Lecture 3:
Chapter 3

C C Moxley

UAB Mathematics

26 January 16

Statistics involves describing data sets and inferring things about them. The first step in understanding a set of data is often to find its 'center.'

§3.2 Measurements of Center

Statistics involves describing data sets and inferring things about them. The first step in understanding a set of data is often to find its ‘center.’

We will discuss the following measurements of center:

§3.2 Measurements of Center

Statistics involves describing data sets and inferring things about them. The first step in understanding a set of data is often to find its 'center.'

We will discuss the following measurements of center:

- mean

§3.2 Measurements of Center

Statistics involves describing data sets and inferring things about them. The first step in understanding a set of data is often to find its ‘center.’

We will discuss the following measurements of center:

- mean
- median

§3.2 Measurements of Center

Statistics involves describing data sets and inferring things about them. The first step in understanding a set of data is often to find its 'center.'

We will discuss the following measurements of center:

- mean
- median
- mode

§3.2 Measurements of Center

Statistics involves describing data sets and inferring things about them. The first step in understanding a set of data is often to find its 'center.'

We will discuss the following measurements of center:

- mean
- median
- mode
- midrange

§3.2 Measurements of Center

Statistics involves describing data sets and inferring things about them. The first step in understanding a set of data is often to find its 'center.'

We will discuss the following measurements of center:

- mean
- median
- mode
- midrange
- weighted mean

§3.2 Mean

Definition (Mean)

The mean (or average) of a dataset is given by $\dfrac{\sum\limits_{i=1}^{n} x_i}{n}$.

§3.2 Mean

Definition (Mean)

The mean (or average) of a dataset is given by $\dfrac{\sum_{i=1}^{n} x_i}{n}$.

Essentially, you simply add up all the values in the data set and divide by the number of values in the data set.

§3.2 Mean

Definition (Mean)

The mean (or average) of a dataset is given by $\dfrac{\sum_{i=1}^{n} x_i}{n}$.

Essentially, you simply add up all the values in the data set and divide by the number of values in the data set.

Example (Mean)

A sample of 10 fun sized candy bags showed that each bag had the following number of candies: 12, 12, 12, 13, 14, 14, 15, 15, 16, 16. What's the average number of candies per bag for this sample?

§3.2 Mean

Definition (Mean)

The mean (or average) of a dataset is given by $\dfrac{\sum_{i=1}^{n} x_i}{n}$.

Essentially, you simply add up all the values in the data set and divide by the number of values in the data set.

Example (Mean)

A sample of 10 fun sized candy bags showed that each bag had the following number of candies: 12, 12, 12, 13, 14, 14, 15, 15, 16, 16. What's the average number of candies per bag for this sample?

$$m = \frac{12+12+12+13+14+14+15+15+16+16}{10} = 13.9.$$

§3.2 Median

Definition (Median)

The median of a dataset is the datum point (or the average of two consecutive data points) which has an equal number of data points above and below it.

§3.2 Median

Definition (Median)

The median of a dataset is the datum point (or the average of two consecutive data points) which has an equal number of data points above and below it.

To find the median, order your data points and find the number in the middle - average two consecutive data points if necessary, i.e. if there are an even number of data points.

§3.2 Median

Definition (Median)

The median of a dataset is the datum point (or the average of two consecutive data points) which has an equal number of data points above and below it.

To find the median, order your data points and find the number in the middle - average two consecutive data points if necessary, i.e. if there are an even number of data points.

Example (Median)

A sample of 10 fun sized candy bags showed that each bag had the following number of candies: 12, 12, 12, 13, 14, 14, 15, 15, 16, 16. What's the median number of candies in a bag?

§3.2 Median

Definition (Median)

The median of a dataset is the datum point (or the average of two consecutive data points) which has an equal number of data points above and below it.

To find the median, order your data points and find the number in the middle - average two consecutive data points if necessary, i.e. if there are an even number of data points.

Example (Median)

A sample of 10 fun sized candy bags showed that each bag had the following number of candies: 12, 12, 12, 13, 14, 14, 15, 15, 16, 16. What's the median number of candies in a bag? These are already ordered! Because there is an even number, the median is the average of the 5th and 6th data points, i.e. it is the average of 14 and 14, which is 14.

§3.2 Mode

Definition (Mode)

The mode of a dataset is the value which occurs the most number of times.

