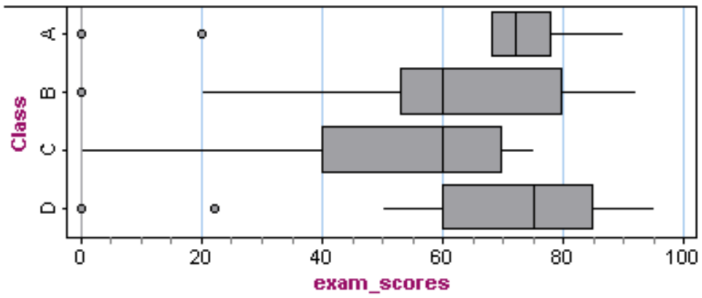
d. If we use the interquartile range as a measure of variability, then College A has less variability.

Problem 8

This histogram shows the distribution of exam scores for a class of 33 students.



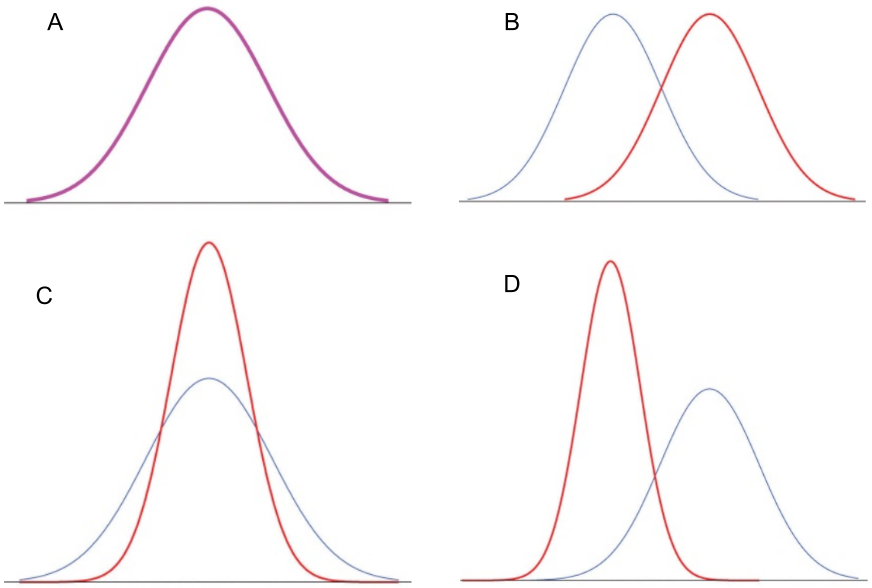
Which of the following boxplots represents the data displayed in the above histogram?



a. Boxplot A
b. Boxplot B
c. Boxplot C
d. Boxplot D

Problem 9

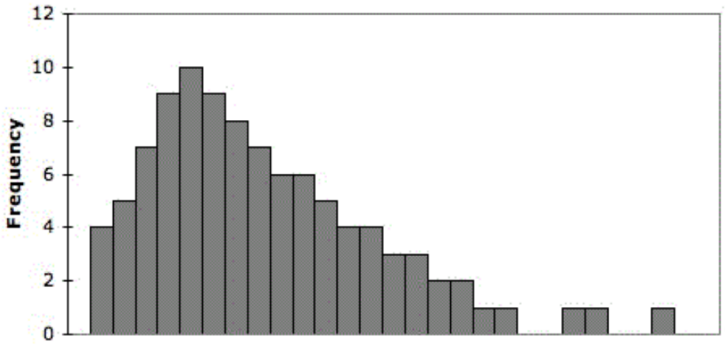
Match the following figures with the corresponding descriptions:



- Centers differ.
- Centers and spread differ.
- Two identical distributions.
- Dispersions differ.

Problem 10

Consider a quantitative data set with histogram shown below.



Which would be a better tool for measuring center and spread? Explain.

a. Median and IQR

b. Mean and standard deviation

Problem 11

Three instructors are comparing scores on their finals: each had 99 students. In class A, one student got 1 points, another 99 points, and the rest got 50 points. In class B, 49 students got a score of 1, one student got a score of 50, and 49 students got a score of 99. In class C, one student got a score of 1, one student got a score of 2, one student got a score of 3, and so forth, all the way through 99. If you need to do calculations, trying doing them in R: the functions `seq()` and `rep()`, for sequencues and repetitions, respectively, are your friends.

a. Which class had the biggest average? or are they the same?

b. Which class had the biggest SD? or are they the same?

c. Which class had the biggest range? or are they the same?

Problem 12

Suppose that 40 students take a quiz and everyone in the class scores an 80.

Which of the following best represents the standard deviation of scores for this quiz?

a. 80

b. 2

c. 0

Problem 13

A study on college students found that the men had an average weight of about 66kg and an SD of about 9kg. The women had an average weight of about 55kg and an SD of 9kg.

a. Find the averages and SDs, in pounds (1kg = 2.2lb)

b. If you took the men and women together, would the SD of their weights be smaller than 9kg, just about 9kg, or bigger than 9kg? Why?

Problem 14

Consider a large group of people and suppose we measure each of their heights.

Which of the following data sets would have a higher standard deviation?

a) The set of the people's heights measured in inches.

b) The set of the people's heights measured in feet.