

## Basic Statistics Vocabulary

Population: The set of all measurements of interest to the sample collector.

Parameter: A characteristic of the population.

Statistic: A characteristic of a sample drawn from a population.

Accuracy: Closeness of a measured or computed value to its true value. How close to the actual population the sample estimates are. This is affected by nonsampling errors (errors from method of measurement, improperly designed or executed sampling plan).

Precision: Closeness of repeated measurements of the same quantity to each other. Consistency of measurements, or reliability. Restricted to errors of sampling, malfunctioning equipment, or observer bias.

- A biased but sensitive scale may give inaccurate but precise weight readings.
- An insensitive scale may give an accurate reading, by chance, but it would be an imprecise reading since a number of readings would likely all yield different weights.
- The 'bullseye' of a dart board is accurate. However, a person who throws consistently off to the left will produce a group of darts clustered together (precise throws, but inaccurate).
- **Unless there is bias in a measuring instrument, precision will lead to accuracy.**

Measures of Central Tendency: There are three commonly used measures: Mean, Median, and Mode. The **Mean** is the average value of a set of measurements (the sum of all the measurements divided by the total number of measurements). The **Median** of a set of measurements is the middle value when the measurements are arranged from lowest to highest. The **Mode** is the measurement that occurs most often.

Measures of spread (Measures of variability):

Range: The difference between the largest and smallest values. A large range indicates a large amount of variability.

The most common measure is the sample standard deviation (s). This is a measure of how far a typical observation deviates from the sample average. Since this is a measure of spread about the *mean*, it should be used only when the mean is used as the measure of center. The standard deviation (s) is the square root of the variance  $s^2$ . The variance will be large if the observations are widely spread about their mean, and small if the observations are all close to the mean. Because the variance involves squaring the deviations it doesn't have the same unit of measurement as the original observations. Lengths measured in centimeters have

a variance measured in squared centimeters. Taking the square root remedies this, so the  $s$  measures dispersion about the mean in the original scale. The standard deviation = 0 only when there is no spread (all observations have the same value). Otherwise,  $s > 0$ .

Sampling error: A standard deviation among estimates rather than among individual observations. Regardless of the sample size (generally, the larger the sample the more likely it is that the measured use will represent the whole population), there is always some difference between the sample statistics and the population parameters being estimated. *The most appropriate sample size depends on how much sampling error you accept.*

Standard error: A mathematical way of expressing sampling error. If it's small, the measure of reliability is good. Standard error is most commonly computed for the mean. It indicates the amount of variation among means taken from many samples. How far do they deviate from the population mean? The greater the variation in values from sample to sample, the larger the standard error, and the less reliable the statistic for estimating parameters. Standard error is effected by sample standard deviation ( $s$ ) and by sample size. If  $s$  is constant but the sample size increases, the standard error decreases.

---

Confidence interval: The range + or – from the sample mean

Confidence level: Usually 95%. The probability that the interval contains the true mean is 95%. A confidence level of 67% may be sufficient for broad policy decisions, while site specific or allocation decisions may require a 95% confidence level.

- To increase the confidence limit (99%), the confidence interval must become wider. However, as confidence increases, the statement becomes vaguer.
- Use is estimated to be 123 + or – 10% at the 67% confidence level. Use is between 111 and 135 67% of the time.

Correlation: The manner in which two or more variables fluctuate with reference to each other. Do they vary together? Also tells us the strength of the association (measures intensity of association observed between two variables, and whether the correlation could be expected to chance alone). One variable may be the cause of the other, but we neither know nor assume this. Both variables must be random.

- Positive correlation – as education increases, income increases
- Negative correlation – as education increases, number of children decreases
- Correlation coefficient - +1 (perfect positive association) to –1 (perfect negative correlation). There is no dispersion about the line of relationship.