CHAPTER 2:
Describing Distributions with Numbers

Measuring center: the mean

The most common measure of center is the arithmetic average, or **mean.**

To find the **mean,** $\bar{x}$ (pronounced "x-bar"), of a set of observations, add their values and divide by the number of observations. If the *n* observations are x_1, $x_2, x_3, \ldots, x_n$, their mean is:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

or, in more compact notation

$$\bar{x} = \frac{1}{n} \sum x_i$$

Measuring center: the median

Because the mean cannot resist the influence of extreme observations, it is not a **resistant measure** of center.

Another common measure of center is the **median.**

The **median**, ***M***, is the midpoint of a distribution, the number such that half of the observations are smaller and the other half are larger.

To find the median of a distribution:

1. Arrange all observations from smallest to largest.
2. If the number of observations ***n* is odd**, the median *M* is the center observation in the ordered list. If the number of observations ***n* is even**, the median *M* is the average of the two center observations in the ordered list.
3. You can always locate the median in the ordered list of observations by counting up $(n + 1)/2$ observations from the start of the list.

Example 1

Here are the data: 5 4 3 2 6 2 3 4 8

1) The mean:

$$\bar{x} = \frac{5+4+3+2+5+2+3+4+8}{9} = \frac{36}{9} = 4$$

2) The median:

n = 9

a) arrange in increasing order: 2 2 3 3 4 4 5 6 8

b) n is odd, location of median = $\frac{9+1}{2} = 5th$ element in the sorted list so M = 4.