§3.2 Mode

Definition (Mode)

The mode of a dataset is the value which occurs the most number of times.

A dataset could be multimodal, bimodal, monomodal, or could have no modes. In what situations to these arise?

§3.2 Mode

Definition (Mode)

The mode of a dataset is the value which occurs the most number of times.

A dataset could be multimodal, bimodal, monomodal, or could have no modes. In what situations to these arise?

Example (Mode)

A sample of 10 fun sized candy bags showed that each bag had the following number of candies: 12, 12, 12, 13, 14, 14, 15, 15, 16, 16. What's the mode of this data set?

§3.2 Mode

Definition (Mode)

The mode of a dataset is the value which occurs the most number of times.

A dataset could be multimodal, bimodal, monomodal, or could have no modes. In what situations to these arise?

Example (Mode)

A sample of 10 fun sized candy bags showed that each bag had the following number of candies: 12, 12, 12, 13, 14, 14, 15, 15, 16, 16. What's the mode of this data set? The mode is 12.

§3.2 Midrange and Weighted Average

- The midrange is the point directly between the lowest and highest data points, i.e. it is the average of the max and the min of a data set.

§3.2 Midrange and Weighted Average

- The midrange is the point directly between the lowest and highest data points, i.e. it is the average of the max and the min of a data set.
- A weighted average places greater or smaller significance on certain data points. (An example would be GPA calculation.)

§3.3 Measures of Variation

The following sets of data both have the same number of data points and the same mean, median, mode, and midrange. But they obviously have a significant difference. This difference is characterized by **variance**

§3.3 Measures of Variation

The following sets of data both have the same number of data points and the same mean, median, mode, and midrange. But they obviously have a significant difference. This difference is characterized by **variance**

1, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 9.

The following sets of data both have the same number of data points and the same mean, median, mode, and midrange. But they obviously have a significant difference. This difference is characterized by **variance**

1, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 9.

1, 1, 2, 2, 3, 3, 4, 5, 5, 5, 5, 6, 7, 7, 8, 8, 9, 9.

§3.3 Range

Definition (Range)

The range of a set of data is the difference between the largest and smallest values.

§3.3 Range

Definition (Range)

The range of a set of data is the difference between the largest and smallest values.

The range is extremely sensitive to outliers.

§3.3 Range

Definition (Range)

The range of a set of data is the difference between the largest and smallest values.

The range is extremely sensitive to outliers.

Example (Range)

What is the range of the data set 1, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 9?

§3.3 Range

Definition (Range)

The range of a set of data is the difference between the largest and smallest values.

The range is extremely sensitive to outliers.

Example (Range)

What is the range of the data set 1, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 9? It's $9 - 1 = 8$.

§3.3 Standard Deviation

Definition (Standard Deviation)

The **standard deviation** of a set of data points is given by

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

§3.3 Standard Deviation

Definition (Standard Deviation)

The **standard deviation** of a set of data points is given by

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

The standard deviation measures how far the data points vary from the **mean**.

§3.3 Standard Deviation

Definition (Standard Deviation)

The **standard deviation** of a set of data points is given by

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

The standard deviation measures how far the data points vary from the **mean**. When is it zero? Is it ever negative?

Example (Calculate Standard Deviation)

Use 1, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 9.

Example (Calculate Standard Deviation)

Use 1, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 9. Well, $\sum_{i=1}^{n} x_i^2 = 492$ and $\bar{x} = 5$, so we have

$$s = \sqrt{\frac{18(492) - 90^2}{18(17)}} \approx 1.57$$

Example (Calculate Standard Deviation)

Use 1, 1, 2, 2, 3, 3, 4, 5, 5, 5, 5, 6, 7, 7, 8, 8, 9, 9.

§3.3. Standard Deviation

Example (Calculate Standard Deviation)

Use 1, 1, 2, 2, 3, 3, 4, 5, 5, 5, 5, 6, 7, 7, 8, 8, 9, 9. Well, $\sum_{i=1}^{n} x_i^2 = 568$ and $\bar{x} = 5$, so we have

$$s = \sqrt{\frac{18(568) - 90^2}{18(17)}} \approx 2.63$$

§3.3. Standard Deviation

How is the standard deviation useful?

