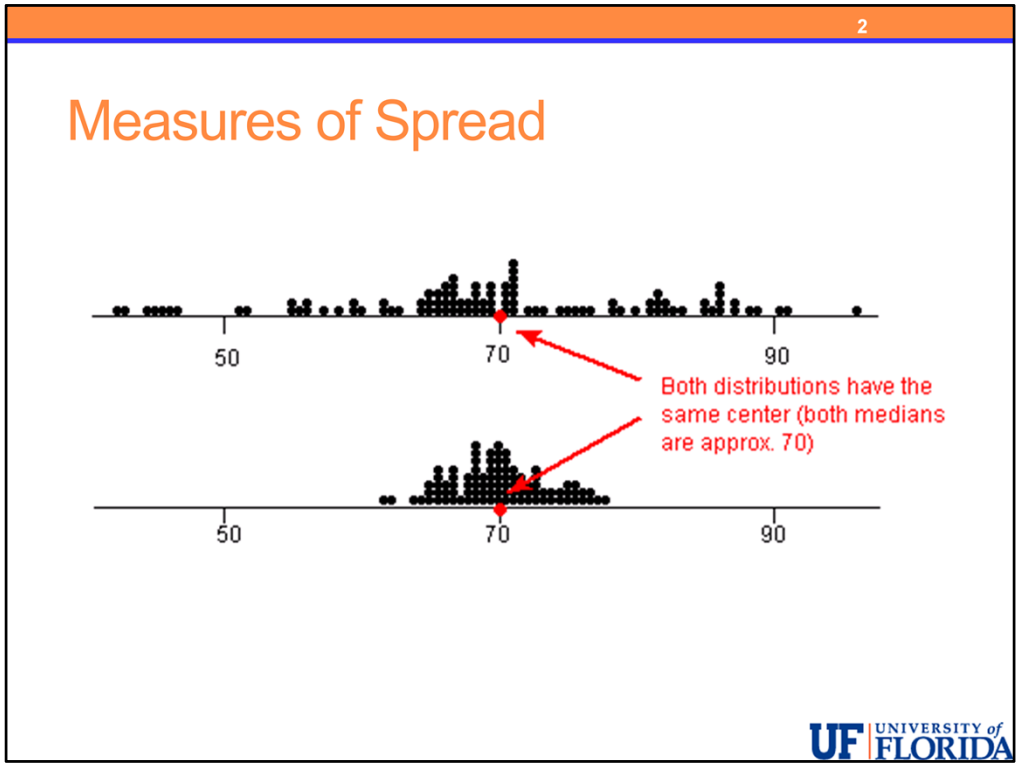
MEASURES OF SPREAD

One Quantitative Variable

Unit 1: Exploratory Data Analysis

We will look the three common and useful measures of spread.

The Range, the Inter-Quartile Range (or IQR), and the Standard Deviation (which we usually denote by a lower case s).



A measure of the center of the distribution is clearly not enough to describe the full picture.

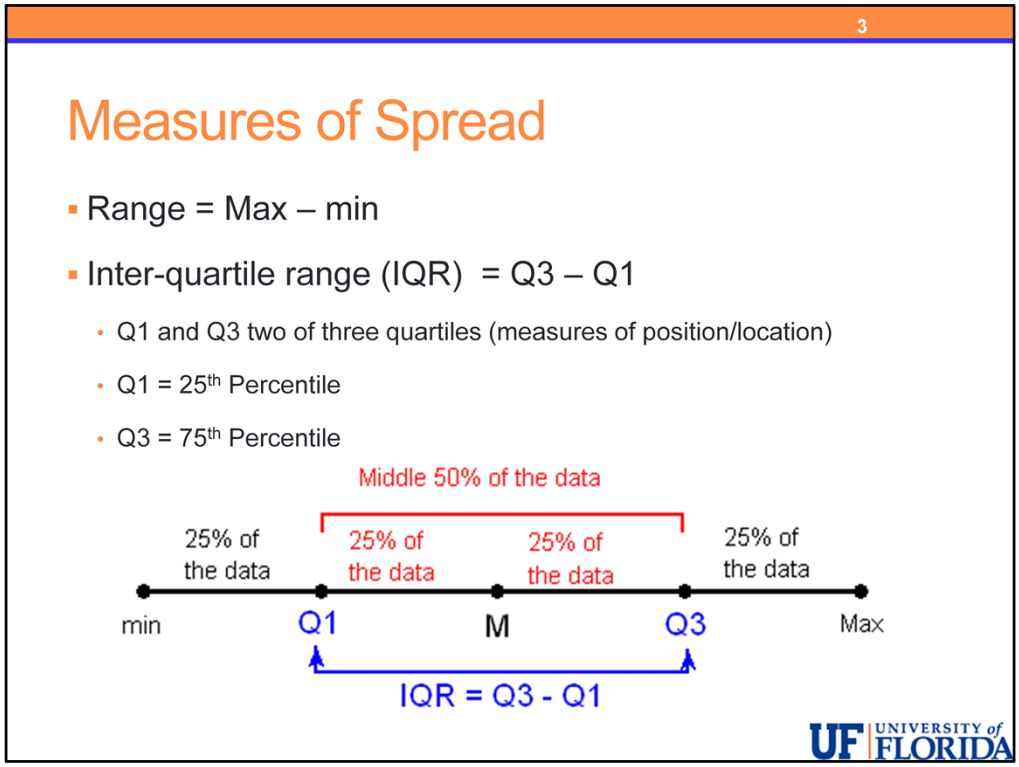
These two distributions have the same center but very different spread!

