Terms Used in Statistics not Defined Elsewhere

Analytic Statistics – Drawing conclusions from data sets that will allow decisions under uncertainty. Some authors include both probability and inductive statistics in this category.
Census – A survey of an entire population, "A 100% sample"
Continuous data – Technically, data made up of variables characterized by the fact that there can be a value of the variable between any two values of the variable, so that the number of values is uncountably infinite. A continuous variable will have a range within the real numbers. As a practical matter a continuous variable that takes its value from a measuring process like height, volume or weight. Dollar amounts are usually considered continuous.
Confidence Level – In hypothesis testing the probability of a given statistical test not rejecting a null hypothesis when the null hypothesis is true. This is logically equivalent to its definition in creating confidence intervals where the confidence level is the probability that a given parameter falls within an interval that is supposed to estimate that parameter. The confidence level is indicated by $(1-\alpha)$, where α is the **significance level**.
Consistent – An estimator is consistent is consistent if the probability that it is within any arbitrarily small distance of the parameter it estimates approaches one as the sample size becomes infinite, for example if the variance of its sampling distribution approaches zero and it is unbiassed.
Cross Section Data – Data on some variable taken at the same point or period in time.
Cumulative Distribution Function – A function $F(x)$ which has the property $F(c) = P(x \leq c)$.
Data – Information collected by a researcher. The definition "facts in the form of numbers" is not strictly correct, but is illustrative.
Data Set – A set of data collected for some task or from some given source.
Deduction – Deductive reasoning is often called "reasoning from the general to the partial." It is the process of reasoning that draws a conclusion from an assumed general truth.
Descriptive Statistics – Summarizing, presenting and organizing quantitative data.
Discrete data – Data made up of variables that can only take a countable number of values. Usually the number of possible values is finite.
Efficiency – A measure of the variance of the sampling distribution of an estimator. The estimator with the smallest variance is called **best.**
Five number summary – The five numbers are a lower limit, the first quartile, the median, the third quartile and an upper limit.
Fractile (or quantile) – A value (x_{1-p}) that has a certain fraction (p) of data below it, for example, $x_{.26}$ the .74 fractile or the $\frac{1}{3}$ fractile. Examples include the following:

Quartiles – Values below which are $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{3}{4}$ of the data. These are $x_{.75}$, $x_{.50}$ and $x_{.25}$ respectively.

Quintiles - Values below which are $\frac{1}{5}$, $\frac{2}{5}$, $\frac{3}{5}$ and $\frac{4}{5}$ of the data.

Deciles - Values below which are $\frac{1}{10}$, $\frac{2}{10}$, $\frac{3}{10}$ ..... and $\frac{9}{10}$ of the data.

Percentiles - Values below which are1%, 2%, 3% ..... and 99% of the data.

Frame – A list of the members of a population. A sample can be selected from this list.
Frequency – The number of items falling into a category. **Relative Frequency** – The fraction or percent of items in a population or sample that fall into a given category. f is generally used for frequency and F is generally used for cumulative frequency, which is the total number of items up to a given point.
Frequency Distribution – A table or chart that shows the classes into which data has been grouped and how many items or what proportion of items there are in each class.

Profit Rate	f	f_{rel}
9-10.99%	3	.200
11-12.99%	3	.200
13-14.99%	5	.333
15-16.99%	3	.200

17-18.99%	1	.067
Total	15	1.000

In the table above, the lower limits of the classes are 9, 11, 13, 15 and 17, the (width of the) class interval is 2 and the midpoints of the classes are 10, 12, 14, 16 and 18.

Grouped data – Data that is only available as a frequency distribution.

Induction - Inductive reasoning is often called "reasoning from the partial to the general." This is often called statistical inference and is the process of reasoning that draws a conclusion about a population from analysis of a sample. In statistics, an explicit or implicit reference to probability is involved.

