MEASURING THE CENTER

Averages, Quartiles, and the Five Number Summary

Descriptive statistics: A way to quickly summarize data within a set using just a few numbers.

Mean: The average of a set calculated by adding all the values in the set and dividing by the number of values in the set.

Outlier: A value or values significantly higher or lower than the rest of the set that can skew the mean of a set.

The mean is very sensitive to outliers, while the median is not.

Median: The middle value in a data set.

Mode: The value that appears most often in the set.

When a set has two modes it is called **bimodal**. When it has more than two modes, it is **multimodal**.

Standard deviation: A measurement of the amount of variation from the mean in a data set.

For example, if a data set has a mean of 50 units and a standard deviation of 20 units, we can conclude that most of the data will fall between 30 and 70 units.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

σ = standard deviation

x_i = the value of the i^{th} observation

N = Number of data points

μ = mean of data values

Five number summary: The minimum, first quartile, median, third quartile, and maximum of a data set.

Each **quartile** represents 25% of the data within a set.

The first and third quartiles can be found by identifying the medians of the lower and upper halves of the data.

Range: The distance between the maximum and minimum.

Interquartile range (IQR): The distance between the third and first quartiles.

Graphical Organization

Boxplot: A graph representing the five number summary.

The boxed area represents the IQR with the median at the center.

Frequency distribution: A table that sorts data into equally-sized classes.

Ages of Mobile Phone Customers

Class	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
$20 \le X < 30$	17	17	34.00%	34.00%
$30 \le X < 40$	16	33	32.00%	66.00%
$40 \le X < 50$	12	45	24.00%	90.00%
$50 \le X < 60$	4	49	8.00%	98.00%
$60 \le X < 70$	1	50	2.00%	100.00%
Total	50		100.00%	

Frequency: The amount of data points that fall into each class.

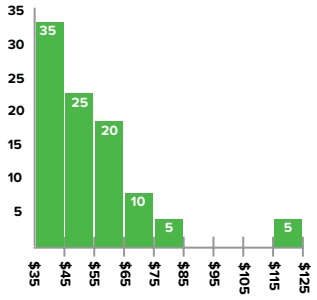
Cumulative frequency: The running total of the frequencies.

Relative frequency: The frequency divided by the total number of data points.

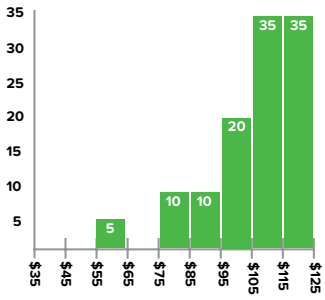
Cumulative relative frequency: The running total of the relative frequencies.

Histogram: A frequency distribution shown in graph form.

Positive skew (right skew): When values pull a chart to the right.



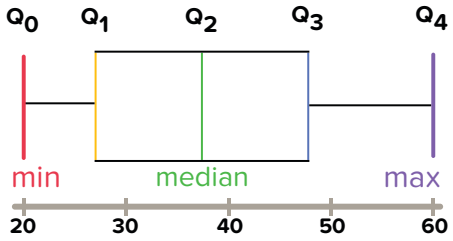
Negative skew (left skew): When values pull a chart to the left.



In a histogram with a positive skew, the mean is greater than the median.

In a histogram with a negative skew, the median is greater than the mean.

Boxplot



If Q1 and the minimum are the same value, you won't see a tail on the left side.

If Q3 and the maximum are the same value, there will be no tail on the right side.

SYMBOLS, RULES, AND LAWS

Introduction to Probability

Experiment: A situation involving chance, like rolling a six-sided die.

An **outcome** is the possible result of an experiment, like rolling a 3.

The **sample space** is all possible outcomes for a particular experiment, written as a **set**—an unordered collection of distinct **elements** that do not repeat.

Probability: The likelihood of something happening. The formula we can use to calculate probability is:

$$\frac{\text{\# of desired outcomes or events}}{\text{\# of elements in the same space}}$$

Note: This equation only works if all of the elements in the sample space are **equally likely outcomes**.

