

Comparing Datasets (Boxplots)

Present: 13.3 #37, and 13.4 #1 and 21

9:55 Kaitlin Hei., Kaitlyn, Mandi, Grace, Molly, Azjja Kjo.

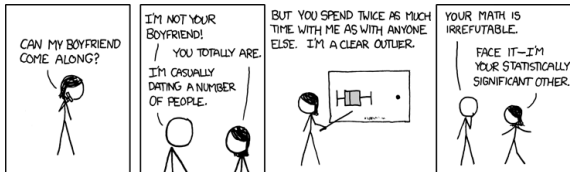
11:00 Caitlyn McL., Gen, Danielle, Steven, Emily, Jordan Pet.

Homework: See handout.

Discussion Leaders Monday:

9:55 Katie Kle., Megan, Katie, Alyssa, Nicole, Rachel Roc.

11:00 Kyle Pol., Rebecca, Emma, Jamie, Erin, Jamie Sch.



Brand A
 $\bar{x} = 43,560$
 $S \approx 2000$

Brand B
 $\bar{x} = 48,175$
 $S \approx 2000$

gap ~ 5000 ,
 or > 2
 st. dev. apart

After reviewing the correct solution (below), write your score on the back of your quiz.

- 0 = no progress at all; just rewrote problem
- 0.5 = false start, not based on relevant principles
- 1 = false start, but sustained effort with some relevant principles
- 1.5 = significant mistake(s), or significant misunderstanding(s)
- 2 = mistake near the end or could not finish; also excessive reliance on calculator or 'brute force' methods
- 2.5 = trivial mistake (e.g. arithmetic error), but work is mostly correct
- 3 = correct answer and work

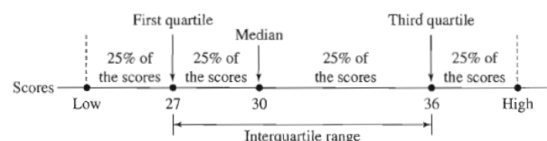
As we have seen, the mean absolute deviation (m.a.d.) of an n item dataset is just the average of the n "absolute deviations" from the mean. Suppose dataset A has 15 items and a m.a.d. of 3.2, and dataset B has 35 items and a m.a.d. of 2.8. Calculate to the nearest hundredth the weighted average of the mean average deviations.

$$\frac{3.2(15) + 2.8(35)}{15 + 35} = \frac{146}{50} = 2.92.$$

The 5-Number Summary, IQR, and Boxplots

Think About...

A data set of test scores has a first quartile score of 27, a median of 30, and a third quartile score of 36. What does each of the numbers 27, 30, and 36 mean? What is the interquartile range? What would the 50th percentile be?



Source: Sowder et al., 2010, p. 698

The **interquartile range** is defined as $Q3 - Q1$. It gives the amount of spread of the middle 50% of the data.

The larger IQR is, the more spread out the (middle 50% of) the data is.

ACTIVITY 5 Finding Quartiles

The following exam scores are from a class of 30 students:

65, 64, 46, 38, 58, 44, 65, 60, 50, 70, 55, 44, 68, 67, 66,
66, 81, 68, 51, 51, 75, 53, 47, 62, 59, 25, 78, 49, 77, 49

1. The first thing to do is order the data from smallest to largest. Making a stem-and-leaf plot can help order the data, as shown on the next page.

2	5
3	8
4	4 4 6 7 9 9
5	0 1 1 3 5 8 9
6	0 2 4 5 5 6 6 7 8 8
7	0 5 7 8
8	1

ACTIVITY 5 Finding Quartiles

The following exam scores are from a class of 30 students:

$\frac{30+1}{2} = 15.5$

65, 64, 46, 38, 58, 44, 65, 60, 50, 70, 55, 44, 68, 67, 66,
66, 81, 68, 51, 51, 75, 53, 47, 62, 59, 25, 78, 49, 77, 49

1. The first thing to do is order the data from smallest to largest. Making a stem-and-leaf plot can help order the data, as shown on the next page.