Although our two measures of center differed, they do attempt to measure the same point in the distribution and thus are comparable.

The three measures of spread, however, provide very different ways to quantify the variability of the distribution.

They do not try to estimate the same quantity, instead they provide information about three different aspects of the spread of the distribution.

When we use these three measure together, we can obtain a more complete picture of the spread of the distribution.



The range is the Maximum value minus the minimum value and gives the full extent of the range of observations. Notice that the range is one number, the difference between the Maximum and the minimum.

For example in the best actress Oscar winners data, we might say "the values ranged from 21 to 80," however, the range is 80-21 = 59 years.

The Inter-Quartile Range (or sometimes Inner-Quartile Range) – we will use the notation IQR – is defined as the difference between the 3rd quartile and the 1st quartile.

Q1 is the 1st quartile which is the 25th percentile. Approximately 25% are below Q1 and approximately 75% are above.

Q3 is the 3rd quartile which is the 75th percentile. Approximately 75% are below Q3 and approximately 25% are above.

We have also already defined the 2nd quartile, Q2. It is the median or 50th percentile with approximately 50% above and approximately 50% below the median.

The five values: minimum, Q1, Median, Q3, and Maximum make up what is commonly called "the five-number summary". Each of these values represents a measure of position or location.

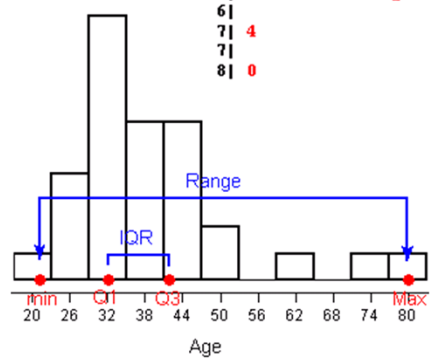
We won't be concerned with calculating Q1 and Q3 by hand. You should be able to find the mean, median, minimum, Maximum, and range by hand for a small dataset. Given the quartiles, you should be able to calculate the IQR.

Notice that the IQR gives the range of the middle 50% of the data.

Age of Best Actress Oscar Winners

- Median = 35 (Mean = 38.5)
- Range = 80 – 21 = 59
- Q1 = 32
- Q3 = 41.5
- IQR = 41.5 - 32 = 9.5

```
2| 1
2| 56669          Bottom Half
3| 013333444
3| 555789
4| 11123
4| 599
5|
5|
6| 1              Top half
6|
7| 4
7|
8| 0
```



Software packages use different formulas to calculate the quartiles Q1 and Q3

UF UNIVERSITY of FLORIDA

Here is our example of the Age of best actress Oscar winners with all of our summary measures so far.

We have a stemplot and histogram.

The median was 35 and the mean was 38.5.

The range is 80 – 21 which is 59 years. This tells us the full length of the data is 59 years.

Q1 is 32, Q3 is 41.5 which gives and IQR of 41.5 – 32 or 9.5. This tells us the length of the middle 50% of the data is 9.5 years.

Here we see these values illustrated on the histogram.

Values for Q1 and Q3 given by different software packages for a given dataset may vary as there are numerous acceptable methods for estimated these values from data.

Standard Deviation

- Quantify spread of distribution by measuring how far observations are from their mean
- Gives the average (or typical distance) between a data point and the mean
- SD or s

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

The final measure of spread is the standard deviation.

This measure is based upon the deviations of each value from the mean or average value.

The interpretation of the standard deviation is as the typical or average distance between observed data values and the sample mean.

We will not calculate the standard deviation by hand, however, you should review the examples presented in the course materials to get a full understanding of the calculation.

To calculate the standard deviation, we take each observation and subtract the sample mean, x-bar.

We square each of the differences, and add these squared deviations.

We divide that total by n – 1 where n is the number of observations.

