Glossary

start

Average

also called mean or arithmetic mean; a number that describes the central tendency of the data

Blinding

not telling participants which treatment a subject is receiving

Categorical Variable

variables that take on values that are names or labels

Cluster Sampling

a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

Continuous Random Variable

a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV.

Control Group

a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups

Convenience Sampling

a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

Cumulative Relative Frequency

The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Data

a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Discrete Random Variable

a random variable (RV) whose outcomes are counted

Double-blinding

the act of blinding both the subjects of an experiment and the researchers who work with the subjects

Experimental Unit

any individual or object to be measured

Explanatory Variable

the **independent variable** in an experiment; the value controlled by researchers

Frequency

the number of times a value of the data occurs

Informed Consent

Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant.

Institutional Review Board

a committee tasked with oversight of research programs that involve human subjects

Lurking Variable

a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

Mathematical Models

a description of a phenomenon using mathematical concepts, such as equations, inequalities, distributions, etc.

Nonsampling Error

an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

Numerical Variable

variables that take on values that are indicated by numbers

Observational Study

a study in which the independent variable is not manipulated by the researcher

Parameter

a number that is used to represent a population characteristic and that generally cannot be determined easily

Placebo

an inactive treatment that has no real effect on the explanatory variable

Population

all individuals, objects, or measurements whose properties are being studied

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion

the number of successes divided by the total number in the sample

Qualitative Data

See Data.

Quantitative Data

See Data.

Random Assignment

the act of organizing experimental units into treatment groups using random methods

Random Sampling

a method of selecting a sample that gives every member of the population an equal chance of being selected.

Relative Frequency

the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes

Representative Sample

a subset of the population that has the same characteristics as the population

Response Variable

the **dependent variable** in an experiment; the value that is measured for change at the end of an experiment

Sample

a subset of the population studied

Sampling Bias

not all members of the population are equally likely to be selected

Sampling Error

the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

Sampling with Replacement

Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

Sampling without Replacement

A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

Simple Random Sampling

a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample.

Statistic

a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Statistical Models

a description of a phenomenon using probability distributions that describe the expected behavior of the phenomenon and the variability in the expected observations.

Stratified Sampling

a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Conditional Probability

the likelihood that an event will occur given that another event has already occurred

Contingency Table

the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities.

Dependent Events

If two events are NOT independent, then we say that they are dependent.

Equally Likely

Each outcome of an experiment has the same probability.

Event

a subset of the set of all outcomes of an experiment; the set of all outcomes of an experiment is called a sample space and is usually denoted by S. An event is an arbitrary subset in S. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital letters such as A, B, C, and so on.

Experiment

a planned activity carried out under controlled conditions

$P(A|B) = P(A)$
$P(B|A) = P(B)$
$P(A \cap B) = P(A)P(B)$

Independent Events

The occurrence of one event has no effect on the probability of the occurrence of another event. Events A and B are independent if one of the following is true:

Mutually Exclusive

Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events A and B are mutually exclusive, then $P(A \cap B) = 0$.

Outcome

a particular result of an experiment

$0 \leq P(A) \leq 1$
If A and B are any two mutually exclusive events, then
$P(A \cup B) = P(A) + P(B)$.
$P(S) = 1$

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur; the foundation of statistics is given by the following 3 axioms (by A.N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S. Then: (1)There are only two possible outcomes called "success" and "failure" for each trial and (2) The probability p of a success is the same for any trial (so the probability $q = 1 - p$ of a failure is the same for any trial).

Bernoulli Trials

an experiment with the following characteristics: There are a fixed number of trials, n. There are only two possible outcomes, called "success" and, "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial. The n trials are independent and are repeated using identical conditions.

Binomial Experiment

a statistical experiment that satisfies the following three conditions:

Binomial Probability Distribution

a discrete random variable (RV) that arises from Bernoulli trials; there are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial one) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Geometric Distribution

a discrete random variable (RV) that arises from the Bernoulli trials; the trials are repeated until the first success. The geometric variable X is defined as the number of trials until the first success. The mean is $\mu = \frac{1}{p}$ and the standard deviation is $\sigma = \sqrt{\frac{1}{p}\left(\frac{1}{p} - 1\right)}$. The probability of exactly x failures before the first success is given by the formula: $P(X = x) = p(1-p)^{x-1}$ where one wants to know probability for the number of trials until the first success: the xth trail is the first success. An alternative formulation of the geometric distribution asks the question: what is the probability of x failures until the first success? In this formulation the trial that resulted in the first success is not counted. The formula for this presentation of the geometric is: $P(X = x) = p(1-p)^x$. The expected value in this form of the geometric distribution is $\mu = \frac{1-p}{p}$. The easiest way to keep these two forms of the geometric distribution straight is to remember that p is the probability of success and $(1-p)$ is the probability of failure. In the formula the exponents simply count the number of successes and number of failures of the desired outcome of the experiment. Of course the sum of these two numbers must add to the number of trials in the experiment.