- *When means are similar*, you can use the standard deviation to see differences in variation in samples.
- The standard deviation is less sentitive than range for measuring variation.
- As a general rule of thumb, you should expect to see about 95% of all data points falling within 2 standard deviations of the mean.

§3.3. Population Standard Deviation

Definition (Standard Deviation)

The **population standard deviation** of a set of data points is given by

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

§3.3. Population Standard Deviation

Definition (Standard Deviation)

The **population standard deviation** of a set of data points is given by

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

What are the differences between population standard deviation and standard deviation?

§3.3. Variance

Definition (Variance)

The **variance** of a set of data points is simply the square of the standard deviation.

§3.3. Variance

Definition (Variance)

The **variance** of a set of data points is simply the square of the standard deviation.

The sample variance is denoted s^2 while the population variance is denoted σ^2.

§3.3. Variance

Definition (Variance)

The **variance** of a set of data points is simply the square of the standard deviation.

The sample variance is denoted s^2 while the population variance is denoted σ^2.

- Variance is more sensitive to outliers than is standard deviation.

§3.3. Variance

Definition (Variance)

The **variance** of a set of data points is simply the square of the standard deviation.

The sample variance is denoted s^2 while the population variance is denoted σ^2.

- Variance is more sensitive to outliers than is standard deviation.
- Variance carries the square of the units of the data it describes.

§3.3. Variance

Definition (Variance)

The **variance** of a set of data points is simply the square of the standard deviation.

The sample variance is denoted s^2 while the population variance is denoted σ^2.

- Variance is more sensitive to outliers than is standard deviation.
- Variance carries the square of the units of the data it describes.
- Variance is an unbiased estimator while standard deviation is a biased estimator (in the sense described later).

§3.3. Variance

Definition (Variance)

The **variance** of a set of data points is simply the square of the standard deviation.

The sample variance is denoted s^2 while the population variance is denoted σ^2.

- Variance is more sensitive to outliers than is standard deviation.
- Variance carries the square of the units of the data it describes.
- Variance is an unbiased estimator while standard deviation is a biased estimator (in the sense described later).
- Variance is always non-negative.

§3.3. Bias

An estimator (statistic) which is **biased** tends to systematically over- or underestimate the parameter to which it corresponds.

§3.3. Bias

An estimator (statistic) which is **biased** tends to systematically over- or underestimate the parameter to which it corresponds. The standard deviation systematically underestimates the population standard deviation whereas the variance does not systematically under- or overestimate the population variance.

§3.3. Bias

An estimator (statistic) which is **biased** tends to systematically over- or underestimate the parameter to which it corresponds. The standard deviation systematically underestimates the population standard deviation whereas the variance does not systematically under- or overestimate the population variance. For certain distributions, these systematic biases can be compensated for.

§3.3. Chebyshev Theorem and Bell-Shaped Distributions

Theorem (Chebyshev's Theorem)

The proportion of any set of data lying within K standard deviations of the mean is always at least $1 - 1/K^2$, where $K > 1$.

§3.3. Chebyshev Theorem and Bell-Shaped Distributions

Theorem (Chebyshev's Theorem)

The proportion of any set of data lying within K standard deviations of the mean is always at least $1 - 1/K^2$, where $K > 1$.

Theorem (The Empirical Rule)

For bell-shaped distributions, we have that

- $\approx 68\%$ *of all values lie within* s *of* m*,*

§3.3. Chebyshev Theorem and Bell-Shaped Distributions

Theorem (Chebyshev's Theorem)

The proportion of any set of data lying within K standard deviations of the mean is always at least $1 - 1/K^2$, where $K > 1$.

Theorem (The Empirical Rule)

For bell-shaped distributions, we have that

- $\approx 68\%$ *of all values lie within s of m,*
- $\approx 95\%$ *of all values lie within $2s$ of m, and*

§3.3. Chebyshev Theorem and Bell-Shaped Distributions

Theorem (Chebyshev's Theorem)

The proportion of any set of data lying within K standard deviations of the mean is always at least $1 - 1/K^2$, where $K > 1$.

Theorem (The Empirical Rule)

For bell-shaped distributions, we have that

- $\approx 68\%$ *of all values lie within s of m,*
- $\approx 95\%$ *of all values lie within $2s$ of m, and*
- $\approx 99.7\%$ *of all values lie within $3s$ of m.*

§3.3. Coefficient of Variation

When comparing the variation of two different types of random variales, you can use the coefficient of variation.

