Chapter 2. Describing Distributions with Numbers

Note. In the previous chapter, we presented various ways to display data and then casually described such things as the center and overall pattern. In this chapter, we make things more quantitative and give formal definitions of various numerical parameters of data.

Measuring Center: The Mean

Definition. If n numerical observations are denoted by $x_1, x_2, \ldots, x_n$, their **mean** is

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

or in more compact notation

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Example S.2.1. Mean Slaps.

According to Michael Fleming's *The Three Stooges—An Illustrated History*, we have the following slap count for the Three Stooges shorts from the 1937 production year.

#	**Title**	**Slap Count**
20.	Grips, Grunts, and Groans	8
21.	Dizzy Doctors	7
22.	Three Dumb Clucks	8
23.	Back to the Woods	6
24.	Goofs and Saddles	8
25.	Cash and Carry	3
26.	Playing the Ponies	8
27.	The Sitter Downers	3

Find the mean number of slaps per film for the 1937 production year.

Solution. We have $n = 8$ data points, so in the notation of the formula for the mean, we have:

$$\begin{aligned} x_1 &= 8 & x_5 &= 8 \\ x_2 &= 7 & x_6 &= 3 \\ x_3 &= 8 & x_7 &= 8 \\ x_4 &= 6 & x_8 &= 3. \end{aligned}$$

So we get the mean of the data is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{8+7+8+6+8+3+8+3}{8} = 6.375.$$

Measuring Center: The Median

Definition. The **median** M is the midpoint of a distribution, the number such that half the observations are smaller and the other half are large. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.

2. If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $(n+1)/2$ observations up from the bottom of the list.

3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. The location of the median is again $(n+1)/2$ from the bottom of the list.

Example. Exercise 2.4 page 41. Find the median for these $n = 19$ (odd) numbers.

Example S.2.2. Median Stooges.

Find the median of the data from Example S.2.1. Notice that $n = 8$ is even in this example.

Comparing the Mean and the Median

Note. The mean and median of a roughly symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is usually farther out in the long talk than is the median.

Note. The median of a data set is not much influenced by very large or very small data values. It is said that the median is “resistant.” However, these extreme data values can have an influence on the mean.

Example. Exercise 2.4 page 41. Work as stated.

Measuring Spread: The Quartiles

Note. The mean and the median only give an idea of the central tendency of a set of data. Another important aspect of a data set is the spread of the data. One measure of the spread is the **range.**

Definition. The *range* of a data set is the difference between the largest and smallest observations.

Note. Other useful measures of the spread of a set of data are the quartiles. They are, informally, the median of the lower half of the data (called the **first quartile** and the median of the higher half of the data (called the **third quartile**).

Definition. The **quartiles** are calculated as:

1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.

2. The **first quartile** Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.

3. The **third quartile** Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Example S.2.3. Quartered Stooges.

Consider the slaps per film data of Fleming's *The Three Stooges—An Illustrated History* again.

Year	Slaps/Film	Year	Slaps/Film
1934	33.5	1947	31.9
1935	23.1	1948	9.9
1936	10.5	1949	14.4
1937	6.4	1950	19.6
1938	8.9	1951	21.1
1939	11.6	1952	13.6
1940	14.0	1953	16.0
1941	10.6	1954	8.5
1942	7.6	1955	14.8
1943	12.2	1956	17.8
1944	13.7	1957	6.4
1945	10.2	1958	11.4
1946	15.6		

Find the median M, the first quartile Q_1, and the third quartile Q_3.

Solution. First, we must arrange the data in order from smallest to largest:

rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Year	1937	1957	1942	1954	1938	1948	1945	1936	1941	1958	1939	1943	1952	1944	1940	1949	1955	1946	1953	1956	1950	1951	1935	1947	1934
slaps/film	6.4	6.4	7.6	8.5	8.9	9.9	10.2	10.5	10.6	11.4	11.6	12.2	13.6	13.7	14.0	14.4	14.8	15.6	16.0	17.8	19.6	21.1	23.1	31.9	33.5

More simply, we get:

6.4 6.4 7.6 8.5 8.9 **9.9** **10.2** 10.5 10.6 11.4 11.6 12.2
13.6
13.7 14.0 14.4 14.8 15.6 **16.0** **17.8** 19.6 21.1 23.1 31.9 33.5

Since there are $n = 25$ (odd) data points, the median is the number in position $(n + 1)/2 = ((25) + 1)/2 = 13$. The data point in the 13th position is 13.6, so this is the median: $M = 13.6$. Since there are 12 (even) data points to the left of the median (in the ordered list), then we find the first quartile by averaging the center **two** numbers in the list of the first 12 data points. These are the values 9.9 and 10.2, so $Q_1 = \dfrac{9.9 + 10.2}{2} = 10.05$. Since there are 12 (even) data points to the right of the median (in the ordered list), then we find the third quartile by averaging the center **two** numbers in the list of the last 12 data points. These are the values 16.0 and 17.8, so $Q_3 = \dfrac{16.0 + 17.8}{2} = 16.9$.

