Name: ________________________________ Block: _______ Date: ____________

Statistical Variation

Notation/Term	*Description*	*Example/Notes*
population	An entire set of data about which we wish to gain information.	The height of every middle school student in the U.S.
sample	A subset of data from a population. We use samples to make inferences about populations.	The heights of students in our class.
x_i	The i^{th} element of data in a sample or population.	In data set {2, 4, 5, 10}, $x_2 = 4$
mean of a population	Represented by symbol μ (Greek letter "mu").	For our purposes, we will use $\bar{x}$ in place of μ in formulas.
mean of a sample	Represented by $\bar{x}$ (x bar)	Symbol our calculator uses.
deviation	The distance a data point is from the mean of the data. Can be positive, negative, or zero.	Find by calculating $x - \mu$ or $x - \bar{x}$
Σ	The Greek uppercase symbol Σ (sigma) is used to indicate summation. The terms inside the sigma are added together based on an indexing scheme surrounding the Σ	$\sum_{i=1}^{5} i = 1 + 2 + 3 + 4 + 5$; $\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4$
mean absolute deviation	Mean of absolute values of the deviations of elements from the mean of a data set.	mean absolute deviation = MAD = $\frac{\sum_{i=1}^{n} \lvert x_i - \mu \rvert}{n}$
variance	The average of the squared deviations from the mean. Variance is represented by the lower case symbol sigma squared (σ^2)	variance = $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$
standard deviation	The square root of the variance, denoted by σ.	standard deviation = $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$
z-score	Also called standard score; measure of position that determines the number of standard deviations an element is above or below the mean of a data set.	z-score = $z = \frac{x - \mu}{\sigma}$

Summation Notation

The Greek uppercase letter Σ (sigma) is used to denote summation.

Examples: $\sum_{i=1}^{5} i = 1+2+3+4+5 = 15$ $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + ... + x_n$

You try: Evaluate the summations…

a) $\sum_{i=1}^{4} 2i$ b) $\sum_{i=1}^{3} i^2$ c) $\sum_{i=1}^{5} (i+1)$ d) Rewrite the formula for mean using Σ notation.

Population vs. Sample

In statistics, a population is a well-defined group of objects about which we wish to gain information.

Examples of populations: the heights of every U.S. President, the ages of every math teacher in Virginia.

It is often very difficult to obtain data for every member of a population. Instead, we take a sample of the data, or a subset of the population. There are various ways of choosing sample data, but statisticians do their best to get the best representative sample they can.

We refer to a specific element, the i^{th} element, of a sample or population as x_i.

Measures of Spread or Variability

We have already studied the measures of central tendency: mean, median, and mode. We use these measurements to determine a number that best describes a typical data point. Measures of variability attempt to describe the spread of a set of data.

We have already looked at some measures of variability: range and interquartile range.

Deviation is the distance a data point is from the mean of the data. It can be positive, negative, or zero, and is found by subtracting the mean (μ) from the data point value (x).

Example: S = {1, 3, 4, 6, 7, 9} μ = 5 (add the numbers and divide by 6)

Deviation = x - μ for each x

x	1	3	4	6	7	9
x - μ	1 – 5 = -4	3 – 5 = -2	4 – 5 = -1	6 – 5 = 1	7 – 5 = 2	9 – 5 = 4

Sum all the deviations – what do you get? _________

Will this always happen? Why? ____

We cannot simply use the deviations to find a mean deviation. Why not? __________________

There are two ways to avoid the problem of having the mean deviation always come out to 0.

Mean Absolute Deviation (MAD)

We have seen that we do not want negative deviations, since then the sum of deviations becomes 0. We can use absolute value to ensure we always have a positive deviation.

We find the absolute deviation of each data point in a sample or population, and then divide by the number of data points to find the mean absolute deviation.

Absolute deviation is defined by $|x - \mu|$, where x is a data point and μ is the mean.

Using our previous example of data: S = {1, 3, 4, 6, 7, 9} $\mu = 5$

x	1	3	4	6	7	9
$\lvert x - \mu\rvert$	$\lvert 1-5\rvert = 4$	$\lvert 3-5\rvert = 2$	$\lvert 4-5\rvert = 1$	$\lvert 6-5\rvert = 1$	$\lvert 7-5\rvert = 2$	$\lvert 9-5\rvert = 4$

To find the mean absolute deviation, we would sum these deviations and then divide by the number of deviations (6). The generic formula for mean absolute deviation is:

$$\text{Mean Absolute Deviation (MAD)} = \frac{\sum_{i=1}^{n} |x_i - \mu|}{n}$$

In the above example, the mean absolute deviation = $\frac{4+2+1+1+2+4}{6} = \frac{14}{6} = \frac{7}{3} \approx 2.3$

Standard Deviation

Another way to ensure that deviations are always positive is to square them. We can sum the squares of the deviations and divide by the number of deviations to get variance, denoted by σ^2:

$$\text{Variance} = \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

We take the square root of variance to get a closer approximation of the deviation of points from the center. We call the square root of variance standard deviation.

$$\text{Standard Deviation} = \sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

In our above example, S = {1, 3, 4, 6, 7, 9}, the variance and standard deviation are as follows:

$$\sigma^2 = \frac{(1-5)^2 + (3-5)^2 + (4-5)^2 + (6-5)^2 + (7-5)^2 + (9-5)^2}{6} = \frac{42}{6} = 7; \qquad \sigma = \sqrt{7} \approx 2.65$$

You try: Find the MAD, variance, and standard deviation of S = {1, 1, 3, 4, 5}

It's a lot of work to find these descriptive statistics by hand! Fortunately, the calculator comes to our rescue. Use the calculator guide to find the mean absolute deviation (MAD), variance, and standard deviation of the following data sets:

a) {10, 20, 30, 40, 50, 60}

b) {10, 12, 7, 11, 20, 7, 6, 8, 9}

c) {1202, 1229, 1012, 1014, 1120, 1429}

Z-Scores (Standard Scores)

We use standard deviations to assess how typical a data point is. How far away is a data value from the mean?

A z-score tells us how many standard deviations an element is above or below the mean of the data set. A z-score is calculated by subtracting a data set's mean from the element of interest, and dividing that difference by the standard deviation of a data set.

$$\text{Standard Score} = \text{z-Score} = z = \frac{x-\mu}{\sigma}$$

A positive z-score means that the point is that many standard deviations above the mean, and a negative z-score means it is that many standard deviations below the mean.

Example: A set of values has a mean of 85 and a standard deviation of 6. Find the z-score of the value 76.

$$\text{z-score} = \frac{x-\mu}{\sigma} = \frac{76-85}{6} = \frac{-9}{6} = -1.5$$

This tells us that 76 is 1.5 standard deviations below the mean.

Example: John weighs 220 lbs; his dog Fido weighs 90 lbs. If human males weigh an average of 160 lbs with a standard deviation of 20 lbs, and all dogs of Fido's breed have an average weight of 80 lbs with a standard deviation of 5 lbs, how do John and Fido compare, relative to their populations, with respect to weight?

John's z-score =

Fido's z-score =

Conclusion?

You try:

a) Find the z-score for a data value of 15 if a set of data has a mean of 75, and a standard deviation of 5.

b) Find the value of an element in a dataset with a mean of 8, a standard deviation of 2, and a z-score of 1.5.

c) Test A has a mean of 50 and $\sigma = 10$. Test B has a mean of 20 and $\sigma = 5$. Which is better: a score of 65 on test A or a score of 29 on test B?