Example 2

Find the median for next data set: 4 6 7 2 8 3

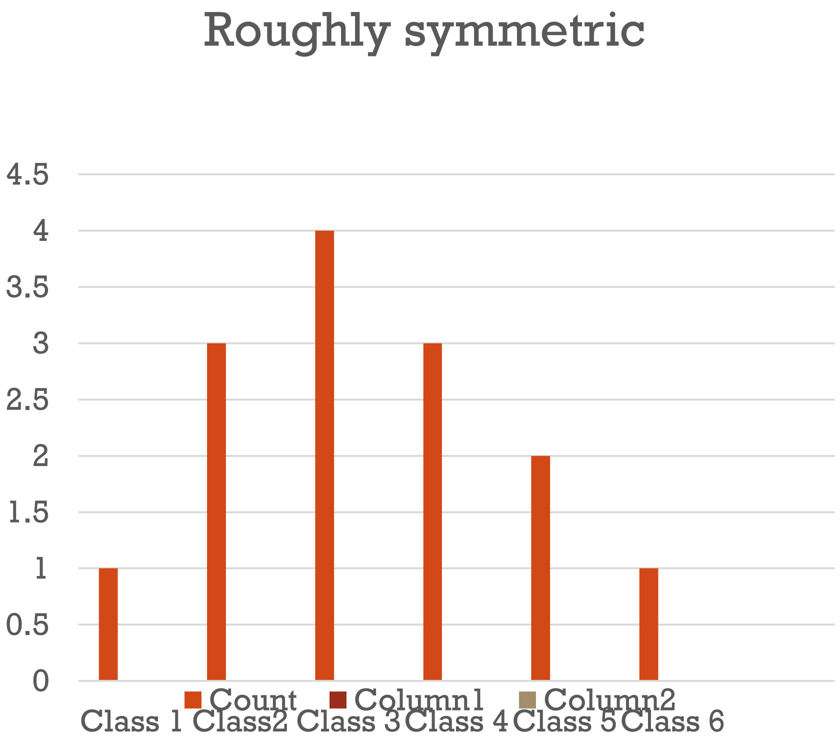
1) 2 3 4 6 7 8

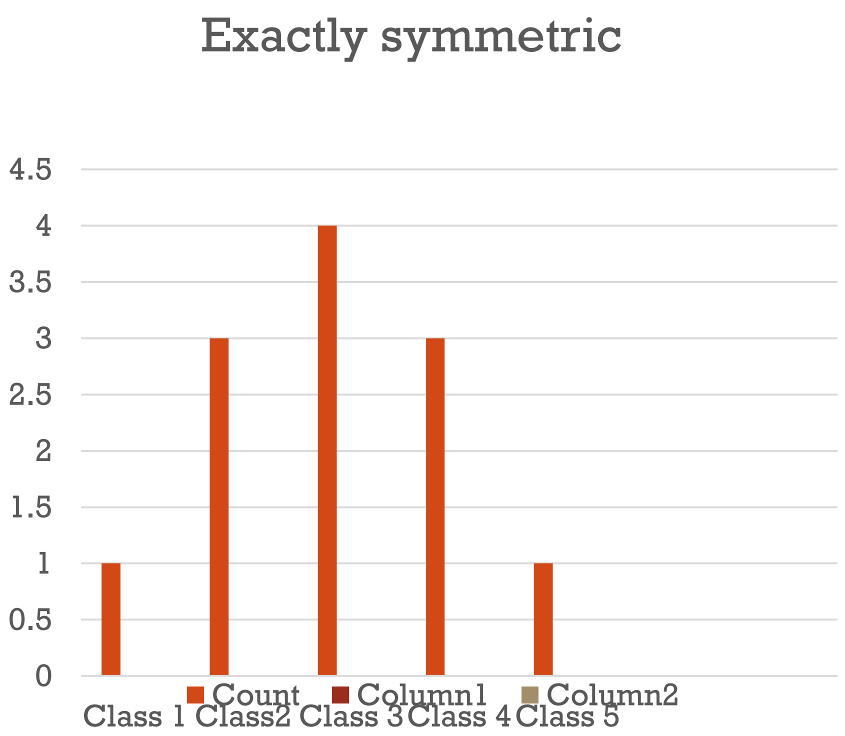
2) n = 6, n is even, location of median = $\frac{6+1}{2} = 3.5$, so median is a mean of 3rd and 4th values

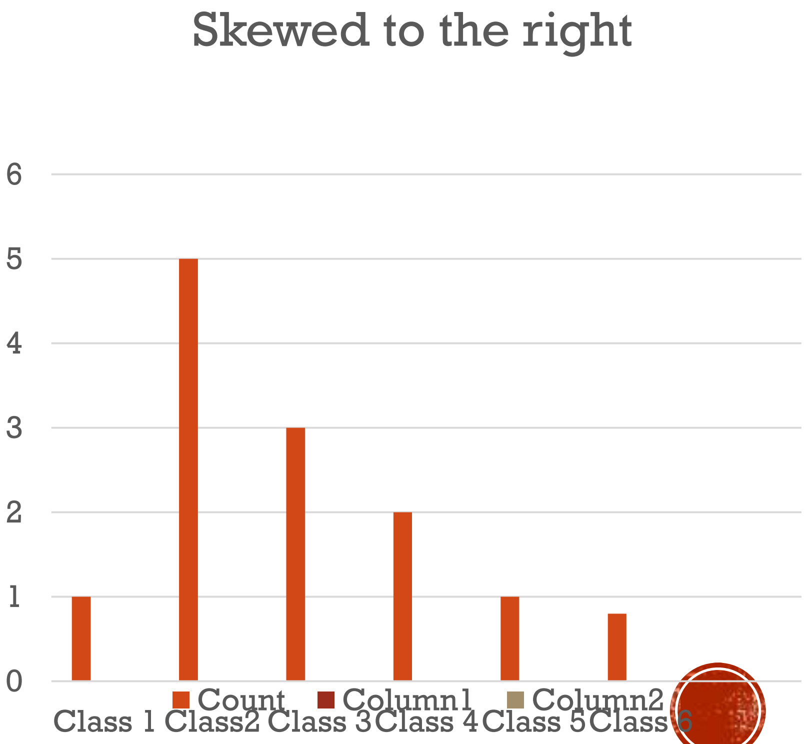
$$M = \frac{4+6}{2} = 5$$

Facts: if a distribution is:

1) Roughly symmetric – the mean and median are close together;
2) Exactly symmetric – the mean and median are exactly the same;
3) Skewed to the right or to the left – the mean is usually farther out in the long tail than is the median







Measuring spread: quartiles

- A measure of center alone can be misleading.
- A useful numerical description of a distribution requires both a measure of center *and a measure of spread*. We could look at the largest and smallest values (and we will!), but like the mean, they are (obviously) affected by extreme values—so we will examine other percentiles.

