15.1: Glossary of Statistical Terms used in Inference

Additive Rule

In probability, for events A and B, P(A or B) = P(A) +P(B) - P(A and B).

Alpha (α) – see Level of Significance

Alternative Hypothesis (H_a) A statement about the value of a population parameter that is assumed to be true if the Null Hypothesis is rejected during testing.

Analysis of Variance (ANOVA)

A group of statistical tests used to determine if the mean of a numeric variable (the Response) is affected by one or more categorical variables (Factors).

Bar Graphs

A graph of categorical data in which the height of the bar represents the frequency of each choice. Bar graphs can be clustered or stacked for multiple categorical variables.

Bernoulli Distribution

A probability distribution function (parameter p) for a discrete random variable that is the numbers of successes in a single trial when there are only two possible outcomes (success or failure).

Beta (β)

The probability, set by design, of failing to reject the Null Hypothesis when it is actually false. Beta is calculated for specific possible values of the Alternative Hypothesis.

Biased Sample

A sample that has characteristics, behaviors and attitudes from the population from which the sample is selected -- in other words a non-representative sample.

Binomial Distribution

A probability distribution function (parameters n, p) for a discrete random variable that is the numbers of successes in a fixed number of independent trials when there are only two possible outcomes (success or failure).

Bivariate Data

Pairs of numeric data; there are two variables or measurements per observation.

Box Plot

A graph that represent the 3 quartiles (Q1, median and Q3), along with the minimum and maximum values of the data.

Blinding

In an experiment, blinding is keeping the participant and/or the administrator unaware as to what treatment is being given. A **single blind study** is when the participant does not know whether the treatment is real or a placebo. A **double blind study** is when neither the administrator of the treatment nor the participant knows whether the treatment is real or a placebo.

Categorical data

Non-numeric values. Some examples of categorical data include eye color, gender, model of computer, and city.

Central Limit Theorem

A powerful theorem that allows us to understand the distribution of the sample mean, $\bar{X}$. If $X_1, X_2, \ldots, X_n$ is a random sample from a probability distribution with mean = μ and standard deviation = σ and the sample size is "sufficiently large", then $\bar{X}$ will have a Normal Distribution with the same mean and a standard deviation of $\sigma/\sqrt{n}$ also known as the Standard Error). Because of this theorem, most statistical inference is conducted using a sampling distribution from the Normal Family.

Class Intervals

For grouped numeric data, one category, usually of equal width, in which values are counted.

Chi-square Distribution (χ^2)

A family of continuous random variables (based on degrees of freedom) with a probability density function that is from the Normal Family of probability distributions. The Chi-square distribution is non- negative and skewed to the right and has many uses in statistical inference such as inference about a population variance, goodness-of-fit tests and test of independence for categorical data.

Chi-square Goodness-of-fit Test

A test that is used to test if **observed** data from a categorical variable is consistent with an **expected** assumption about the distribution of that variable.

Chi-square Test of Independence

A test to determine if there is a relationship between two randomized categorical variables

Chi-square Test of Homogeneity

A test that is run the same way as a **Chi-square Test of Independence**, but in which only one of the categorical variables is randomized.

Classical probability (also called Mathematical Probability)

Determined by counting or by using a mathematical formula or model..

Cluster Sample

A sample that is created by first breaking the population into groups called clusters, and then by taking a sample of clusters.

Complement of an Event

The complement of an event means that the event does not occur. If the event is labeled A, then the complement of A is labeled A' and read as "not A".

Conditional Probability

The probability of an event A occurring given that another event B has already occurred. This probability is written as **P(A|B)** which is read as **P(A given B)**.

Confidence Interval

An interval estimate that estimates a population parameter from a random sample using a predetermined probability called the level of confidence.

Confidence Level

see Level of Confidence

Confounding Variable

A lurking variable that is not known to the researcher, but that affects the results of the study.

Contingency Tables

A method of displaying the counts of the responses of two categorical variables from data, also known as **cross tabulations**, or **two-way tables**.

Control Group

In an experiment, the group that receives no treatment giving the researcher a baseline to be able to compare the treatment and placebo groups.

Continuous data

Quantitative based on the real numbers. Some examples of continuous data include time to complete an exam, height, weight. Continuous data are values that are measured, or answers the question "How much"?