Event: A particular collection of possible outcomes within a sample space, also written as a set.

The probability of rolling an even number is the ratio of elements in event E to elements in sample space S:

$$P(E) = \{2, 4, 6\}/\{1, 2, 3, 4, 5, 6\} = 3/6 = 0.5$$

Experiment: rolling a 6-sided die

Sample space	$D = \{1, 2, 3, 4, 5, 6\}$
Probability of rolling a 4	$P(4) = \frac{\{4\}}{\{1, 2, 3, 4, 5, 6\}} = \frac{1}{6}$

Rolling an even number could be considered an event, with: $E = \{2, 4, 6\}$.

Theoretical and Experimental Probabilities

Theoretical probabilities are those we can calculate through knowledge of a situation alone (i.e., the outcome of a coin flip).

Experimental probabilities are those we derive from repeated trials of real-world experiments. They are calculated by:

$$\frac{\text{\# of times desired outcome occurred}}{\text{total \# of trials}}$$

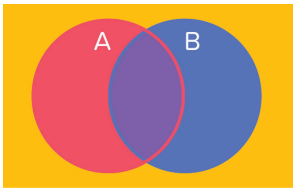
Subjective probabilities are those based on intuition or gut feel.

Theoretical probabilities are the most reliable, followed by experimental probabilities. Subjective probabilities are the least reliable.

Many real-world situations don't have associated theoretical probabilities. So, experimental probabilities are often used in business to make predictions and optimize performance.

Venn Diagrams

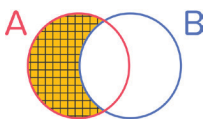
A **Venn diagram** shows logical relationships between events so we can better understand the probabilities beneath them.



- The box in a Venn Diagram represents an entire sample space.
- Circles within the box represent different events.
- Everything outside of A's circle is A^C, event A's **complement.**
- The overlap of two circles is the **intersection** of these events, written $A \cap B$ (*A intersect B*).
- Everything contained in multiple circles is the **union** of the events, written $A \cup B$ (*A union B*).

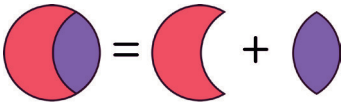
These symbols can be used to represent complicated expressions:

$A \cap B^C$



Or simple ones:

$A = (A \cap B) \cup (A \cap B^C)$



Basic Probability Rules

1. There are no negative probabilities.
2. The probability of something occurring is 100%.
3. The total probability of an event either happening or not happening is 100%.

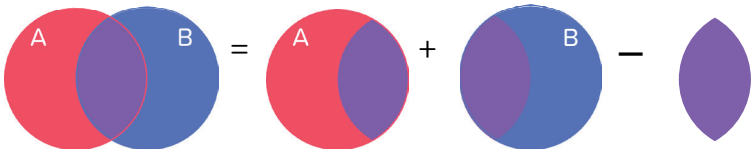
The probability of summer snow in San Diego can't be negative. The probability that it will either snow or not snow is 100%.

The Addition Rule

The addition rule helps us find the probability of the union of two events—the chance of one event or the other (or both) happening.

Addition rule for disjoint events: $P(A \cup B) = P(A) + P(B)$

Addition rule (general version): $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Disjoint / Mutually exclusive events: Events that cannot happen at the same time.

An event and its complement are disjoint events.

The Multiplication Rule for Independent Events

Independent events do not influence each other.

Independent and disjoint events are not the same thing. In fact, two disjoint events cannot be independent.

Multiplication rule for independent events: If two events are independent, the probability of both of them occurring is:

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

Events that are independent:
A: heads on your first coin flip
B: heads on your second flip

Events that aren't independent:
A: heads on your first coin flip
B: tails on your first coin flip
(these are disjoint)

Conditional Probabilities and the Multiplication Rule

Marginal probability: The probability of a single event, P(A).

