MEASURES OF CENTER

One Quantitative Variable

Unit 1: Exploratory Data Analysis

UF UNIVERSITY of FLORIDA

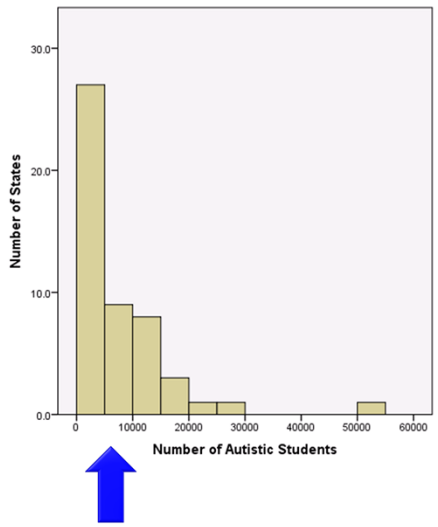
We will look the two most common and useful measures of center.

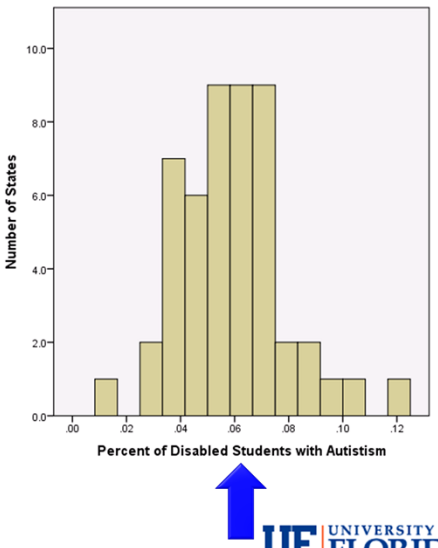
The sample mean and the sample median.

These two measures approach the question of “center” differently.

It’s important to understand the differences between these two measures of center.

Where is the "middle" of the distribution?





UF UNIVERSITY of FLORIDA

In our quest to find the middle or center of the distribution, we desire a measure of center that locates the bulk of the data.

Measures of Center: Sample Mean

- Average, sum of observations divided by total number
- Each observation is equally weighted in the calculation
- Sensitive to skewness and outliers
- For Data: $x_1, x_2, \ldots, x_n$
- The formula can be written mathematically as:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

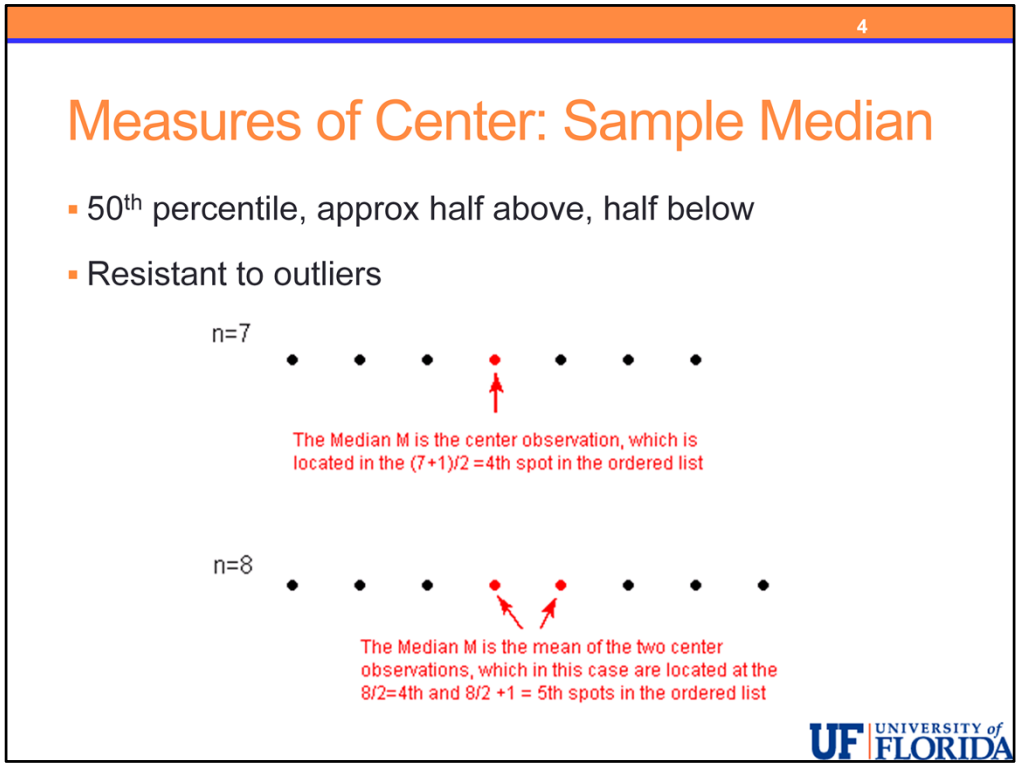
UF UNIVERSITY of FLORIDA

The sample mean is also commonly known as the average. To calculate the sample mean we add all of the observations and divide by the number of observations.