Continuous Random Variable

A random variable that has only continuous values. Continuous values are uncountable and are related to real numbers.

Correlation Coefficient

A measure of correlation(represented by the letter r) that measures both the direction and strength of a linear relationship or association between two variables. The value r will always take on a value between -1 and 1. Values close to zero imply a very

weak correlation. Values close to 1 or -1 imply a very strong correlation. The correlation coefficient should not be used for non-linear correlation.

Critical value(s)

The dividing point(s) between the region where the Null Hypothesis is rejected and the region where it is not rejected. The critical value determines the decision rule

Cross Tabulations

see Contingency Tables

Cumulative Frequency

In grouped data, the number of times a particular value is observed in a class interval or in any lower class interval.

Cumulative Relative Frequency

In grouped data, the proportion or percentage of times a particular value is observed in a class interval or in any lower class interval.

Data Dredging

see p-hacking

Decision Rule

The procedure that determines what values of the result of an experiment will cause the Null Hypothesis to be rejected. There are two methods that are equivalent decision rules:

1. If the test statistic lies in the Rejection Region, Reject H_o (Critical Value method).
2. If the p-value < α, Reject H_o (p-value method).

Dependent Events

Two events are dependent if the probability of one event occurring is changed by knowing if the other event occurred or not. Events that are not dependent are called independent.

Dependent Sampling

A method of sampling in which 2 or more variables are related to each other (paired or matched). Examples would be the "Before and After" type models using the Matched Pairs t-test.

Discrete data

Quantitative natural numbers (0, 1, 2, 3, ...). Some examples of discrete data include number of siblings, friends on Facebook, or bedrooms in a house. Discrete data are values that are counted, or where you might ask the question "How many"?

Discrete Random Variable

A random variable that has only discrete values. Discrete values are related to counting numbers.

Dot Plot

A graph of numeric data in which each value is represented as a dot on a simple numeric scale. Multiple values are stacked to create a shape for the data. If the data set is large, each dot can represent multiple values of the data.

Effect Size

The "practical difference" between a population parameter under the Null Hypothesis and a selected value of the population parameter under the Alternative Hypothesis.

Empirical probability

Probability that is based on the relative frequencies of historical data, studies or experiments.

Empirical Rule

(Also known as the 68-95-99.7 Rule). A rule used to interpret standard deviation for data that is approximately bell-shaped. The rule says about 68% of the data is within one standard deviation of the mean, 95% of the data is within two standard deviations of the mean, and about 99.7% of the data is within three standard deviations of the mean.

Estimation

An inference process that attempts to predict the values of population parameters based on sample statistics

Event

A result of an experiment, usually referred to with a capital letter A, B, C, etc.

Expected Value

A value that describes the central tendency of a random variable, also known as the population mean and that is expressed by the symbol (pronounced mu). The expected value is a parameter, meaning a fixed quantity

Experiment

A study in which the researcher will randomly break a representative sample into groups and then apply treatments in order to manipulate a variable of interest. The goal of an experiment is to find a cause and effect relationship between a random variable in the population and the variable manipulated by the researcher.

Exponential Distribution

A probability distribution function (parameter μ) for a continuous random variable that models the waiting time until the first occurrence of an event defined by a Poisson Process.

Explanatory Variable

The variable that the researcher controls or manipulates

***F* Distribution**

A family of continuous random variables (based on 2 different degrees of freedom for numerator and denominator) with a probability density function that is from the Normal Family of probability distributions. The F distribution is non-negative and skewed to the right and has many uses in statistical inference such as inference about comparing population variances, ANOVA, and regression.

Factor

In ANOVA, the categorical variable(s) that break the numeric response variable into multiple populations or treatments.

Frequency

In grouped data, the number of times a particular value is observed.

Frequency distribution

An organization of numeric data into class intervals.

Geometric Distribution

A probability distribution function (parameter p) for a discrete random variable that is the number of independent trials until the first success in which there are only two possible outcomes (success or failure).

Hypothesis

A statement about the value of a population parameter developed for the purpose of testing

Hypothesis Testing

A procedure, based on sample evidence and probability theory, used to determine whether the hypothesis is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

Independent Events

Two events are independent if the probability of one event occurring is not changed by knowing if the other event occurred or not. Events that are not independent are called dependent.