There are one or more Bernoulli trials with all failures except the last one, which is a success.

In theory, the number of trials could go on forever. There must be at least one trial.

The probability, p, of a success and the probability, q, of a failure do not change from trial to trial.

Geometric Experiment

a statistical experiment with the following properties:

Hypergeometric Experiment

a statistical experiment with the following properties:

1. You take samples from two groups.
2. You are concerned with a group of interest, called the first group.
3. You sample without replacement from the combined groups.
4. Each pick is not independent, since sampling is without replacement.

Normal Distribution

a continuous random variable (RV) with pdf $f(x) =$

$$\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

, where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV, Z, is called the standard normal distribution. Standard Normal Distribution a continuous random variable $(RV) X \sim N(0, 1)$; when X follows the standard normal distribution, it is often noted as $Z \sim N(0, 1)$. z-score the linear transformation of the form $z = \frac{x-\mu}{\sigma}$ or written as $z = \frac{|x-\mu|}{\sigma}$; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma)$ the result is the standard normal distribution $Z \sim N(0, 1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently. A z-score is the number of standard deviations a particular x is away from its mean value.

Binomial Distribution

a discrete random variable (RV) which arises from Bernoulli trials; there are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is:

$X \sim B(\mathbf{n}, \mathbf{p})$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- the desired confidence level,
- information that is known about the distribution (for example, known standard deviation),
- the sample and its size.

Confidence Level (CL)

the percent expression for the probability that the confidence interval contains the true population parameter; for example, if the CL = 90%, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Degrees of Freedom (df)

the number of objects in a sample that are free to vary

Error Bound for a Population Mean (EBM)

the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation.

Error Bound for a Population Proportion (EBP)

the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes.

Inferential Statistics

also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of the production is defective.

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Binomial Distribution

a discrete random variable (RV) that arises from Bernoulli trials. There are a fixed number, n, of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n, p) \mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\overline{X}$. If the size n of the sample is sufficiently large, then $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means will approximate a normal distribution regardless of the shape of the population. The expected value of the mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

The desired confidence level.

Information that is known about the distribution (for example, known standard deviation).

The sample and its size.

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

Critical Value

The t or Z value set by the researcher that measures the probability of a Type I error, σ.

Hypothesis

a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternative hypothesis (notation H_a).

Hypothesis Testing

Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

Cohen’s d

a measure of effect size based on the differences between two means. If d is between 0 and 0.2 then the effect is small. If d approaches is 0.5, then the effect is medium, and if d approaches 0.8, then it is a large effect.

a is the symbol for the Y-Intercept

Sometimes written as b_0, because when writing the theoretical linear model β_0 is used to represent a coefficient for a population.

b is the symbol for Slope

The word coefficient will be used regularly for the slope, because it is a number that will always be next to the letter “x.” It will be written as b_1 when a sample is used, and β_1 will be used with a population or when writing the theoretical linear model.

Bivariate

two variables are present in the model where one is the “cause” or

independent variable and the other is the "effect" of dependent variable.

Linear

a model that takes data and regresses it into a straight line equation.

Multivariate

a system or model where more than one independent variable is being used to predict an outcome. There can only ever be one dependent variable, but there is no limit to the number of independent variables.

R2R2 – Coefficient of Determination

This is a number between 0 and 1 that represents the percentage variation of the dependent variable that can be explained by the variation in the independent variable. Sometimes calculated by the equation $R^2 = \frac{SSR}{SST}$ where SSR is the "Sum of Squares Regression" and SST is the "Sum of Squares Total." The appropriate coefficient of determination to be reported should always be adjusted for degrees of freedom first.

Residual or "error"

the value calculated from subtracting $y_0 - \hat{y}_0 = e_0$. The absolute value of a residual measures the vertical distance between the actual value of y and the estimated value of y that appears on the best-fit line.

RR – Correlation Coefficient

A number between −1 and 1 that represents the strength and direction of the relationship between "X" and "Y." The value for "r" will equal 1 or −1 only if all the plotted points form a perfectly straight line.