It is clear from this process that each observation is given an equal weight in the calculation. As we will see, this results in the sample mean being sensitive to skewness and outliers.

We use the symbol x with a line over it, called x-bar, to represent the sample mean.

If we write each observation as x-1, x-2, up to x-n then we can express the calculation in either of these forms where the capital greek letter sigma represents the summation, in this case, of the data points x-1 up to x-n.



The sample median is the value which splits the data in half (once it is sorted in order). Approximately half of the data points will be below this value and approximately half above.

If there are an odd number of observations, the sample median is the middle observation after sorting the data.

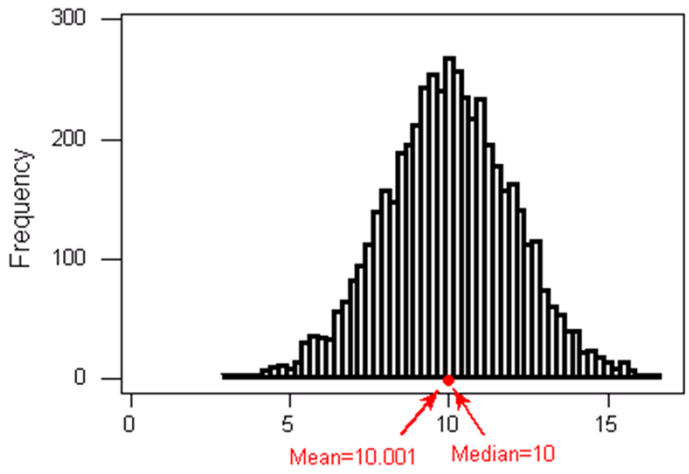
If there are an even number of observations, the sample median is taken to be the average of the two middle observations in the sorted data.

In the display here, the dots are not necessarily equally spaced and could even be equal, we simply illustrate the process regardless of the data values, as long as they are sorted.

The sample median can also be called the 50^{th} percentile or the 2^{nd} quartile. Percentiles and quartiles (along with the median) are measures of position and will be discussed in more detail later.

Because the magnitude of the observations is ignored in determining the sample median, only their relative positions are used, the sample median is very resistant or robust to skewness and outliers.

Measures of Center



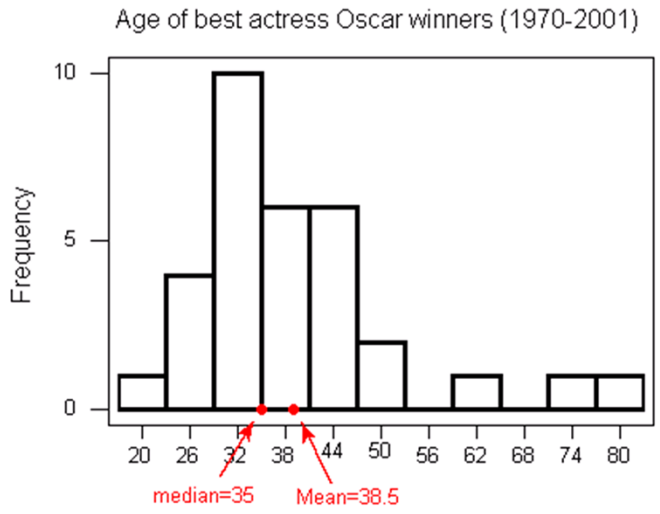
UF UNIVERSITY of FLORIDA

For symmetric distributions, the sample mean and sample median will be nearly identical.

We will see that in this case, we tend to use the sample mean, however, the sample median could also be used.

For symmetric distributions, both methods of defining the center are able to locate the bulk of the data easily.

