DESCRIPTIVE STATISTICS

MAT 152 – Statistical Methods I

Lecture 4

Instructor: Dustin Roten

Fall 2020

Measuring the “Spread” of the Data

A simple way to measure the **spread** of data is to calculate its **range**.

$$\text{Range} = \max - \min$$

Consider the data:

$$1, 1, 2, 3, 4, 7, 8, 8, 10$$
$$\text{Range} = 10 - 1 = 9$$

Calculating the range of the data provides information about the overall spread of the data. However, no information is given about the spread <u>within</u> the data.

Measuring the “Spread” of the Data

Consider the following data:

3, 5.2, 7, 9, 10.5, 11.2

This dataset’s **location** and **center** can be described using the values of **min**, **max**, **quartiles**, **mean**, **median**, and **mode**.

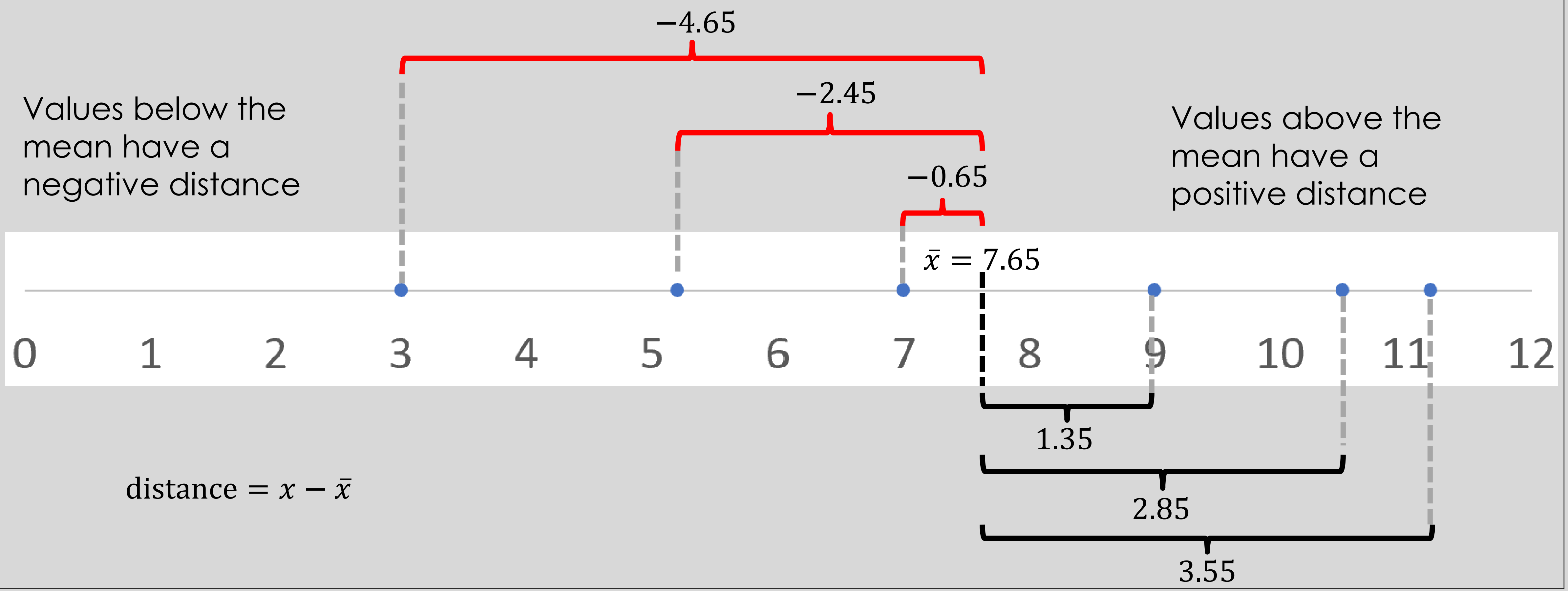
But, how can the **spread** be measured?

