AP Statistics Important Vocabulary Terms
Prepared by Woody Nivens, Laurens High School

Alternative Hypothesis: states that a treatment has had an effect or caused a change in the population

Bias: describes a study which systematically favors certain outcomes

Binomial Distribution: the distribution of the probabilities of X successes out of n trials, calculated using *p* as the probability of any single success – B(*n*, *p*)

Blind: describes an experiment in which the subjects do not know which treatment they are getting

Blocking: a statistical design which creates groups that are similar in some way, and then randomizes the treatments within each block

Central Limit Theorem: states that when an SRS is drawn from a population with mean μ and standard deviation σ, the sampling distribution for the sample mean will be approximately normally distributed, and have a mean μ and a standard deviation $\sigma/\sqrt{n}$

Chi-Square Distributions: a family of skewed-right distributions which take on only positive values and are defined by their degrees of freedom – the specific shape of the Chi-Square Distribution changes as the sample size changes

Chi-Square Goodness-of-Fit Test: used to determine if a population has a certain hypothesized distribution

Chi-Square Test for Homogeneity: used to determine if every category in the population has the same population

Chi-Square Test for Independence: used to determine if there is a relationship between two categorical variables – also known as *Chi-Square Test for Association*

Coefficient of Determination: tells what percent of the change in the response variable can be attributed to the change in the explanatory variable – symbolized as $\boldsymbol{r^2}$

Complement of an Event: the set of all outcomes not defined as successful outcomes for any event

Conditional Probability: the probability of an event occurring if it is known that another specific event has already occurred

Confidence Interval: an interval estimate of a parameter calculated using a sample from that population

Confidence Level: the probability that the desired parameter will fall into a confidence interval if many intervals were calculated from samples of the same size

Confounding Variable: a variable which could affect the result of a statistical test but has not been controlled for

Continuous Random Variable: a random variable which takes on all values in an interval of numbers

Control Group: any group of subjects who receive either a placebo or no treatment at all during an experiment

Correlation: measures the direction and strength of the linear relationship between two quantitative variables – symbolized as ***r***

Critical Value: a value (z-score, t-score, or χ^2 value) used in a hypothesis test to help determine if the null hypothesis should be rejected

Cumulative Distribution Function: A function which calculates the sum of the probabilities for each possible value for any random variable X

Degrees of Freedom: a value used to help determine significance for a t-test or a Chi-Square test – measured as n-1 in most cases, or (r-1)(c-1) when dealing with two-way tables

Dependent Trials: trials whose probability is affected by the outcome of previous trials

Dependent Variable: see *Response Variable*

Density Curve: a curve used to represent a distribution – a density curve is always on or above the horizontal axis and has a total area of exactly 1 underneath it

Discrete Random Variable: a random variable with countable outcomes

Disjoint Events: events which cannot occur at the same time – also known as *Mutually Exclusive Events*

Distribution: a list of what values a variable takes on and how often it takes on each one of those values

Double Blind: describes an experiment in which neither the subjects nor the researcher know which treatment each subject is getting

Empirical Rule: also known as the 68-95-99.7 rule – is used as an approximation for what percent of the data falls within 1, 2, or 3 standard deviations of the mean in any normal distribution

Expected Value: see *Mean*

Experimental Units: the individuals on which an experiment is conducted – if the test is being conducted on humans, the units are called **Subjects**

Explanatory Variable: attempts to explain the observed outcomes in a statistical study – also known as the *Independent Variable*

Exploratory Data Analysis: uses graphs and numerical summaries to describe the variables in a data set and the relationships among them

Factor: any explanatory variable in an experiment

Five Number Summary: a method to describe a data set using the minimum, first quartile, median, third quartile, and maximum points in the data set

Geometric Distribution: a distribution of probabilities of when the first successful outcome occurs in a probability experiment

Hypothesis Test: a type of inference used to determine the feasibility of an assumed population parameter – also known as a *Significance Test*

Independent Trials: trails whose probabilities are not affected by the outcome of previous trials

Independent Variable: see *Explanatory Variable*

Individuals: people or objects described by a set of data

Inference: the statistical process of drawing conclusions about a population by examining data from a sample

Influential Point: a point which, if removed from the data set, would markedly change the regression equation for that data set

Interquartile Range (IQR): the difference between the third and first quartiles of a data set

Law of Large Numbers: states that as increased numbers of observations are drawn from any population, the mean of the observations eventually approaches the mean of the population as closely as we would like to estimate it, and remains that close or closer

Least Squares Regression Line: a regression line which makes the sum of the squares of the vertical distances from the data points to the line as small as possible

Level: a numerical value of a factor of an experiment

Matched Pairs: a statistical design which compares two treatments – this is usually done with one sample receiving each treatment over a different time period

Mean: the "average" of a data set – also known as the *Expected Value*

Median: the point at which 50% of the data is above and 50% of the data is below

Mutually Exclusive Events: see *Disjoint Events*

Nonresponse: a type of bias that occurs when an individual chosen for a sample cannot be contacted or chooses not to participate