Measures of Center



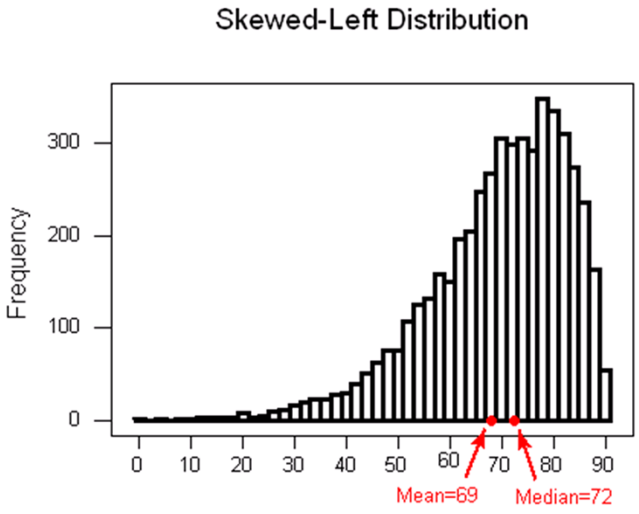
UF UNIVERSITY of FLORIDA

For a skewed distribution or even a symmetric distribution with outliers in only one direction, the sample mean will be pulled in the direction of the skewness or extreme observations.

Here for the skewed right distribution of age of best actress Oscar winners, the sample mean age is 38.5 whereas the sample median age is 35.

In trying to locate the bulk of the data the sample median tends to do a better job than the sample mean.

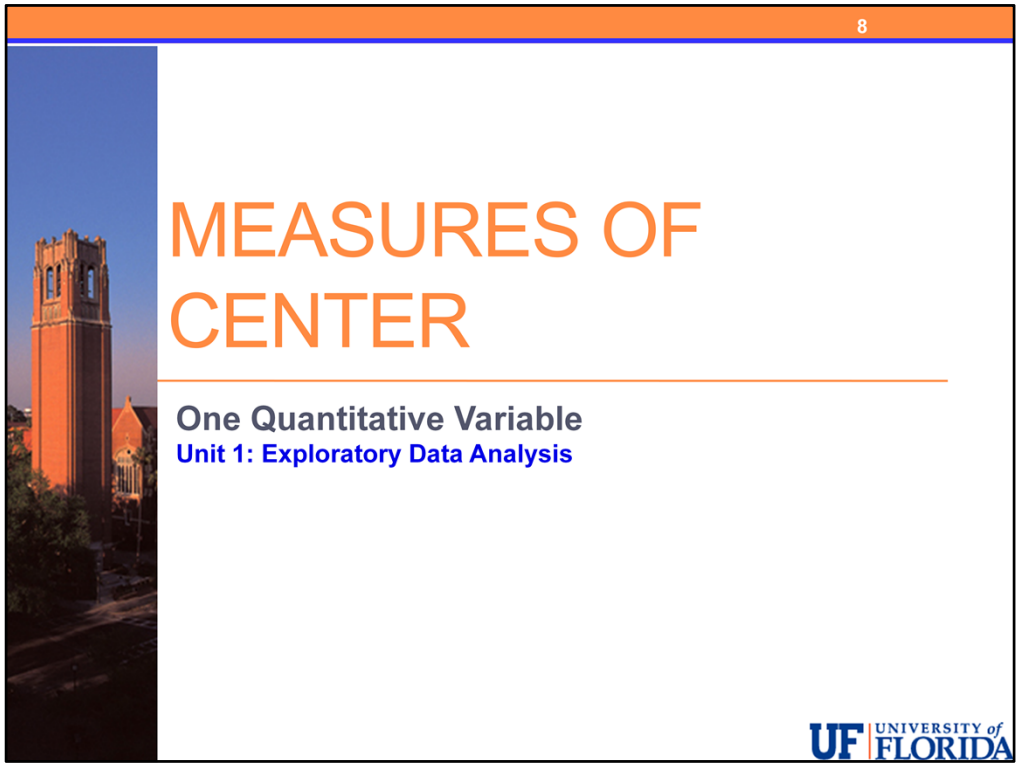
Measures of Center



UF UNIVERSITY of FLORIDA

Similarly for a skewed left distribution, the sample mean is pulled down. In this distribution, the sample mean is 69 whereas the sample median is 72. Again, the sample median seems to do a better job of locating the bulk of the data.

The general recommendation is that it is fine to use the sample mean in situations which are symmetric with no outliers. In other cases, one should report the sample median or possibly both the sample median and sample mean.



Remember our broad goal of examining and describing the distribution of one quantitative variable.

We have introduced the main features of shape, center, spread, and outliers.

We have discussed shape, including symmetry, skewness, and modality. And now we have covered the two measures of center, the sample mean or average and the sample median.

One goal of numerical measures such as the sample mean and sample median is to quantify aspects of the distribution we see in our graphical displays. Next we will discuss measures of spread.