First, the mean of the data can be calculated:

$$\bar{x} = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{1}{6}(3 + 5.2 + 7 + 9 + 10.5 + 11.2) = 7.65$$

Measuring the "Spread" of the Data

The "distance" of each value from the mean can be used to measure the spread.



Measuring the "Spread" of the Data

Value	$x - \bar{x}$
3	-4.65
5.2	-2.45
7	-0.65
9	1.35
10.5	2.85
11.2	3.55

The values of $x - \bar{x}$ are called the "deviations"

Perhaps the **mean** of the deviations can be used as a measure for **spread**?

It turns out that the sum of all deviations is always 0.
That's because the **mean** of the data is the **center**. There is just as much negative deviation as positive deviation.

The negative signs can be dealt with by squaring the deviation!

Measuring the "Spread" of the Data

Value	$x - \bar{x}$	$(x - \bar{x})^2$
3	-4.65	21.6625
5.2	-2.45	6.0025
7	-0.65	0.4225
9	1.35	1.8225
10.5	2.85	8.1225
11.2	3.55	12.6025
Total =	0	50.595

Mean of the SQUARED deviations

$$\frac{50.595}{6} = 8.4325$$

The mean of the squared deviations is called the **variance**.

A Note About the Variance

Value	$x - \bar{x}$	$(x - \bar{x})^2$
3 ft	-4.65 ft	21.6625 ft^2
5.2 ft	-2.45 ft	6.0025 ft^2
7 ft	-0.65 ft	0.4225 ft^2
9 ft	1.35 ft	1.8225 ft^2
10.5 ft	2.85 ft	8.1225 ft^2
11.2 ft	3.55 ft	12.6025 ft^2
Total =	0 ft	50.595 ft^2

Suppose the previous data were measurements of distance.

The units of the variance is squared

$$\frac{50.595\ \text{ft}^2}{6} = 8.4325\ \text{ft}^2$$

Taking the square root of the variance provides the **standard deviation**.

The units of the **standard deviation** are not squared.

$$\sqrt{(\text{Variance})} = \sqrt{8.4325\ \text{ft}^2} = 2.9039\ \text{ft}$$

The Variance and Standard Deviation

Formally, the **variance** and **standard deviation** of a population is given by:

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} \text{ and } \sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

$\sigma^2 = \text{population variance}$
$\sigma = \text{population standard deviation}$
$N = \text{population size}$

Formally, the **variance** and **standard deviation** of a sample is given by:

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} \text{ and } s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

$s^2 = \text{sample variance}$
$s = \text{sample standard deviation}$
$n = \text{sample size}$

The calculations for the population requires a division by N.

The calculations for the sample requires a division by (n-1).

This is known as Bessel's Correction

Example

Consider the following sample data:

3, 0, 8, 3, 5, 3, 8, 9, 7, 6

What is the **variance**?

What is the **standard deviation**?

First, determine if the data represents a **sample** or a **population**.

Sample

Next, the **center** must be found.

$\bar{x} = 5.2$

Example (Cont.)

Value	$x-\bar{x}$ $(x-5.2)$	$(x-\bar{x})^2$ $(x-5.2)^2$
3	-2.2	4.84
0	-5.2	27.04
8	2.8	7.84
3	-2.2	4.84
5	-0.2	0.04
3	-2.2	4.84
8	2.8	7.84
9	3.8	14.44
7	1.8	3.24
6	0.8	0.64
Total =	0	75.6

First, calculate the **deviation** of each value. This requires the **mean** of the data.

Next, square the deviations.

Since this is a SAMPLE, $n-1$ must be used in the denominator.

$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{\Sigma(x-5.2)^2}{10-1} = \frac{75.6}{9} = 8.4$ (Variance)

$s = \sqrt{s^2} = \sqrt{8.4} \approx 2.898$ (Standard Deviation)

Using the Standard Deviation

The standard deviation is useful when comparing values from different datasets.

A **z-score** can be calculated to determine how far above or below the mean a value is.