The Five-Number Summary and Boxplots

Definition. The *five-number summary* of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols the five-number summary is:

$$\text{Minimum} \quad Q_1 \quad M \quad Q_3 \quad \text{Maximum}$$

Example S.2.4. Five-Numbered Stooges
What is the five-number summary for the slaps per film data of Example S.2.3?

Solution. The minimum number in the data is 6.4 and the maximum is 33.5, so the five-number summary is:

$$6.4 \quad 10.05 \quad 13.6 \quad 16.9 \quad 33.5$$

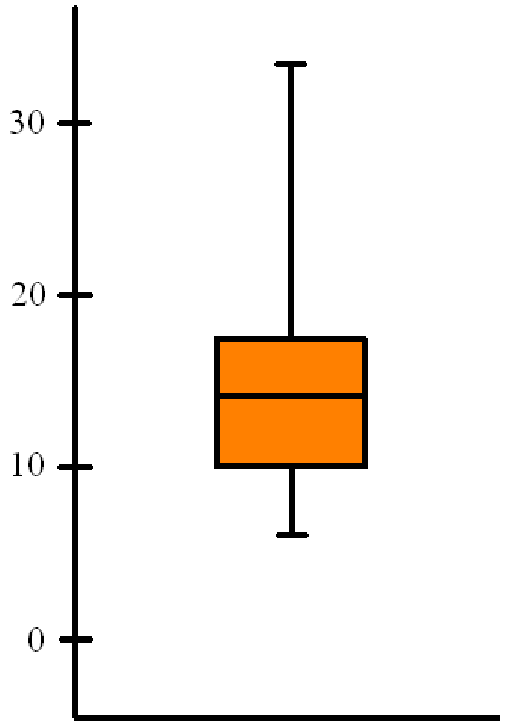
Definition. A **boxplot** is a graph of the five-number summary. It includes the properties:

- A central box spans the quartiles Q_1 and Q_3.
- A line in the box marks the median M.
- Lines extend from the box out to the smallest and largest observations.

Example S.2.5. Boxing Stooges.

Create a boxplot for the five-number summary of Example S.2.4.

Solution. We create a vertical axis calibrated with numerical values and then plot the five-number summary as indicated above to get:



Note. Sometimes it is helpful to plot two boxplots side-by-side to compare properties of different populations. See, for example, Figure 2.2 on page 46.

Spotting Suspected Outliers

Definition. The **interquartile range** is the distance between the first and third quartiles:

$$IQR = Q_3 - Q_1.$$

Note. The following is simply a "rule of thumb" for candidate outliers:

The 1.5 × IQR Rule for Outliers. Call an observation a suspected outlier if it falls more than $1.5 \times IQR$ above the third quartile of below the first quartile.

Example S.2.6. Stooge Outliers.

Does the slaps per film data of Example S.2.3 have any outliers, according to the $1.5 \times IQR$ rule?

Measuring Spread: The Standard Deviation

Note. A much more common description of the spread of a set of data is the standard deviation and (quoting Moore) “its close relative, the variance.”

Definition. The **variance** s^2 of a set of observations is an average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations $x_1, x_2, \ldots, x_n$ is

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

or, more compactly,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

The **standard deviation** s is the square root of the variance s^2:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

(Some texts call these the “sample variance” and “sample standard deviation — versus the population variance and standard deviation.)

Example. Example 2.7 page 48. Consider the data set:

1792 1666 1362 1614 1460 1867 1439

The mean is $\overline{x} = 1600$. We can calculate the variance as:

Observations x_i	Deviations $x_i - \overline{x}$	Squared Deviations $(x_i - \overline{x})^2$
1792	$1792 - 1600 = 192$	$192^2 = 36,864$
1666	$1666 - 1600 = 66$	$66^2 = 4,356$
1362	$1362 - 1600 = -238$	$(-238)^2 = 56,644$
1614	$1614 - 1600 = 14$	$14^2 = 196$
1460	$1460 - 1600 = -140$	$(-140)^2 = 19,600$
1867	$1867 - 1600 = 267$	$267^2 = 71,289$
1439	$1439 - 1600 = -161$	$(-161)^2 = 25,921$
	sum = 0	sum = 214,870

So the variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{6}(214,870) = 35,811.67.$$

The standard deviation is

$$s = \sqrt{35,811.67} = 189.24.$$

Note. Some properties of the standard deviation are:

- s measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- s (since it is the square root of something) is always greater than zero. $s = 0$ only when there is *no spread*. This happens only when all observations have the same value. Otherwise $s > 0$. As the observations become more spread out about their mean, s gets larger.
- s has the same units of measurement as the original observations.
- Like the mean $\overline{x}$, s is not resistant. A few outliers can make s very large.

Choosing Measures of Center and Spread

Note. The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution. Use $\overline{x}$ and s only for reasonably symmetric distributions that are free of outliers.

Organizing a Statistical Problem

Note. The "Four-Step Process" for organizing a statistical problem is:

State: What is the practical question, in the context of the real-world setting?

Formulate: What specific statistical operations does this problem call for?

Solve: Make the graphs and carry out the calculations needed for this problem.

Conclude: Give your practical conclusion in the setting of the real-world problem.

Example. Exercise 2.12 page 55.

rbg-12-13-2008