Independent Sampling

A method of sampling in which 2 or more variables are not related to each other. Examples would be the “Treatment and Control” type models using the independent samples *t*-test.

Inference

– see Statistical Inference

Interquartile Range (IQR)

A measure of variability that is calculated by taking the difference of the 1st quartile and 3rd quartiles.

Interval Estimate

A range of values based on sample data that is used to estimate a population parameter

Interval Level of Data

Quantitative data that have meaningful distance between values, but that do not have a "true" zero. Interval data are numeric, but zero is just a place holder. Examples of interval data include temperature in degrees Celsius and year of birth.

Joint Probability

The probability of the union or intersection of multiple events occurring. If A and B are multiple events, then P(A or B) and P(A and B) are examples of joint probability.

Level

In ANOVA, a possible value that a categorical variable factor could be. For example, if the factor was shirt color, levels would be blue, red, yellow, etc.

Level of Confidence

The probability, usually expressed as a percentage, that a Confidence Interval will contain the true population parameter that is being estimated.

Level of Significance (α)

The maximum probability, set by design, of rejecting the Null Hypothesis when it is actually true (maximum probability of making Type I error).

Levels of Data

The four levels of data are Nominal, Ordinal, Interval and Ratio.

Lurking Variable

see Confounding Variable

Margin of Error

The distance in a symmetric Confidence Interval between the Point Estimator and an endpoint of the interval. For example a confidence interval for μ may be expressed as $\bar{X} \pm$Margin of Error.

Marginal Probability

The probability a single event A occurs, written as P(A).

Mean

see Population Mean or Sample Mean

Median

see Population Median or Sample Median

Mode

see Population Mode or Sample Mode

Model Assumptions

Criteria that must be satisfied to appropriately use a chosen statistical model. For example, a Student's t statistic used for testing a population mean vs. a hypothesized value requires random sampling and that the sample mean has an approximately Normal Distribution.

Multiplicative Rule

In probability, for events A and B, **P(A and B) = P(A)P(B|A) = P(B)P(A|B)**.

Mutually Exclusive Events

Events that cannot both occur; the intersection of two events has no possible outcomes.

Nominal Level of Data

Qualitative data that only define attributes, with no hierarchal ranking. Examples of nominal data include hair color, ethnicity, gender and any yes/no question.

Non-probability Sampling Methods

Non-scientific methods of sampling that have immeasurable biases and should not be used in scientific research. These methods include Convenience Sampling and Self-selected sampling.

Non-response Bias

A type of sampling bias that occurs when people are intentionally or non-intentionally excluded from participation or choose not to participate in a survey or poll. Sometimes people will lie to pollsters as well.

Normal Distribution

Often called the "bell-shaped" curve, the Normal Distribution is a continuous random variable which has Probability Density Function $X = \exp\left[-(x-\mu)^2/2\sigma^2\right]/\sigma\sqrt{2\pi}$ .The special case where $\mu = 0$ and $\sigma = 1$, is called the Standard Normal Distribution and is designated by Z.

Normal Family of Probability Distributions

The Standard Normal Distribution (Z) plus other Probability Distributions that are functions of independent random variables with Standard Normal Distribution. Examples include the t, the F and the Chi-square distributions.

Null Hypothesis (H_o)

A statement about the value of a population parameter that is assumed to be true for the purpose of testing

Observational Study

A study in which the researcher takes measurements from a representative sample, but does not manipulate any of the variables with treatments. The goal of an observational study is to interpret and analyze the measured variables, but it is not possible to show a cause and effect relationship.

Ogive

A line graph in which the vertical axis is cumulative relative frequency and the horizontal axis is the value of the data, specifically the endpoints of the class intervals. The left end point of the first class interval will have a cumulative relative frequency of zero. All other endpoints are given the right endpoint of the corresponding class interval. The points are then connected by line segments. The ogive can be used to estimate percentiles.

Outcome

A result of the experiment which cannot be broken down into smaller events.

Ordinal Level of Data

Qualitative data that define attributes with a hierarchal ranking. Examples of nominal data include movie ratings (G, PG, PG13, R, NC17), t-shirt size (S, M L, XL), or your letter grade on a term paper.

Outlier

A data point that is far removed from the other entries in the data set.

***p*-value**

The probability, assuming that the Null Hypothesis is true, of getting a value of the test statistic at least as extreme as the computed value for the test.

