Key Terms used in Statistics

Stats is like a language. Remember that the statistics definition of a word may be vastly different from the dictionary definition. Here are some key vocabulary used in statistics. (The terms are listed in alphabetical order.)

Assumptions – Will the data collected represent the population well? Assumptions will help us decide. Each test has different assumptions, but most assumptions include that the data is random and have some way of checking if the data set is large enough to measure the population. If sample data meets assumptions, it may give us good information about the population and allow us to make a confidence interval or test a claim with at hypothesis test. If data does not meet most of the assumptions, it is often garbage and does not represent the population. So we are wasting our time analyzing it in the first place.

Biased Sample – A sample is biased if it does not represent the population. An incorrectly collected data set will often only represents certain groups of the population and leave other out.

Boot-Strapping – Sampling distributions are key to understanding sampling variability, calculating standard error and creating confidence intervals. How can a sampling distribution be created when we do not know the population? Most of the time, we only have sample data. A technique was developed to create a sampling distribution by taking random values from the sample with replacement. This is called "Boot-Strapping". It is a little controversial, but boot-strapping does give pretty accurate estimates of standard error and pretty accurate confidence intervals.

Categorical Data (also called Qualitative Data) - Information that puts elements or people in categories of groups. For example the city a person lives in. In categorical data we are interested in exploring percentages in the various groups.

Census – Attempting to get data from everyone in the population. This tends to be the most accurate data you can get, (even better than a random sample.)

Center – The average of a data set. There are different types of center and different types of averages. We want the best average to represent the middle of the data. In Ch. 3, we see that when a data set is bell shaped, we like to use the mean as our center or average, but if a data set is skewed or irregular shaped, we like to use the median as our center or average.

Claim – What we want to find out about a population or what someone guesses a population value is.

Conclusion – In general conclusions should explain the meaning of something in a non-technical way. Someone who does not know statistics should be able to understand the conclusion. As statisticians, we need to analyze complicated ideas like variability and patterns and p-value and explain the meaning of the statistics to people so that they can understand and make sense of the world around them.

Confidence intervals – Two numbers that we think the population might be in between. It is often calculated by taking the sample statistic and then adding and subtracting the margin of error from it. The tricky part is that our interval may or may not contain the population value.

Confident (95% confidence interval) – 95% of confidence intervals created contain the population value and 5% of confidence intervals created do not contain the population value. When we create an interval, we are never sure if it does or does not contain the population value. Confidence in Statistics is not about being sure you got it right, instead it admits that we could have gotten it wrong.
Data - Information in many forms

Hypothesis Test – We look at sample data and compare it to the population value in the claim. The trick is to figure out if there is or is not a significant difference between the sample values and the population value guessed at in the claim.

Margin of Error – Think of this as the approximate difference between the sample statistic and the population parameter. We want to know how far off our point estimate might be.

Outlier – a data value that differs dramatically from the usual values in the data set. We often look at graphs to spot outliers. These are often mistakes in the data collection or just very interesting cases.
Parameter – a number describing a population, often theoretical or guessed at. We tend to use Greek letters for parameters in Statistics.

Point Estimate – When trying to approximate a population parameter, we often start with the sample statistic. If the data was collected randomly, then the statistic will hopefully be somewhat close to the actual population value. Sadly many newspapers and articles state a sample value and tell you it is a population value. This is called a point estimate. They are often very wrong.

Population - All elements or people to be studied. For example a population may be everyone in the U.S.A.

P-value – Generally one of the most difficult concepts for students to understand. P-value has far ranging implications and some deep meaning behind it. In a general sense, the P-value is a conditional probability. If the null hypothesis is true, the p-value is the probability of getting the sample statistic(s) (sample data) or more extreme. Or more extreme refers to getting the sample data or any data that would significantly disagree with the null hypothesis. P-value is best understood through simulation. Simulate the null hypothesis and then calculate the chances of getting the sample data or any other simulation more extreme than the sample data. (Traditional explanations of P-value involve calculating area under a curve corresponding to the test statistic. Simulation techniques give a much better visual for student understanding.) The key is that the lower the P-value gets, the more of a discrepancy there is between the sample value and the population value in the null hypothesis. So a low P-value (close to zero) means we can reject the null hypothesis.

Quantitative Data - Information that measures something, it usually involves units and the average makes sense in the context. For example the distance people can throw a baseball in feet. In quantitative data we are concerned with shape, center, spread, and outliers.
Random- A sampling method where everyone in the population has an equal chance of being in the sample. Very difficult and expensive. A mistake students tend to make is that they say things like "I bumped into this person at the mall. It was totally random." That is wrong. Not everyone in the population had a chance to bump into them at the mall. When estimating or doing a hypothesis test about a population, we require the sample data to be collected randomly. Data not collected randomly is very biased and does not represent the population.

Sample – a subset of the population. We often cannot get information from everyone in a population, but we can get information from some of the population. For example, if the population is everyone in the U.S.A., then we may get data from 2500 people in the U.S.A. The trick is that we want the sample to represent the population.

Sampling Distribution – Taking lots and lots of samples from a population and can graphing all the sample values from each sample. For example a sampling distribution of sample means takes lots and lots of random samples and then looks at a graph of all the sample means. We can do the same for sample percentages. Understanding sampling distributions is key to understanding sampling variability. It also helps in understanding of standard error, margin of error and confidence intervals.

Sampling Variability – The study of how well samples estimate a population. This is one of the most crucial topics for students to understand how samples work. For example many stat students think that random samples are a perfect approximation of the population value. They do not know how wrong they are. Random sample values can have a huge amount of variability from the population value. The study of this variability is a crucial part of inferential statistics.

Shape – Quantitative data has a general shape to it. Data describing one quantitative variable may look bell shaped (normal), skewed, or uniform. We learn to recognize shapes when we see them. Ordered pair data may have a linear or nonlinear shape or be scattered with no relationship. The key is that shape determines the most accurate statistics to use.

Spread – In a quantitative data set, think of spread as how far away numbers are from the center. In statistics, we are often looking for typical spread or how far typical numbers are from the center. In Ch. 3, we see that when a data set is bell shaped, we like to use the standard deviation as our best measure of spread, but if a data set is skewed or irregular shaped, we like to use the IQR as our measure of spread. In a general sense spread or variability is the most important concept in statistics. For example, what is the average yearly salary of people living in the U.S? Let's suppose we think that the average yearly salary in the U.S. is $42,000. Does everyone in the U.S. make $42,000? Of course not. There is variability. Individual salaries may vary dramatically from $42,000. We need to understand why and how data varies.

Standard Error – The standard deviation of a sampling distribution. For example, If we look at a sampling distribution of lots and lots of sample percentages from a population, the standard error is the standard deviation of all of those sample percentages.

Statistic – a number describing a sample.

Statistics - The study of information or data to learn about the world around us.

Test Statistics – Think of it like a ruler. It measures how far sample data is from the population value, usually in terms of the number of standard errors. The bigger the test statistic, the more discrepancy there is between the sample data and the population value (null hypothesis).

Unbiased Sample – A sample is unbiased if it represents the population well. For example large random samples tend to be unbiased estimators, as they represent the population rather well.