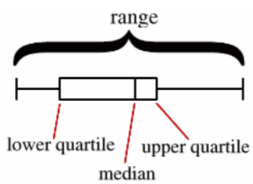
MDM4U: Mathematics of Data Management

Measures of Spread (Part 2)

Standard Deviation and z-Scores

MDM4U: Data Management



Standard Deviation

A *deviation* is the difference between any value in a data set and the mean.

For a population, a deviation is $x - \mu$, while for a sample, it is $x - \bar{x}$.

A data set with larger deviations has a greater spread.

Values less than the mean have negative deviations, while those above the mean have positive deviations.

The most common measure of deviation within a data set is the *standard deviation*, which measures the average distance of a datum from the mean of the data set.

Standard Deviation

Standard Deviation of a Population

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

Since a sample tends to underestimate the deviations in a population, the formula is slightly different for samples.

Standard Deviation of a Sample

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

Standard Deviation

Example

Calculate the standard deviation for the following data.

5	7	7	8	10	14	19

Solution: Calculate the mean of the data.

$$\bar{x} = \frac{5+7+7+8+10+14+19}{7} = 10.$$

Make a table, with columns for x, $x - \bar{x}$, and $(x - \bar{x})^2$.

Standard Deviation

Datum	x	$x - \bar{x}$	$(x - \bar{x})^2$
x_1	5	-5	25
x_2	7	-3	9
x_3	7	-3	9
x_4	8	-2	4
x_5	10	0	0
x_6	14	4	16
x_7	19	9	81
			$\sum(x-\bar{x})^2 = 144$

Therefore, $s = \sqrt{\frac{144}{7-1}} = \sqrt{24} \approx 4.899$.

Standard Deviation

There is a faster method of computing the standard deviation, developed prior to the emergence of statistical software.

This computational formula deals with the squares of each datum, rather than any differences from the mean.

Computational Formula for Standard Deviation (Sample)

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

Standard Deviation

Example

Verify the computational formula using the earlier data.

Datum	x	x^2
x_1	5	25
x_2	7	49
x_3	7	49
x_4	8	64
x_5	10	100
x_6	14	196
x_7	19	361
		$\sum x^2 = 844$

Therefore, $s = \sqrt{\frac{844-7(10)^2}{7-1}} \approx 4.899$.

Variance

Variance in a data set is a measure of dispersion of the data.

Mathematically, variance is the square of the standard deviation.

Variance of a Population

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

Variance of a Sample

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

Variance

Example

Calculate the variance of the earlier data.

Since the standard deviation is $s = \sqrt{24}$, the variance is $s^2 = 24$.

Note that the variance is *always* calculated as part of the process of calculating the standard deviation.

Also note that for both the standard deviation and the variance, we will *almost always* be using the formula for a sample, since we do not often have data for the entire population.

Variance

Your Turn

Calculate the variance and standard deviation of the following data.

3	7	9	10	13	24

Solution: The mean is $\overline{x} = \frac{3+7+9+10+13+24}{6} = 11$.

Use the computational formula to calculate the variance and standard deviation.

Variance

Datum	x	x^2
x_1	3	9
x_2	7	49
x_3	9	81
x_4	10	100
x_5	13	169
x_6	24	576
		$\sum x^2 = 984$

Therefore, the variance is $s^2 = \frac{984-6(11)^2}{6-1} = \frac{258}{5} = 51.6$.

The standard devtaion is $s = \sqrt{\frac{258}{5}} \approx 7.183$.

z-Scores

A *z-score* measures the number of standard deviations a datum is from the mean.

z-Score for a Population

$$z = \frac{x-\mu}{\sigma}$$

z-Score for a Sample

$$z = \frac{x-\overline{x}}{s}$$

A negative z-score indicates a datum is below the mean, while a positive z-scores indicates it is above.

z-Scores

Example

A data set has a mean of 5 and a standard deviation of 1.2. Determine the z-scores for data with values of 6.2 and 3.

Solution: The first datum is above the mean, so its z-score will be positive.

The datum is $z = \frac{6.2 - 5}{1.2} = \frac{1.2}{1.2} = 1$ standard deviation above the mean.

z-Scores

The second datum is below the mean, so its z-score will be negative.

$z = \frac{3 - 5}{1.2} = -\frac{2}{1.2} = -\frac{5}{3}$. The datum is one-and-two-thirds standard deviations below the mean.

z-scores will play a *very* important role in the last unit of this course when we deal with continuous probability distributions.

Questions?