Normal Distribution: a symmetric, bell-shaped distribution in which approximately 68% of the data lies within one standard deviation of the mean, 95% lies within two standard deviations of the mean, and 99.7% lies within three standard deviations of the mean – all normal distributions can be defined by their mean and standard deviation

Null Hypothesis: states that either a treatment has had no effect on a population, or that the population has not changed

Observation: any single point from a data set

Outlier: an individual observation that falls outside the pattern of the data set – often defined as any number that is 1.5(IQR) outside of Q1 or Q3

P-value: the probability that the observed outcome would take on a value as extreme or more extreme than observed if the null hypothesis were true

Parameter: a number that describes a population

Percentile: tells what percent of a data set falls below the given observation

Placebo: a false treatment which should have no effect on an experiment – placebos should appear to be the same as the actual treatment

Pooled Procedures: occurs when separate samples are combined into a single sample for analysis – this should only be done if it is known that the variances of the two populations are equal

Population: the entire group of individuals that we want information about

Power of a Hypothesis Test: the probability that the test will reject the null hypothesis when the null hypothesis is false – the power is equal to 1 minus (probability of a Type II error for the given alternative)

Probability: the proportion of times an outcome would occur over a large number of trials

Probability Distribution Function: a function which assigns a probability for each possible value for any discrete random variable X

Proportion: tells what percent of a data set falls into a given category

Qualitative Variable: a variable which takes on a non-numeric description

Quantitative Variable: a variable which takes on a numeric value

Quartiles: observations which fall at the 25^{th}, 50^{th}, and 75^{th} percentiles of a data set

Range: the difference between the maximum and minimum values of a data set

Random: when individual outcomes are uncertain, but there is a pattern to the distribution of the outcomes over time

Random Variable: a variable whose value is a numeric outcome of a random phenomenon

Randomization: using the laws of probability – this is done to select members for a sample and also to assign treatments to specific samples in experiments

Regression Line: a straight line that describes how a response variable changes as the explanatory variable changes

Residual: the difference between and observed value of a response variable and its predicted value from a regression equation

Response Variable: measures the outcome of a statistical study – also known as the *Dependent Variable*

Robustness: a measure of how much the P-value of a test is affected if the conditions of the hypothesis test are not met

Sample: a part of the population used to gather information about the entire population

Sample Space: a list of all possible outcomes for a random event

Sampling Distribution: a distribution of values taken by a statistic in all possible samples of the same size from the same population

Sampling Frame: a list from which a sample is chosen – ideally the sampling frame consists of the entire population

Significance Level: the point at which it will be determined that a result is statistically significant

Significance Test: see *Hypothesis Test*

Simple Random Sample (SRS): a sample in which every member of the population has the same probability to be chosen, and every group of size *n* has the same probability to be chosen

Simulation: a method for collecting data which uses the laws of probability to represent all possible outcomes of an experiment

Skewed: describes a distribution whose histogram extends much farther to one side of the mean than the other – the distribution is said to be skewed in the direction of this "tail"

Standard Deviation: square root of the variance – used as a common measure of spread for a data set

Standard Error: the standard deviation of a sampling distribution – measures the amount of expected error per standard deviation from the mean of the distribution

Standard Normal Distribution: a normal distribution with a mean of zero and a standard deviation of one

Standardized Score: see *z*-Score

Statistic: a number that describes a sample

Statistically Significant: an observed effect so far removed from the mean that it would be unlikely to occur by chance alone

Stratified Random Sample: a sample chosen by splitting the population into several well-defined groups, then taking an SRS from each group

Subjects: see *Experimental Units*

Symmetric: describes a distribution whose histogram has its left and right sides as mirror images of each other

***t*-Distributions:** a family of symmetric, bell-shaped distributions with a standard deviation larger than that of the standard normal distribution – the specific shape of the *t*-distribution changes as the sample size changes – this distribution is defined by its degrees of freedom

Treatment: a specific experimental condition applied to an experimental unit or subject

Treatment Group: a group of subjects who receive an actual treatment during an experiment

Type I Error: when the null hypothesis is rejected but it is in fact true – the probability of a Type I Error is the significance value for that test

Type II Error: when the null hypothesis is not rejected but it is in fact false – the probability of a Type II Error must be calculated for a specific alternative test value

Unbiased Statistic: a statistic from a sampling distribution whose mean must be equal to the mean of the population

Undercoverage: a type of bias that occurs when some groups of a population are left out of the selection process for the sample

Variability: describes the spread of a data set

Variable: any characteristic of an individual

Variance: the average of the squares of the deviations of the observation from their mean – used as a measure of spread for a data set

Voluntary Response Sample: consists only of people who choose to participate – a poor method for collecting meaningful data

***z*-Score:** a measure used to tell how many standard deviations above or below the mean an observation lies – also known as a *Standardized Score*