Infinite and finite populations – A population is usually considered infinite if removing a sample from it will have little effect. If we sample with replacement, we in effect observe individual units of the population and then, in effect, throw them back into the population, leaving it unchanged, so that we can consider it infinite. If we sample without replacement we are taking a sample of n items from a population of N items in such a way that, if an item is chosen to be in the sample, it cannot be chosen again as part of the same sample. As a rule of thumb, if we are sampling without replacement we can usually get away with considering the population infinite if the sample is less that $1/20$ or 5% of the population

Index – This is defined as a number that is used to express the relationship between two values of one variable or between two variables simply. Most commonly in time series, one year is designated as the base year and the values for every other year are expressed as a percentage of the base year value.

Maximum Likelihood Estimator – The value of a parameter most likely to have produced the data actually observed.

Observation – In a table or the equivalent all the numbers relating to one unit of observation at a given time. For example in a table giving the GDP and population of every country, the GDP and population of the US would be one observation.

Parameter – A number that characterizes a population, such as a population mean or variance. These are often represented by small Greek letters.

Per capita – Per person. For example, income per capita is some measure of total income like GDP divided by total (human) population. **Question:** What is the population per capita of the US?

Population – All of the persons or things that are under investigation. Also called a Universe.

Primary source – The original source of a data set. Presumably, this is the best place to assess methodology and its use minimizes transcription errors.

Qualitative data – Data which is not quantitative. This refers to data that is descriptive and cannot usefully be manipulated by mathematical operations like addition, subtraction, multiplication and division.

Quantitative data – Ordinary numerical data. Such data should be, at least manipulable by addition and subtraction.

Real value – This is not the same as 'real' in a mathematical sense. This refers to an economic quantity that has been adjusted to eliminate price changes. For example, real GDP is calculated by evaluating goods and services at the prices of a base year.

Secondary source – A source that is not the original source of data.

Significance – In this course always **statistical significance.**

Significance Level – The probability of a given statistical test rejecting a null hypothesis when the null hypothesis is true. Usually indicated by α .

Stability of Mass Data – The idea that most randomly selected parts of a population will display characteristics similar to that of the population as a whole.

Statistic – Usually a sample statistic that describes a characteristic of a sample, like the sample mean and variance. Conventionally a sample statistic is indicated by a Latin letter.

Statistical significance – A difference between a statistic and a parameter is statistically significant if it is larger than would be expected by chance alone. Two statistics are **significantly different** if their difference is significantly different from zero at some significance level. If we say that two estimates of the mean are significantly different at the 5% level, we mean that there is at most a 5% chance that the difference is as large as or larger than the one actually observed. This means that we have done a statistical (hypothesis) test or the equivalent (generally a confidence interval). If we say that a coefficient or difference is (statistically) significant it means that we have shown by a statistical test or a confidence interval that it is not zero. Statistical significance does not imply economic or practical significance.

Statistical test – A method for evaluating a statistical hypothesis. Usually this consists of stating null and alternate hypotheses and evaluating them using a test statistic, a test ratio or a confidence interval. An explicit or implicit reference to a table or a probability is expected. A confidence interval is usually considered equivalent to a statistical test. In this course the same as a **hypothesis test.**
Statistics – The description and analysis of data by mathematical means.
Target and sampled populations – A target population is the population of interest and the sampled population is the population that is actually sampled, which may not be exactly the same as the target population.
Time series data – Data that is observed or reported at regular intervals of time, for example, end of month inventory reported over 120 months.
Unbiassed – An estimator is unbiased if its expected value is the parameter it is supposed to estimate.
Unit of observation – What is being investigated. Usually the persons, things or groups that make up a population. For example, in a survey of households the unit of observation will generally be the household.
Weighted Average – A sum of the form $\mu = \frac{\sum wx}{\sum w}$ or $\bar{x} = \frac{\sum wx}{\sum w}$ where the WS are positive numbers. Equivalently $\mu = \sum wx$ or $\bar{x} = \sum wx$ where the WS are positive and $\sum w = 1$.
Width of a class interval – The distance between lower limits of adjacent classes in a frequency distribution.