Sum of Squared Errors (SSE)

the calculated value from adding up all the squared residual terms. The hope is that this value is very small when creating a model.

X – the independent variable

This will sometimes be referred to as the "predictor" variable, because these values were measured in order to determine what possible outcomes could be predicted.

Y – the dependent variable

Also, using the letter "y" represents actual values while $\hat{y}$ represents predicted or estimated values. Predicted values will come from plugging in observed "x" values into a linear model.

all populations of interest are normally distributed.

the populations have equal standard deviations.

samples (not necessarily of the same size) are randomly and independently selected from each population.

there is one independent variable and one dependent variable.

The test statistic for analysis of variance is the F-ratio.

Analysis of Variance

also referred to as ANOVA, is a method of testing whether or not the means of three or more populations are equal. The method is applicable if:

One-Way ANOVA

a method of testing whether or not the means of three or more populations are equal; the method is applicable if:

all populations of interest are normally distributed.

the populations have equal standard deviations.

samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the F-ratio.

Variance

mean of the squared deviations from the mean; the square of the standard deviation. For a set of data, a deviation can be represented as $x - \overline{x}$ where x is a value of the data and $\overline{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

Contingency Table

a table that displays sample values for two different factors that may be dependent or contingent on one another; it facilitates determining conditional probabilities.

Goodness-of-Fit

a hypothesis test that compares expected and observed values in order to look for significant differences within one non-parametric variable. The degrees of freedom used equals the (number of categories – 1).

Test for Homogeneity

a test used to draw a conclusion about whether two populations have the same distribution. The degrees of freedom used equals the (number of columns – 1).

Test of Independence

a hypothesis test that compares expected and observed values for contingency tables in order to test for independence between two variables. The degrees of freedom used equals the (number of columns – 1) multiplied by the (number of rows – 1).

Independent Groups

two samples that are selected from two populations, and the values from one population are not related in any way to the values from the other population.

Matched Pairs

two samples that are dependent. Differences between a before and after scenario are tested by testing one population mean of differences.

Pooled Variance

a weighted average of two variances that can then be used when calculating standard error.

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution, and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

Student's t-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

It is continuous and assumes any real values.

The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.

It approaches the standard normal distribution as n gets larger.

There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data items.

Test Statistic

The formula that counts the number of standard deviations on the relevant distribution that estimated parameter is away from the hypothesized value.

Type I Error

The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.

Type II Error

The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

Parameter

a numerical characteristic of a population

Point Estimate

a single number computed from a sample and used to estimate a population parameter

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and \sigma for population standard deviation

Student's t-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of this random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero.
- It approaches the standard normal distribution as n get larger.
- There is a "family" of t–distributions: each representative of the family is completely defined by the number of degrees of freedom, which depends upon the application for which the t is being used.

Average

a number that describes the central tendency of the data; there are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Central Limit Theorem

Given a random variable with known mean μ and known standard deviation, σ, we are sampling with size n, and we are interested in two new RVs: the sample mean, $\overline{X}$. If the size (n) of the sample is sufficiently large, then $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. If the size ($n$) of the sample is sufficiently large, then the distribution of the sample means will approximate a normal distributions regardless of the shape of the population. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Finite Population Correction Factor

adjusts the variance of the sampling distribution if the population is known and more than 5% of the population is being sampled.

Mean

a number that measures the central tendency; a common name for mean is "average." The term "mean" is a shortened form of "arithmetic mean." By definition, the mean for a sample (denoted by $\overline{x}$) is $\overline{x} = \overline{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Normal Distribution

a continuous random variable with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution and σ is the standard deviation.; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the random variable, Z, is called the **standard normal distribution**.

Sampling Distribution

Given simple random samples of size n from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a sampling distribution.

Standard Error of the Mean

the standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$.

Standard Error of the Proportion

the standard deviation of the sampling distribution of proportions

Conditional Probability

the likelihood that an event will occur given that another event has already occurred.

decay parameter

The decay parameter describes the rate at which probabilities decay to zero for increasing values of x. It is the value m in the probability density function $f(x) = me^{(-mx)}$ of an exponential random variable. It is also equal to $m = \frac{1}{\mu}$, where μ is the mean of the random variable.

