Variation and Statistics

Mean (Average) – A mean is the sum of a set of values divided by the number of values in the set. $\overline{x} = \frac{\Sigma x}{n}$

Mode – The mode of a set of values is the value *or* values that appear most often (highest frequency). A set of values can have more than one mode.

Median – The median of a set of values is the value that appears in the middle when the values are listed in numerical order. If there is an even number of values in the set, the median is the average of the two middle values.

Quartiles – The *first* quartile is the median (middle value) of the first half of a set of values. The *third* quartile is the median of the second half.

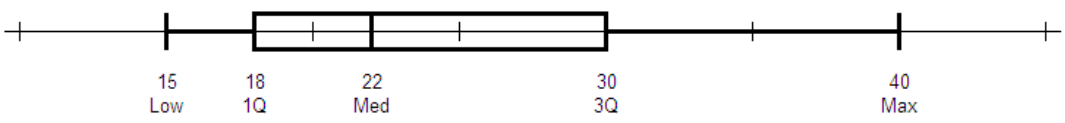
Range – The range of a set of values is the difference in the highest and lowest values in the set.

Five-Number Summary – The five-number summary consists of the following five numbers:

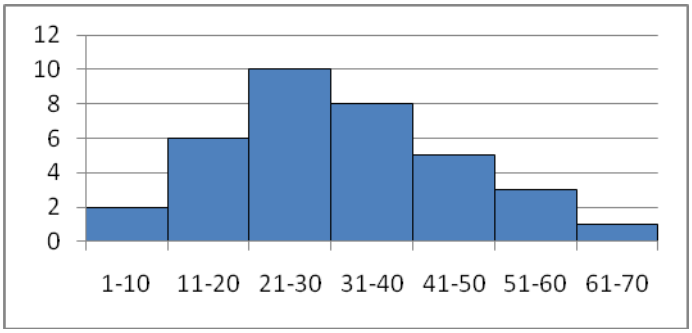
Lowest – First Quartile – Median – Third Quartile – Highest

Box-and-Whisker Plot (Boxplot) – A graphical representation of the five-number summary. A box around the 1st and 3rd quartiles, and whiskers out to the low and high numbers.

Lowest - 15
1st Quartile - 18
Median - 22
3rd Quartile - 30
Highest - 40



Histograms – A graphical representation of a distribution of numbers. In order to create a histogram, first find the range of the numbers (high – low). Then decide how many rectangles you want in your histogram and divide the range by that number (rounding *up* to a multiple of 5 is often convenient). Start with the lowest number and add the previously calculated number until you reach the highest number. These are the separating numbers for your rectangles.



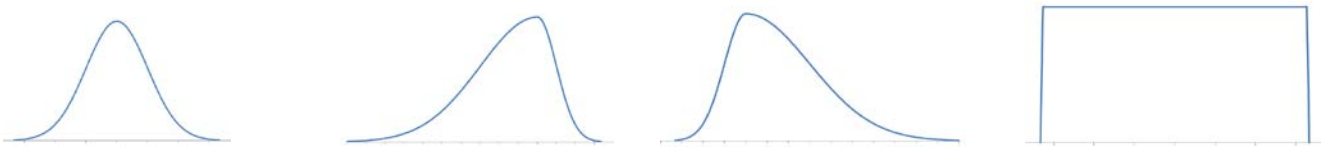
Standard Deviation (Average Error)

The standard deviation of a set of values is the average error of the set of values or the total amount of error distributed evenly to each value. To calculate it, we find the mean: μ, we find each value's error: (x- μ), we square those values in order to make them all positive: $(x- \mu)^2$, we add them up: $\Sigma(x- \mu)^2$, we average them out by dividing by n and then square root that number.

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

Range Rule of Thumb for Standard Deviation – The standard deviation can be estimated by dividing the *range* by 4.

