Basic Statistics Terms and Calculations

Statistics - The science of collecting, organizing, describing, and interpreting data or information. In the study of statistics, it is important to be familiar with a variety of terms.

Data Set - A collection of information.

Variable - The characteristics about which information can be collected.

Distribution - The way a variable's values are spread over the possible values. The distribution can be displayed in a table or a graph.

Qualitative Variable - Variables that are classified into categories (i.e. colors, sports, makes of cars, etc.)

Quantitative Variable - Variables that are numerical and describe how much or how many of something there is (i.e. peoples' ages, heights, salaries, test scores, etc.).

Mode -The most common value that shows up in a data set. There *can be* more than one mode. If all of the data values appear only once, then there is *no mode*. If the data set is bar graphed, the mode(s) will show up as a high point/peak. Both Qualitative and Quantitative variables can have a mode.

Note: *There are certain terms used in statistical analysis that **only** apply to Quantitative Variables. Often, these terms involve a mathematical computation as part of their definition. For example…*

Mean – The average value of a data set. This can be calculated by summing up all of the individual values in the data set and dividing the total by the number of data values (n) in the set.

Median - The middle value in a sorted (i.e. low to high) data set. If there is an even number of values, then it is the average of the two middle values.

Range – The difference between the highest and lowest values of a data set.

Variation - A measure of how widely data values are spread out from the center of a data set.

Variance - A measure of how far the values in a data set are from the mean, on the average. *To complete the calculation, it is necessary to know whether the data set is from a population or a sample.*

Standard Deviation – A measure of how far data values are spread around the mean of a data set. It is computed as the square root of the variance. *Therefore, to complete the calculation, it is necessary to know whether the data set is from a population or a sample.*

z-score - A measure of how many standard deviations a specific value (x) in the data set is from the mean of the data set. The z-score is positive if the data value is greater than the mean and negative if it is less than the mean. This term can also be referred to as "the standard score."

Note: *When analyzing a data set, it is necessary to identify if the data is from a population or from a sample. Different symbols and formulas are used to represent certain statistical terms depending on whether the data is from an entire population or a sampling of the population.*

Population – The complete set of people or things being studied.

Sample – A subset of the population from which the raw data are actually obtained. Sampling techniques are often utilized if it is not feasible to gather the entire population of data.

Term	***Symbol(s)***	***Formula(s)***
Sigma	Σ	*This symbol is used to represent the sum of a specific set of values.*
Mean	μ *(population)* $\overline{x}$ *(sample)*	*Mean* = $\frac{sum\ of\ all\ values\ in\ the\ data\ set}{total\ number\ of\ values\ in\ the\ data\ set}$ $\mu = \overline{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + \cdots}{n}$, where x_i = individual data points
Variance	σ^2 *(population)* s^2 *(sample)*	*Population Variance* = $\frac{\sum (x_i - \mu)^2}{n}$ *Sample Variance* = $\frac{\sum (x_i - \overline{x})^2}{n-1}$
Standard Deviation	σ *(population)* s *(sample)*	*Population Standard Deviation* (σ) $\sigma = \sqrt{population\ variance} = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$ *Sample Standard Deviation* (*s*) $s = \sqrt{sample\ variance} = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$
Z-Score	z	$z = \frac{x - \mu}{\sigma} = \frac{x - \overline{x}}{s}$ *x = specific value from the data set.*

Example - The population of class grades on a recent exam is displayed in the table. Using the following normally distributed data set, identify or calculate the following -

a. Mode
b. Mean
c. Median
d. Range
e. Variance
f. Standard Deviation
g. z-score for the student who received an 88 on the exam.

Students Grades on the most recent Statistics Exam

96	74	92	75	64
80	70	54	81	80
72	88	61	76	92

Note: *Before* analyzing the data, it can be helpful to *re-order* it from smallest to biggest! This visually allows the identification of some important information.

Re-ordering the data from low to high makes it easier to identify the High, Low, any repeated numbers, and the "middle" of the data set.

Students Grades on the most recent Statistics Exam
(Data Re-ordered from low to high)

54	61	64	70	72
74	75	76	80	80
81	88	92	92	96

Lowest data point

The Median – since there is an odd number of data points in the set, it is the point in the middle.

There are **2 modes** (80 and 92) because they both show up twice.

Highest data point

Answers –

a. **Mode(s) →** 80 and 92 *(The data values that show up the most often)*

b. **Mean →** $\mu = \frac{\sum x_i}{n} = \frac{1155}{15} = 77$

Sum all of the data points together

Divide by the total number of data points

c. **Median →** 76 *(Center of data set)*

d. **Range →** $Range = High - Low = 96 - 54 = 42$

e. **Variance** → If calculating by hand, it can be helpful to create a table.

	Table for Calculating Variance			
	Data Point *(x)*	**Mean (μ)**	$(x - \mu)$	$(x - \mu)^2$
1	54	77	-23	529
2	61	77	-16	256
3	64	77	-13	169
4	70	77	-7	49
5	72	77	-5	25
6	74	77	-3	9
7	75	77	-2	4
8	76	77	-1	1
9	80	77	3	9
10	80	77	3	9
11	81	77	4	16
12	88	77	11	121
13	92	77	15	225
14	92	77	15	225
15	96	77	19	361

n

2008

Sum of this column

Sum

Variance →

$$Variance = \frac{\sum (x-\mu)^2}{n} = \frac{2008}{15} = 133.87$$

f. Standard Deviation → $\sigma = \sqrt{population\ variance} = \sqrt{133.87} = 11.57$

g. Z-score for student who received an 88 on the exam → $z = \frac{x-\mu}{\sigma} = \frac{88-77}{11.57} = 0.95$

Try this problem on your own!

Find the following for the given data set

a. Mode *(Answ: No Mode)*
b. Mean *(Answ: 10.08)*
c. Median *(Answ: 9.2)*
d. Range *(Answ: 58)*
e. Variance *(Answ: 163.49)*
f. Standard Deviation *(Answ: 12.79)*
g. Z-score for Walmart *(Answ: 0.44)*

Corporate Profit Levels in billions in 2012
(Source: CNN Money)

Exxon	41.1	Ford	20.2
Walmart	15.7	H-P	7.1
Chevron	26.9	AT&T	3.9
Conoco	12.4	Valero	2.1
GM	9.2	BoA	1.5
GE	14.2	McKesson	1.2
Berk.Hath.	10.2	Verizon	2.4
Fannie Mae	-16.9		