2	5
3	8
4	4 4 6 7 9 9
5	0 1 1 3 5 8 9
6	0 2 4 5 5 6 6 7 8 8
7	0 5 7 8
8	1

Median = $\frac{59+60}{2} = 59.5$
 Q1 = $\frac{49}{1}$
 Q3 = $\frac{67}{1}$
 IQR = $67 - 49 = 18$

Handwritten notes showing the formula for finding the position of the median:

For $n=4$: $\frac{n+1}{2} = \frac{4+1}{2} = 2.5$

For $n=5$: $\frac{n+1}{2} = \frac{5+1}{2} = 3$

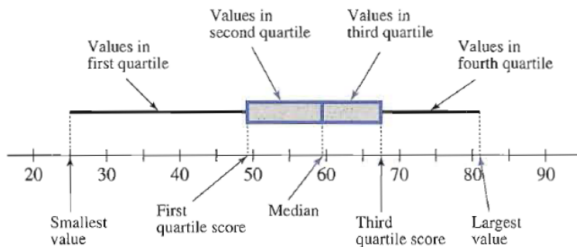
↓ gives the position of the median.

When presenting data, it is often helpful to give what is called the **five-number summary**. This summary consists of the smallest data value, the first quartile score, the median, the third quartile score, and the largest data value listed in order from smallest to largest. For the exam scores in Activity 5, the five-number summary is 25, 49, 59.5, 67, and 81.

* Start w/ a number line.

The horizontal scale is not plotted on a number line, so this box plot is distorted. It is wise to start with a carefully drawn number line before plotting the points for the five number summary.

Quartile scores are often depicted graphically through the use of a **box-and-whiskers plot** or more simply a **box plot**. Box plots are made using the five-number summary. Figure 9 below shows a box plot of the data set consisting of the 30 exam scores with the five-number summary 25, 49, 59.5, 67, 81.



Outliers

About half of our data values fall between the first and third quartile scores, so the smaller the interquartile range is, the closer these data values are together. Furthermore, statisticians often use the interquartile range to make judgments about what data values in a data set are extreme and thus warrant closer scrutiny, because extreme values might be errors or just not good representatives of the data set. Values that are either less than the first quartile score or greater than the third quartile score by more than one and a half times the length of the interquartile range are judged to be extreme and are labeled as *outliers*.

An **outlier** is a data value such that the

$$\text{data value} < \text{first quartile score} - 1\frac{1}{2} \times \text{IQR} \quad \text{or}$$

$$\text{data value} > \text{third quartile score} + 1\frac{1}{2} \times \text{IQR}$$

where IQR is the interquartile range.

Outliers are not used in a five-number summary. If the first score had been 20 in the exam data, 20 would be an outlier. We would then use the next value, 38, as the first number in our five-number summary.

1. Find Q1, Q2, Q3, and the IQR:

$$Q_2 = \text{Med} = 5.0$$

$$Q_1 = \frac{3.2 + 2.1}{2} = 2.65$$

$$Q_3 = \frac{6.5 + 5.7}{2} = 6.1$$

Chocolate sales dataset:

Country	Sales (billions of \$)
A	\$2.0
B	\$2.1 ← Q ₁
C	\$3.2
D1	\$4.9
D2	\$5.0 ← Med = Q ₂
E	\$5.1
F	\$5.7 ← Q ₃
G	\$6.5
H	\$16.6

IQR = 6.10 - 2.65 = 3.45

2. Show that \$16.6 billion is an outlier for this data set:

$$16.6 > Q_3 + 1.5(\text{IQR})$$

$$= 6.1 + 1.5(3.45)$$

$$= 6.1 + 5.175 = 11.275$$

So 16.6 is a high outlier. (Q₁, 2.65, 5, 6.1, 6.5)

3. Write the 5-number summary, and make corrections to the following box plot.

Chocolate Sales