Types of Distribution



Normal (Symmetrical/Bell Shaped) | Negatively (Left) Skewed | Positively (Right) Skewed | Uniform

The Normal Curve

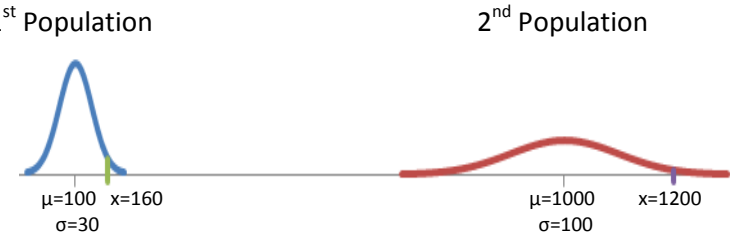
The normal curve or bell-curve describes the probability distribution for a continuous variable. There are more values near the mean or middle, and less and less as the values get higher or lower than the mean. Because it illustrates probability, the area under the ***any distribution*** curve adds up to 1.

Z-scores (Standardized Scores)

A z-score is the number of standard deviations that a certain value is away from the mean (average). "Standardized scores" are used to have a standard way of comparing seemingly incomparable values. To find a z-score, we calculate the value's distance from the mean: x- μ, then we divide that by the standard deviation: σ.

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

For example: Say we have two populations. The first has a mean of 100 (μ=100) and a standard deviation of 30 (σ=30). The second has a mean of 1000 (μ=1000) and a standard deviation of 100 (σ=100). How does a value of 160 (x=160) from the first population, compare to a value of 1200 (x=1200) from the second population?



Let's calculate the two values' z-scores.

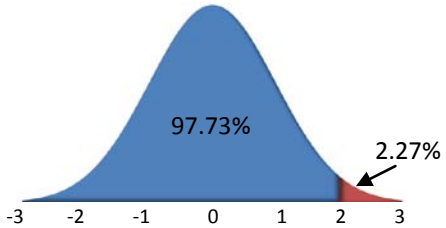
For the first: $z = \frac{160-100}{30} = \frac{60}{30} = 2.00$

For the second: $z = \frac{1200-1000}{100} = \frac{200}{100} = 2.00$

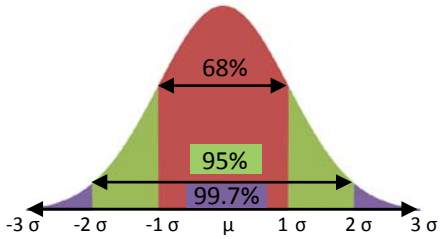
From these calculations, we can see that both values are exactly 2 standard deviations away from their means, so they are equivalent in a statistical interpretation.

Probabilities and z-scores

Because z-scores are "standard", we are able to calculate the area or probability that another value falls above, below or between other values. Using our previous example, we can calculate the probability that a value falls above 160 in the first population. Using a table of z-scores, we find that the area below a z-score of 2 is .9773. So, we have a 97.73% chance of getting a value below 160 in our first population. Because all area under the normal curve adds up to 1, we have a 2.27% chance of getting a value above 160. (1-.9773 = .0227)



Empirical Rule (68-95-99% Rule)
68% of the data falls within 1 standard deviation of the mean, 95% falls within 2 standard deviations of the mean, and 99.7% falls within 3 standard deviations of the mean.



Types of Data – *Nominal* (data are named/labeled but do not have numerical value); *Ordinal* (data is ranked, but does not have true numerical value); *Interval* (numerical data on a defined scale; ex: temperature); *Ratio* (numerical data similar to Interval data but the scale has a true zero; ex: weight)

Population – All items or individuals that you care to study or describe. Values describing a population are called *parameters*.

Samples – A selected number of items from the population that are used to make inferences or predictions about the entire population. Values describing a sample are called *statistics*.