To calculate the quartiles:

- Arrange the observations in increasing order and locate the median *M*.
- The first quartile, Q_1, is the median of the observations located to the left of the median in the ordered list.
- The third quartile, Q_3, is the median of the observations located to the right of the median in the ordered list.

Measuring spread: the quartiles

The quartiles are the 3 points that divide the data set into four equal groups.

To calculate the quartiles:

1. Sort the data in increasing order.
2. Find the median of the data set. It will be the second quartile, so $Q_2 = M$.
3. The first quartile (lower quartile) Q_1 is the middle number (median) between the smallest number and the median of the data set.
4. The third quartile (upper quartile) Q_3 is the middle number (median) between the highest number and the median of the data set.

Example 3

Find quartiles for the data set: 7 12 5 2 9 10 1

1. 1 2 5 7 9 10 12, n is odd
2. $Q_2 = M = 7$
3. $Q_1 = 2$, median of the data set: 1 2 5
4. $Q_3 = 10$, median of the data set: 9 10 12

Example 4

Find quartiles for the data set: 1 3 7 1 10 10 10 13 8 1

1. 1 1 1 3 7 8 10 10 10 13, n is even
2. $Q_2 = M = \frac{7+8}{2} = 7{,}5$
3. $Q_1 = 1$, median of the data set: 1 1 1 3 7
4. $Q_3 = 10$, median of the data set: 8 10 10 10 13

Let's change the last number in the third example from 12 to 20, the quartiles will not change. The quartiles are resistant because they are not affected by a few extreme observations.

Five-number summary

- The minimum and maximum values alone tell us little about the distribution as a whole. Likewise, the median and quartiles tell us little about the tails of a distribution.
- To get a quick summary of both center and spread, combine all five numbers.
- 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.

Minimum $\quad Q_1 \quad M \quad Q_3 \quad$ Maximum

Boxplots

- The five-number summary divides the distribution roughly into quarters. This leads to a new way to display quantitative data, the boxplot.