Conditional probability: The probability of A happening given B has happened, written P(A|B) (*A given B*).

Joint probability: The probability of both A and B happening, written $P(A \cap B)$.

Multiplication rule (general version): $P(A \cap B) = P(A|B) \times P(B)$

Three ways to check if A and B are independent events:

$$P(A|B) = P(A)$$
$$P(A|B) = P(A|B^C)$$
$$P(A \cap B) = P(A) \times P(B)$$

Bayes' Rule and the Law of Total Probability

When we get new information, we can update our **prior probability** to get a more accurate **posterior probability**.

To do this, multiply the prior probability by a **likelihood ratio**.

Bayes' rule: Formula for calculating conditional probabilities.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Law of Total Probability (LTP): Formula for finding marginal probabilities derived by splitting a sample space into partitions and adding the probability of each partition.

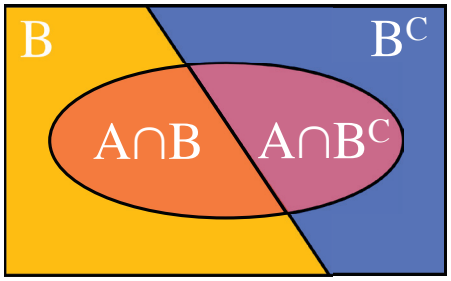
A sample space can be split into more than one partition, with the LTP formulas expanding accordingly.

Combining Bayes' rule and the LTP gives a new formula, which is useful in testing scenarios where we know the false positive rate of a test.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

Remember: There are two versions of the LTP formula.

$$P(A) = P(A \cap B) + P(A \cap B^C)$$
$$P(A) = P(A|B)P(B) + P(A|B^C)P(B^C)$$



Tip: When choosing a probability rule to apply, figure out what you already know and what you want to know. Then, find the formula that can bridge that gap.

PROBABILITY DISTRIBUTIONS: THE BASICS

Random Variables

Random variable: A variable whose value is determined by the outcome of an experiment that involves chance. Random variables are denoted with capital letters.

Discrete random variable: If the number of possible values of a random variable is countable on a given interval, it is discrete.

Continuous random variable: If the number of possible values of a random variable is uncountable on a given interval, it is continuous.

Infinite random variable: If the possible values of a random variable do not exist on a closed range, it is infinite.

Finite random variable: If the possible values of a random variable can be counted within a closed range, it is finite.

Discrete random variables can be either finite (as above) or infinite.

Continuous random variables are always infinite, since their outcomes are never countable (even on a closed range).

Probability Distributions

Probability distribution: A set of probabilities representing how likely the possible outcomes of a random variable are to happen. They can be represented as graphs, equations, or tables.

$P(X = x)$**:** The probability that the random variable, denoted with the capital letter, will have a specific value, denoted with the lowercase letter.

Discrete probability distribution: A countable set of probabilities associated with the possible outcomes of a discrete random variable. In graph form, these are represented as bar graphs.

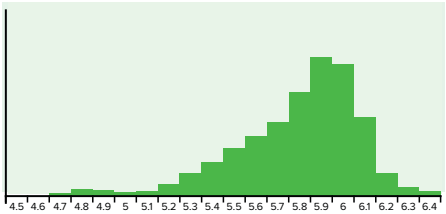
Continuous probability distribution: An uncountable set of probabilities associated with the possible outcomes of a continuous random variable. In graph form, these are represented with smooth, continuous curves.

The sum of the probabilities in a discrete probability distribution add up to 1, and the area under the curve of a continuous probability distribution is 1.

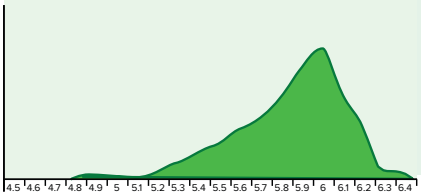
Weighted average: Multiply each outcome by its probability and add them together. The mean of a discrete distribution is the weighted average of all the possible outcomes of the discrete random variable multiplied by their probabilities.