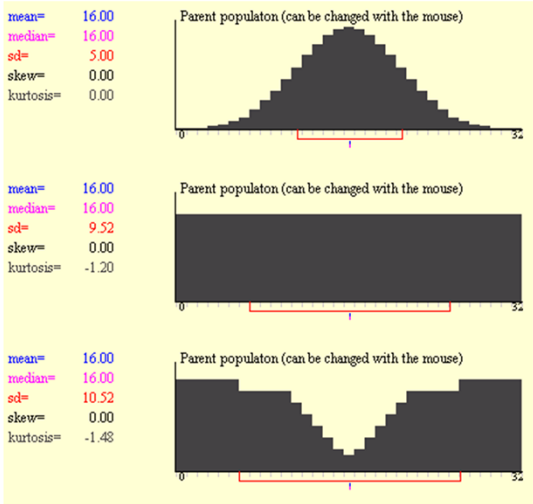
At this stage we have what is called the variance, or s-squared. This is sometimes used in statistical methods, however, it is in units which are the square of the original units which makes it difficult to interpret.

We take the square root to obtain the final result for the standard deviation, s. This value will have the units of the original data and have the interpretation of the typical distance an

observation differs from the sample mean.

Because each value is weighted equally, the standard deviation is influenced by outliers and extreme values. In some scenarios, the standard deviation is less reliable than the mean in this regard and should be used with caution for highly skewed distribution or distributions with extreme outliers. Even moderate skewness and relatively mild outliers can have a dramatic impact on the standard deviation.

Standard Deviation



http://onlinestatbook.com/stat_sim/sampling_dist/index.html

UF UNIVERSITY of FLORIDA

The following three distributions were created using the sampling distribution applet we will use later in the course.

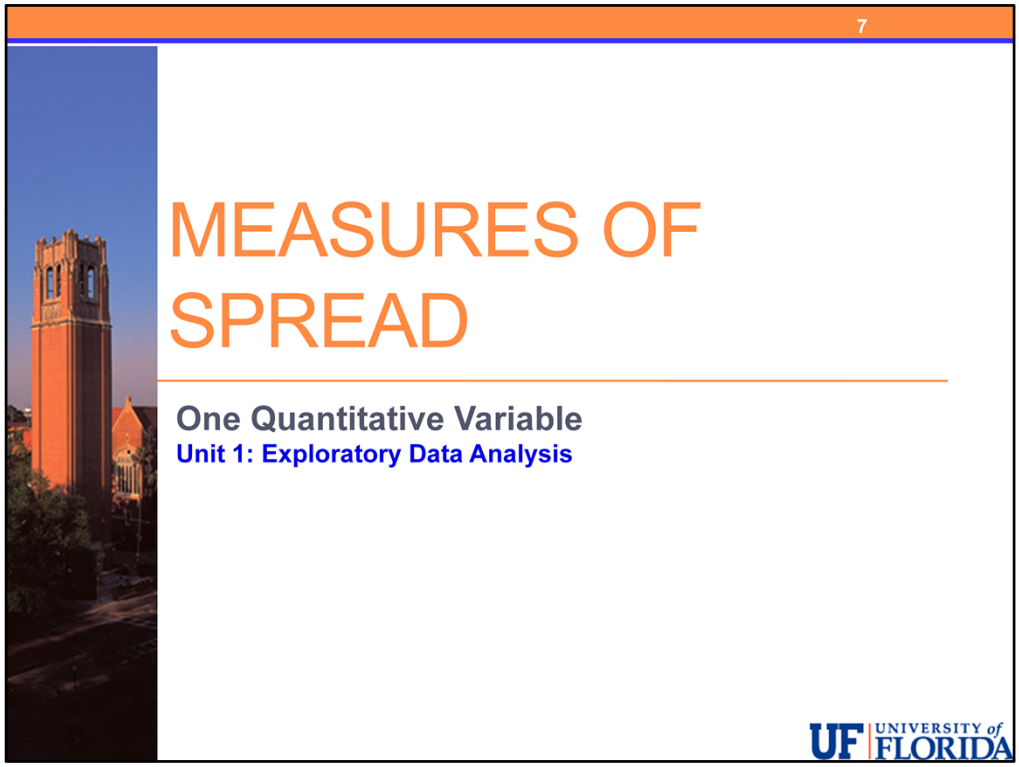
Note that all three distributions have the same mean and median and are perfectly symmetric - you could also create a similar comparison for skewed distributions using this applet.

The more centrally located the values, the lower the standard deviation. The nice “normal” distribution on top, has a standard deviation of 5.

The completely uniform distribution in the middle, has a standard deviation of around 9.5 and the strange V-shaped or inverted distribution has a standard deviation of around 10.5.

Measures of spread are most useful for comparing distributions on the same or similar scales and for scaling measures of position for more general comparisons and decision making.

We will see when we discuss sampling variability, sampling distributions, and inferential statistics that measure of spread play a key role in this analysis.



Again, remember our broad goal of examining and describing the distribution of one quantitative variable.

Now we have discussed shape, center, and spread.

To finish our discussion on exploratory data analysis for one quantitative variable, we need to discuss a few additional topics.

Measures of position, outliers, boxplots, and an introduction to the normal distribution.