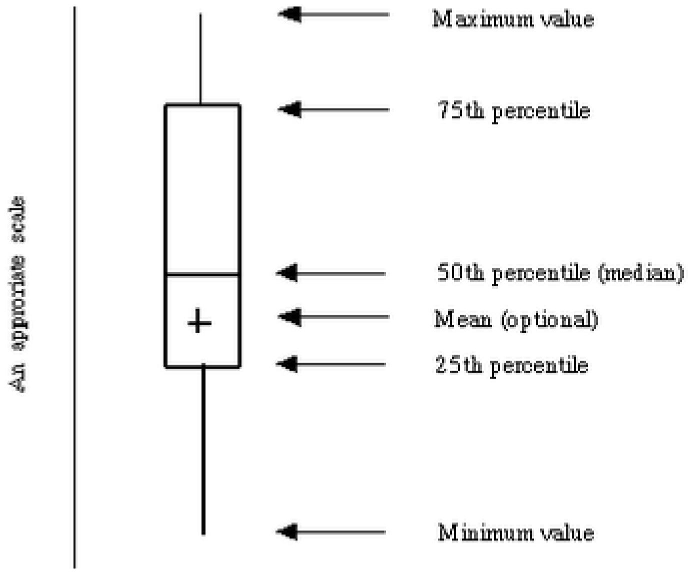
HOW TO MAKE A BOXPLOT

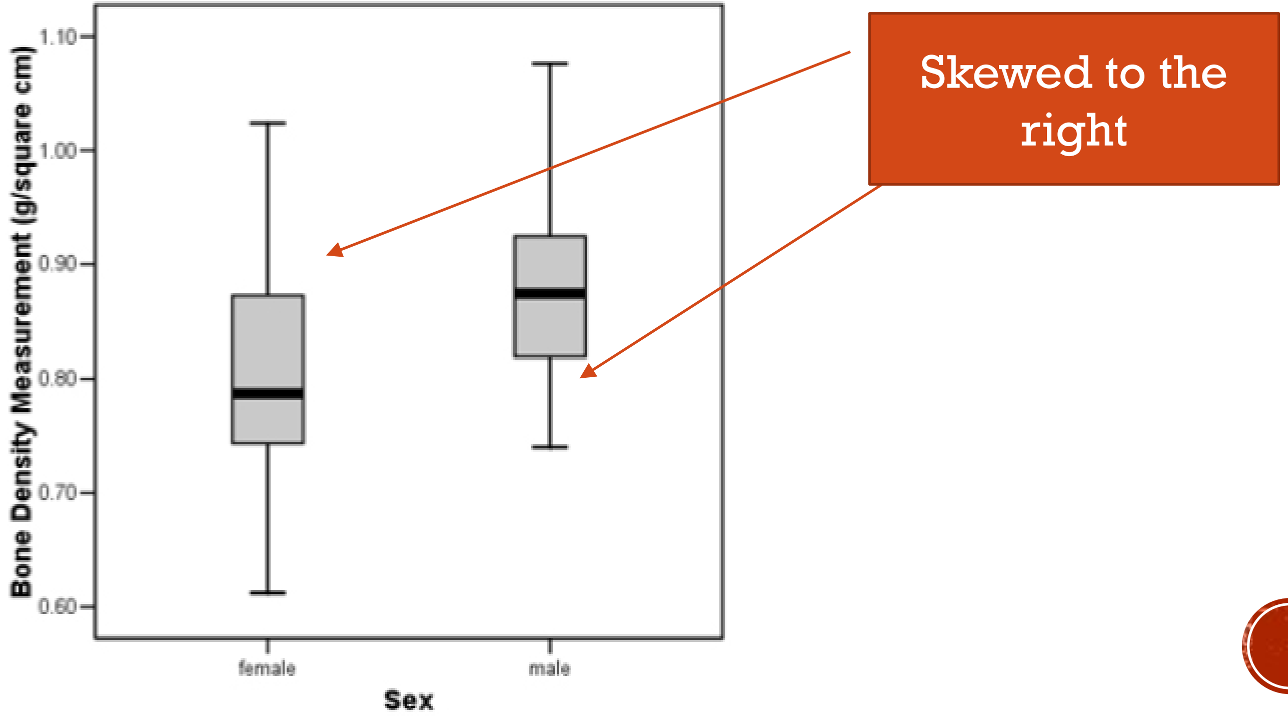
- A central box spans the quartiles Q_1 and Q_3.
- A line in the box marks the median *M*.
- Lines extend from the box out to the smallest and largest observations.

Boxplot is a graph of the five-number summary. How to make a boxplot?

1. A central box spans the lower and upper quartiles.
2. A line in the box marks the median.
3. Lines extend from the box out to the smallest and largest observations.

Boxplots show less detail than histograms or stemplots, so they are best used for side-by-side comparison of more than one distribution.





Spotting suspected outliers

The interquartile range IQR is the distance between the first and third quartiles,

$IQR = Q_3 - Q_1$

THE 1.5 × IQR RULE FOR OUTLIERS

Call an observation a suspected outlier if it falls more than 1.5 × IQR above the third quartile or below the first quartile.

Any values not falling between $Q_1 - (1.5\ x\ IQR)\ and\ Q_3 + (1.5\ x\ IQR)$ are flagged as suspected outliers. The 1.5 × IQR rule is not a replacement for looking at the data. It is most useful when large volumes of data are scanned automatically.