Consider the data from the previous example:

3, 0, 8, 3, 5, 3, 8, 9, 7, 6

$\bar{x} = 5.2$
$s = 2.898$

Each value is +/- a number of standard deviations from the mean.

Using the Standard Deviation

Each value can be defined as the **mean** plus an amount of the standard deviation.

$$x = \bar{x} + z \cdot s$$

Consider the value 9 from the data below:

3, 0, 8, 3, 5, 3, 8, 9, 7, 6

$$9 = 5.2 + z \cdot 2.898$$

What value of z will make this equation true? $\mathrm{z} \approx 1.311.$

The value of 9 is roughly 1.3 standard deviations above the mean.

The z-score for a value in a <u>sample</u> is calculated by: $z = \frac{x-\bar{x}}{s}$

The z-score for a value in a <u>population</u> is calculated by: $z = \frac{x-\mu}{\sigma}$

Using the Standard Deviation

Suppose two students, Alice and Bob, are applying for a scholarship. Their applications are nearly identical so their test scores will be used as a tiebreaker. Alice took the SAT and Brian took the ACT. These two tests are scored on different scales so a method of standardization must be used.

The **z-scores** of Alice and Brian's test scores can be calculated for comparison.

Alice's z-score: $z_A = \frac{x-\mu}{\sigma} = \frac{1345-1081}{176} = 1.5$

Brian's z-score: $z_B = \frac{x-\mu}{\sigma} = \frac{24-20.8}{5.3} \approx 0.603$

Student	Score	Mean Score	Standard Deviation
Alice	1345	1081	176
Brian	24	20.8	5.3

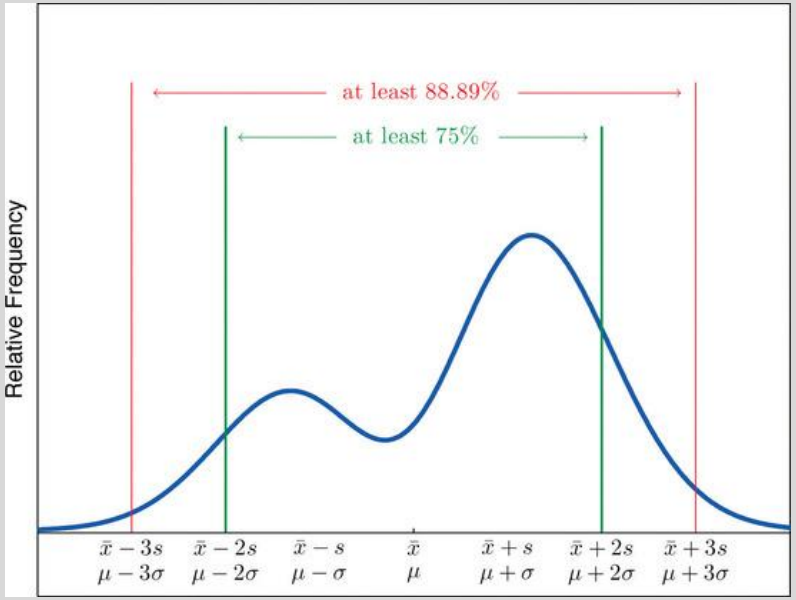
$Z_A > Z_B$

Since the z-score of Alice's test score is higher than Brian's, Alice scored better between the two.

Chebyshev's Rule

For any dataset:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.



A Quick Review

- $\text{Range} = \max - \min$
- Formally, the **variance** and **standard deviation** of a population is given by:

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} \text{ and } \sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

$\sigma^2 = \text{population variance}$

$\sigma = \text{population standard deviation}$

$N = \text{population size}$

- Formally, the **variance** and **standard deviation** of a sample is given by:

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} \text{ and } s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

$s^2 = \text{sample variance}$

$s = \text{sample standard deviation}$

$n = \text{sample size}$

- The z-score for a value in a population is calculated by: $z = \frac{x-\mu}{\sigma}$
- The z-score for a value in a sample is calculated by: $z = \frac{x-\bar{x}}{s}$