§3.3. Coefficient of Variation

When comparing the variation of two different types of random variales, you can use the coefficient of variation.

Definition (Coefficient of Variation)

$$CV = \frac{s}{\bar{x}} \cdot 100\% \qquad CV = \frac{\sigma}{\mu} \cdot 100\%$$

Example (Comparing apples and oranges)

Apples weigh an average of 7 ounces with a standard deviation of 2.1 ounces. Oranges have an average volume of 190 mL with a standard deviation of 2.1 mL.

§3.3. Coefficient of Variation

When comparing the variation of two different types of random variales, you can use the coefficient of variation.

Definition (Coefficient of Variation)

$$CV = \frac{s}{\bar{x}} \cdot 100\% \qquad CV = \frac{\sigma}{\mu} \cdot 100\%$$

Example (Comparing apples and oranges)

Apples weigh an average of 7 ounces with a standard deviation of 2.1 ounces. Oranges have an average volume of 190 mL with a standard deviation of 2.1 mL. Because $CV_A \approx 30\%$ whereas $CV_O \approx 1.1\%$, we know that apples vary in weight much more than oranges vary in volume.

§3.4 Measures of Relative Standing

How do we tell where a piece of data 'fits' into the larger data set?

§3.4 Measures of Relative Standing

How do we tell where a piece of data 'fits' into the larger data set? We use measures of **relative standing**.

§3.4 Measures of Relative Standing

How do we tell where a piece of data 'fits' into the larger data set? We use measures of **relative standing**.

Definition (z-score)

The z-score for a datum is a how many standard deviations it is above or below the mean.

§3.4 Measures of Relative Standing

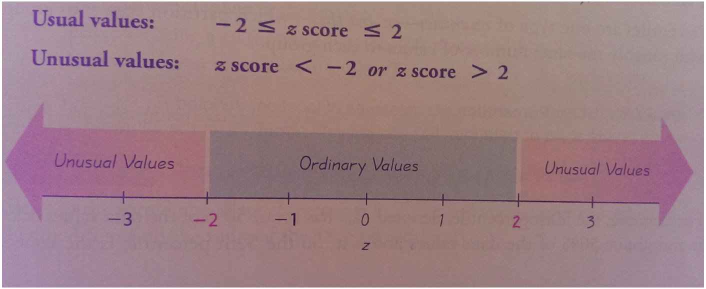
How do we tell where a piece of data 'fits' into the larger data set? We use measures of **relative standing**.

Definition (z-score)

The z-score for a datum is a how many standard deviations it is above or below the mean.

$$z = \frac{x - \bar{x}}{s}, \qquad z = \frac{x - \mu}{\sigma}$$

‘Typical’ values have a z-score of absolute value less than or equal to 2.



§3.4 Percentiles/Quantiles/Quartiles

Definition (Percentile)

Percentiles divide a dataset into 100 parts (P_1, P_2, ..., P_{100}) which each have about 1% of the total data points.

Definition (Percentile)

Percentiles divide a dataset into 100 parts (P_1, P_2, ..., P_{100}) which each have about 1% of the total data points.

Saying that something is in the 35th percentile means, roughly, that 35% of the values fall below it.

§3.4 Percentiles/Quantiles/Quartiles

Definition (Percentile)

Percentiles divide a dataset into 100 parts (P_1, P_2, ..., P_{100}) which each have about 1% of the total data points.

Saying that something is in the 35th percentile means, roughly, that 35% of the values fall below it.

$$\text{percentile value of } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

§3.4 Percentiles/Quantiles/Quartiles

Definition (Percentile)

Percentiles divide a dataset into 100 parts (P_1, P_2, ..., P_{100}) which each have about 1% of the total data points.

Saying that something is in the 35th percentile means, roughly, that 35% of the values fall below it.

$$\text{percentile value of } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

First, second, and third quartiles are P_{25}, P_{50}, and P_{75}, respectively.

Briefly describe interquartile range, semi-interquartile range, midquartile, and 10-90 percentile range.

§3.4 Percentiles/Quantiles/Quartiles

Briefly describe interquartile range, semi-interquartile range, midquartile, and 10-90 percentile range. Discuss boxplots.