Example: find suspected outliers for these data set: 2 8 5 19 45 10

1) Sort: 2 5 8 10 19 45
2) Q_1 is 5, Q_3 is 19
3) $IQR = Q_3 - Q_1$

IQR = 19 – 5 = 14

Any values not falling between $Q_1 - (1.5\ x\ IQR)\ and\ Q_3 + (1.5\ x\ IQR)$ are flagged as suspected outliers.

$Q_1 - (1.5\ x\ IQR)$ = 5 – 1.5x14 = 5 – 21 = -16

$Q_3 + (1.5\ x\ IQR)$ = 19 + 1.5x14 =19 + 21 = 40

45 doesn’t fall in interval (-16, 40), so 45 is a suspected outlier.

Measuring spread: standard deviation

- The most common measure of spread looks at how far each observation is from the mean. This measure is called the standard deviation.
- 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 the *n* observations $x_1, x_2, x_3, \ldots, x_n$, is

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

Again, more briefly:

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

- The standard deviation, s, is the square root of the variance, s^2.

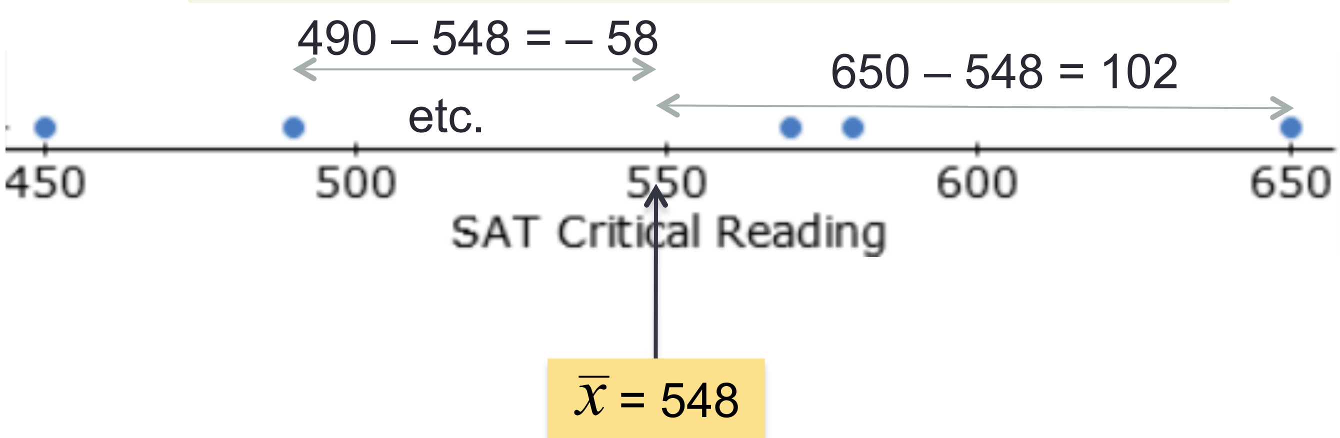
$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

Calculating the Standard Deviation

- **Example:** Consider the following data on the SAT critical reading scores for 5 Georgia Southern University freshman in 2010.

1) Calculate the mean.
2) Calculate each *deviation.*

   deviation = observation – mean



Example: find the standard deviation for these data set: 4 7 10 8 6

n = 5, here the mean:

$$\bar{x} = \frac{4+7+10+8+6}{5} = \frac{35}{5} = 7$$

$s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$ - standard deviation

$$s = \sqrt{\frac{(4-7)^2+(7-7)^2+(10-7)^2+(8-7)^2+(6-7)^2}{5-1}} = \sqrt{\frac{20}{4}} = \sqrt{5} = 2.24$$

Answer standard deviation is 2.24.

Here are the most important properties of the standard deviation:

1. s measures spread about the mean and should be used only when the mean is chosen as the measure of center.
2. s is always zero or 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.
3. s has the same units of measurement as the original observations.
4. Like the mean, s is not resistant. A few outliers can make s very large.

Choosing measures of center and spread

- We now have a choice between two descriptions for center and spread
  - mean and standard deviation
  - median and interquartile range

CHOOSING A SUMMARY

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