***p*-hacking**

An improper research method that uses repeated experiments or multiple measures analysis until the researcher obtains a significant p-value. Also known as Data Dredging.

Parameter

A fixed numerical value that describes a characteristic of a population.

Percentile

The value of the data below which a given percentage of the data fall.

Pie Chart

A circular graph of categorical data where each slice of the pie represents the relative frequency or percentage of data in each category.

Placebo

A treatment with no active ingredients.

Placebo Effect

In an experiment, when a participant responds in a positive way to a placebo, a treatment with no active ingredients.

Placebo Group

In an experiment, the group that receives the treatment with no active ingredients.

Point Estimate

A single sample statistic that is used to estimate a population parameter. For example, $\bar{X}$ is a point estimator for μ.

Poisson Distribution

A probability distribution function (parameter μ) for a discrete random variable that is the number of occurrences in a fixed time period or region, over which the rate occurrences is a constant.

Poisson Process

Counting methods that are modeled by random variables that follow a Poisson Distribution.

Population

The set of all possible members, objects or measurements of the phenomena being studied.

Population Mean

see Expected Value

Population Median

A value that describes the central tendency of a random variable that represents the 50th percentile. The population median is a parameter, meaning a fixed quantity.

Population Mode

The maximum value or values of a probability density function.

Population Variance

The expected value of the squared deviation from the mean, a value that describes the variability of a random variable expressed by the symbol σ^2 (pronounced sigma-squared). The population variance is a parameter, meaning a fixed quantity.

Population Standard Deviation

The square root of the population variance, a value that describes the variability of a random variable expressed by the symbol σ (pronounced sigma).

Power (or Statistical Power)

The probability, set by design, of rejecting the Null Hypothesis when it is actually false. Power is calculated for specific possible values of the Alternative Hypothesis and is the complement of Beta (β).

Probability

The measure of the likelihood that an event A will occur. This measure is a quantity between 0 (never) and 1 (always) and will be expressed as P(A) (read as “The probability event A occurs.”)

Probability Density Function (pdf)

A non-negative function that defines probability for a Continuous Random Variable. Probability is calculated by measuring the area under a probability density function.

Probability Distribution Function (PDF)

A function that assigns a probability to all possible values of a discrete random variable. In the case of a continuous random variable (like the Normal Distribution), the PDF refers to the area to the left of a designated value under a Probability Density Function.

Probability Sampling Methods

Sampling methods that will usually produce a sample that is representative of the population. These methods are also called scientific sampling. Examples include Simple Random Sampling, Systematic Sampling, Stratified Sampling and Cluster Sampling.

Qualitative Data

Non-numeric values that describe the data. Note that all quantitative data is numeric, but some numbers without quantity (such as Zip Code or Social Security Number) are qualitative. When describing categorical data, we are limited to observing counts in each group and comparing the differences in percentages.

Quantitative Data

Measurements and numeric quantities that can be determined from the data. When describing quantitative data, we can look at the center, spread, shape and unusual features.

Quartile

The 25th, 50th and 75th percentiles, which are usually called, respectively, the 1st quartile, the median, and the 3rd quartile.

Radix

A convenient total used in creating a hypothetical two-way table.

Random Sample

see Simple Random Sample

Range

For numeric data, the maximum value minus the minimum value.

Random Variable

A variable in which the value depends upon an experiment, observation or measurement.

Ratio Level of Data

Quantitative data that have meaningful distance between values, and have a "true" zero. Examples of ratio data include time to drive to work, weight, height, or number of children in a family. Most numeric data will be ratio.

Raw Data

Sample data presented unsorted.

Regression Analysis

A method of modeling correlated bivariate data.

Relative frequency

In grouped data, the proportion or percentage of times a particular value is observed.

Replicate

In ANOVA, the sample size for a specific level of factor. If the replicates are the same for each level, the design is balanced.

Rejection Region

Statistical Model region(s) which contain the values of the Test Statistic in which the Null Hypothesis will be rejected. The total area of the Rejection Region $= \alpha$.

Representative Sample

A sample that has characteristics, behaviors and attitudes similar to the population from which the sample is selected.

Response Variable

The numeric variable that is being tested under different treatments or populations.