Exponential Distribution

a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$ or $f(x) = \frac{1}{\mu}e^{-\frac{1}{\mu}x}, x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$ or $P(X \leq x) = 1 - e^{-\frac{1}{\mu}x}$.

memoryless property

For an exponential random variable X, the memoryless property is the statement that knowledge of what has occurred in the past has no effect on future probabilities. This means that the probability that X exceeds $x + t$, given that it has exceeded x, is the same as the probability that X would exceed t if we had no knowledge about it. In symbols we say that $P(X > x + t | X > x) = P(X > t)$.

Poisson distribution

If there is a known average of \mu events occurring per unit time, and these events are independent of each other, then the number of events X occurring in one unit of time has the Poisson distribution. The probability of x events occurring in one unit time is equal to $P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$.

Uniform Distribution

a continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$; it is often referred as the rectangular distribution because the graph of the pdf has the form of a rectangle. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is \(f(x)=\frac{1}{b-a} \text { for } a

Hypergeometric Probability

a discrete random variable (RV) that is characterized by:

1. A fixed number of trials.
2. The probability of success is not the same from trial to trial.

We sample from two groups of items when we are interested in only one group. X is defined as the number of successes out of the total number of items chosen.

Poisson Probability Distribution

a discrete random variable (RV) that counts the number of times a certain event will occur in a specific interval; characteristics of the variable:

- The probability that the event occurs in a given interval is the same for all intervals.
- The events occur with a known mean and independently of the time since the last event.

The distribution is defined by the mean μ of the event in the interval. The mean is $\mu = np$. The standard deviation is $\sigma = \sqrt{\mu}$. The probability of having exactly x successes in r trials is $P(x) = \frac{\mu^x e^{-\mu}}{x!}$. The Poisson distribution is often used to approximate the binomial distribution, when n is "large" and p is "small" (a general rule is that np should be greater than or equal to 25 and p should be less than or equal to 0.01).

Probability Distribution Function (PDF)

a mathematical description of a discrete random variable (RV), given either in the form of an equation (formula) or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome.

Random Variable (RV)

a characteristic of interest in a population being studied; common notation for variables are upper case Latin letters X, Y, Z,...; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x, y, and z. For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3,.... Variables in statistics differ from variables in intermediate algebra in the two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if $X =$ hair color then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x the random variable X takes only after performing the experiment.

Sample Space

the set of all possible outcomes of an experiment

Sampling with Replacement

If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once.

Sampling without Replacement

When sampling is done without replacement, each member of a population may be chosen only once.

The Complement Event

The complement of event A consists of all outcomes that are NOT in A.

The Conditional Probability of $A|B$

P(A||B) is the probability that event A will occur given that the event B has already occurred.

The Intersection: the ∩ Event

An outcome is in the event |(A \cap B\) if the outcome is in both $A \cap B$ at the same time.

The Union: the ∪ Event

An outcome is in the event $A \cup B$ if the outcome is in A or is in B or is in both A and B.

Tree Diagram

the useful visual representation of a sample space and events in the form of a “tree” with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)

Venn Diagram

the visual representation of a sample space and events in the form of circles or ovals showing their intersections

Survey

a study in which data is collected as reported by individuals.

Systematic Sampling

a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let k = (number of individuals in the population)/(number of individuals needed in the sample). Choose every kth individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

Treatments

different values or components of the explanatory variable applied in an experiment

Variable

a characteristic of interest for each person or object in a population

Frequency

the number of times a value of the data occurs

Frequency Table

a data representation in which grouped data is displayed along with the corresponding frequencies

Histogram

a graphical representation in *x*-*y* form of the distribution of data in a data set; *x* represents the data and *y* represents the frequency, or relative frequency. The graph consists of contiguous rectangles.

Interquartile Range

or *IQR*, is the range of the middle 50 percent of the data values; the *IQR* is found by subtracting the first quartile from the third quartile.

Mean (arithmetic)

a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by $\overline{x}$) is $\overline{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$

Mean (geometric)

a measure of central tendency that provides a measure of average geometric growth over multiple time periods.

Median

a number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Midpoint

the mean of an interval in a frequency table

Mode

the value that appears most frequently in a set of data

Outlier

an observation that does not fit the rest of the data

Percentile

a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Quartiles

the numbers that separate the data into quarters; quartiles may or may not be part of the data. The second quartile is the median of the data.

Relative Frequency

the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: *s* for sample standard deviation and σ for population standard deviation.

Variance

mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x - \overline{x}$ where *x* is a value of the data and $\overline{x}$ is the sample mean. The sample variance is equal to

the sum of the squares of the deviations divided by the difference of the sample size and one.

end