Types of Samples – *Convenience* (selected based on accessibility and ease of selection); *Stratified* (selected from unique subgroups of the population); *Cluster* (selected from subgroups similar to the population); *Systematic* (selecting items in a population that are equidistant, ex: every 10^{th} term); *Random* (selected from the population so that every item has an equal chance of being selected; unbiased)

Sampling Distribution – The distribution of a characteristic from all possible samples of size *n* within a population. Given large enough samples, the sampling distribution should be normal, resembling the populations distribution and having the same mean.

Standard Error – The standard deviation of the sampling distribution. It is calculated by dividing the standard deviation by the square root of the sample size. This means that the larger the sample size, the smaller the standard error.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Confidence Intervals – An interval that estimates a population parameter within a given confidence level. To calculate a confidence interval, you must find the standard error, multiply that by the standardized value (z-value) associated with the confidence level, and then add and subtract that from the "point estimate" for the population parameter (the sample statistic).

$$\mu = \bar{X} \pm z\frac{\sigma}{\sqrt{n}}$$

T-values – These are less accurate z-values that are used with samples, size *n*, whose population standard deviation, σ, is unknown. The larger the sample size, the more the t-distribution resembles the z-distribution.

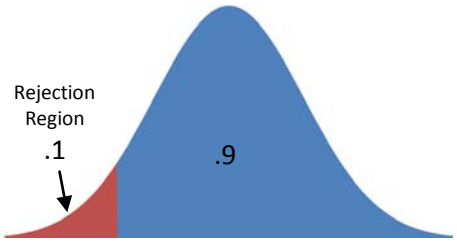
Hypothesis Tests – Used when you need to test a claim about population means or proportions, by using a sample.

Null hypothesis (H_0) – The statement about the population that will be tested. It *must* include equality.
Alternative hypothesis (H_A) – The statement that includes all values of the population that are not in the null.
Rejection Region – The tail or tails that correspond to the alternative hypothesis. A value that falls in this region will cause the null hypothesis to be rejected.
Significance Level (α) – The area associated with the rejection region, and the allowable Type I error.
Critical Value – The value of the statistic (z, t, etc) corresponding to the significance level.

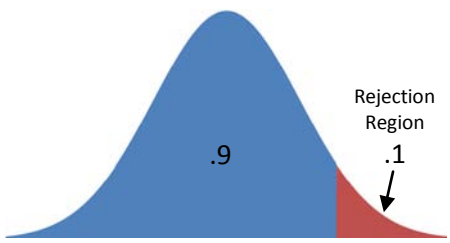
1-Tailed versus 2-Tailed Tests

One-Tailed Tests

$H_0: \mu \geq \#$
$H_A: \mu < \#$
$\alpha = .1$

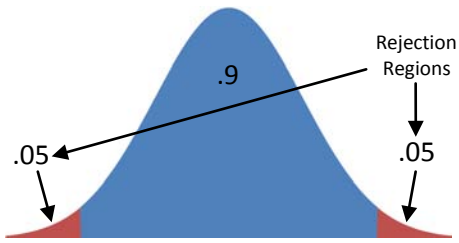


$H_0: \mu \leq \#$
$H_A: \mu > \#$
$\alpha = .1$



Two-Tailed Test

$H_0: \mu = \#$
$H_A: \mu \neq \#$
$\alpha = .1$



If the calculated statistic lies in the appropriate rejection region or if the p-value > α, then H_0 should be rejected.

Types of Error – As an example, let's say that a company has hired you to test their water filtration system. There are four (4) possible outcomes from your tests.

		Actual State of Water	
		Water is ok	Water is not ok
Test Results	Water is ok (Do not reject Ho)	Correct Conclusion	Type II Error
	Water is not ok (Reject Ho)	Type I Error	Correct Conclusion

Type I error is committed when your tests conclude that you should reject the null hypothesis, but it was unnecessary. This error would only cost the company money because, based on your suggestion, they will recalibrate their machines, though they did not have to.

Type II error is committed when your tests conclude that you should not reject the null hypothesis, but it should have been rejected. This error would potentially cost the lives of consumers because the company will not recalibrate their machines, though the water was actually unsafe.