Response bias

A type of sampling bias that occurs when the responses to a survey are influenced by the way the question is asked, or when responses do not reflect the true opinion of the respondent. When conducting a survey or poll, the type, order, and wording of questions are important considerations. Poorly worded questions can invalidate the results of a survey.

Rule of Complement

If the events A and A' are complements, then **P(A) +P(A') =1**.

Sample

A subset of the population that is studied to collect or gather data.

Sample Size

The number of observations in your sample size, usually represented by n.

Sample Mean

a. The arithmetic average of a numeric data set.
b. A random variable that has an approximately Normal Distribution if the sample size is sufficiently large.
c. An unbiased estimator for the population mean

Sample Median

The value that represents the exact middle of data, when the values are sorted from lowest to highest.

Sample Mode

The most frequently occurring value in the data. If there are multiple values that occur most frequently, then there are multiple modes in the data.

Significance Level

see Level of Significance

Sample Space

In probability, the set of all possible outcomes of an experiment.

Sample Standard Deviation

The square root of the sample variance, which measures the spread of data and distance from the mean. The units of the standard deviation are the same units as the data.

Sample Variance

A measure of the mean squared deviation of the data values from the mean. The units of the variance are the square of the units of the data.

Scatterplot

A graph of bivariate data used to visualize correlation between the two numeric variables.

Selection Bias

A type of sampling bias that occurs when the sampling method does not create a representative sample for the study. Selection bias frequently occurs when using convenience sampling

Self-selection Bias

A type of sampling bias that occurs when individuals can volunteer to be part of the study. Volunteers will often have a stronger opinion about the research question and will usually not be representative of the population.

Simple Random Sample

A subset of a population in which each member of the population has the same chance of being chosen and is mutually independent from all other members.

Skewness

A measure of how asymmetric the data vales are.

Standard Deviation

see Sample Standard Deviation or Population Standard Deviation

Standard Normal Distribution

A special case of the **Normal Distribution** where $\mu = 0$ and $\sigma = 1$. The symbol Z is usually reserved for the Standard Normal Distribution.

Statistic

A value that is calculated from only the sample data, and that is used to describe the data. Examples of statistics are the sample mean, the sample standard deviation, the range, the sample median and the interquartile range. Since statistics depend on the sample, they are also random variables.

Statistical Inference

The process of estimating or testing hypotheses of population parameters using statistics from a random sample.

Statistical Model

A mathematical model that describes the behavior of the data being tested.

Stem and Leaf Plot

A method of tabulating data by splitting it into the "stem" (the first digit or digits) and the "leaf" (the last digit, usually). For example, the stem for 102 minutes would be 10 and the leaf would be 2.

Stratified Sample

A sample that is designed by breaking the population into subgroups called strata, which are then sampled so that the proportion of each subgroup in the sample matches the proportion of each subgroup in the population.

Student's t Distribution (or t Distribution)

A family of continuous random variables (based on degrees of freedom) with a probability density function that is from the Normal Family of Probability Distributions. The t distribution is used for statistical inference of the population mean when the population standard deviation is unknown.

Subjective probability

Probability that is a "one-shot" educated guess based on anecdotal stories, intuition or a feeling as to whether an event is likely, unlikely or "50-50". Subjective probability is often inaccurate.

Systematic Sample

A subset of the population in which the first member of the sample is selected at random and all subsequent members are chosen by a fixed periodic interval.

t Distribution

see Student's t Distribution

Test Statistic

A value, determined from sample information, used to determine whether or not to reject the Null Hypothesis.

Treatment Group(s)

In an experiment, the group(s) that receive the treatment that the researcher controls.

Tukey HSD Test

In ANOVA, a post-hoc collection of tests that report honest significant differences in pair of means.

Tree Diagram

A simple way to display all possible outcomes in a sequence of events. Each branch will represent a possible outcome. Using the Multiplicative Rule, the probability of each possible outcome can be calculated.

Two-way Tables

see Contingency Tables

Type I Error

Rejecting the Null Hypothesis when it is actually true.

Type II Error

Failing to reject the Null Hypothesis when it is actually false.

Uniform Distribution

A probability distribution function (parameters a, b) for a continuous random variable in which all values between a minimum value and a maximum value have the same probability.

Variance

see Sample Variance or Population Variance

***Z*-score**

A measure of relative standing that shows the distance in standard deviations that a particular data point is above or below the mean.