X = chin-ups a client does after a Hup-hup Protein Shake

x	6	7	8	9
P (X = x)	0.1	0.2	0.4	0.3



Discrete Probability Distribution



Continuous Probability Distribution

Shapes

Unimodal: A probability distribution with only one peak, or "mode."

Bimodal: A probability distribution with two peaks, or "modes."

Multimodal: A probability distribution with two or more peaks, or "modes."

Skewed right: A probability distribution with the bulk of the data on the left of the graph, and a tail on the right.

Skewed left: A probability distribution with the bulk of the data on the right of the graph, and a tail on the left.

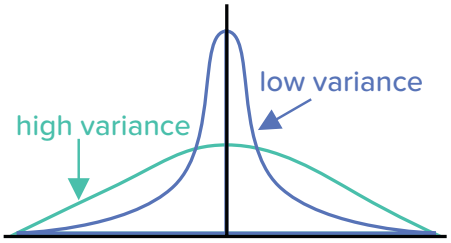
Uniform distribution: A probability distribution in which the probability of any outcome happening is equal.

Expected Value and its Associates

Expected value, $E(X)$: The mean, calculated via weighted average, of all the possible outcomes of a probability distribution. It helps us know what to expect in the long run.

$E(X)$ is a weighted average calculation, meaning it's found by multiplying each outcome by its probability and taking the sum.

$E(cX) = cE(X)$ and $E(X+Y) = E(X) + E(Y)$, thanks to the linearity property.



Variance: A measure of how spread out a distribution is from its mean. A distribution with high variance is more spread out, while one with low variance is narrower. Standard deviation is the square root of the variance.

Skewness: A measure of the symmetry of a distribution. A good rule of thumb is that a skewness above 1 or below -1 is highly asymmetrical.

Variance is always positive, while skewness can be positive or negative to indicate the direction of the skew.

BINOMIAL, POISSON, AND NORMAL DISTRIBUTIONS

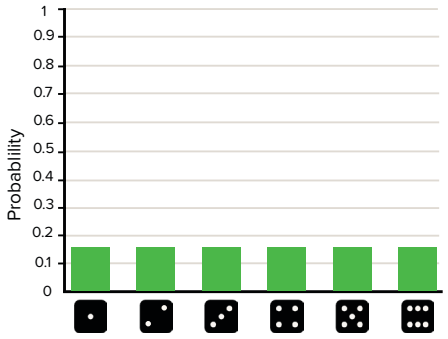
Probability Distribution Families

Probability distribution family: A way to model a group of situations that show similar predictable patterns. Every probability distribution family has a set of conditions that a situation must meet in order to be modeled with that distribution.

Outcome of interest: The particular outcome in a probability distribution whose probability you're trying to find, often denoted as k.

Parameter: A value that relates a probability distribution family to a specific situation. It tells us how to tweak the distribution to properly model this situation.

Probability distribution function: A formula that uses the values of a specific situation's parameters to tell us the probability of an outcome of interest.



Predicting outcomes of a dice roll with the discrete uniform distribution family. $P(X = x) = 1/k$

Binomial Distributions

Binomial distribution: The spread of successes in independent yes/no experiments that have a constant probability.

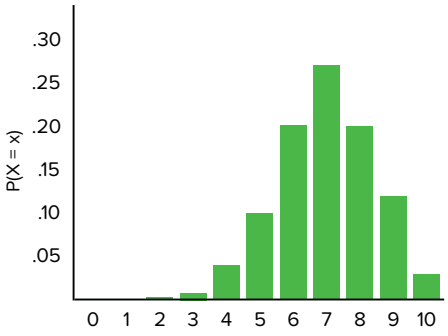
Trial: An independent experiment with only two possible outcomes. The number of total trials in a binomial distribution is written as n.

Success: The desired outcome of a trial. The number of successes you're interested in is often written as k.

P**:** the probablity of success in each trial.

Binomial random variable: A random variable with n independent trials with two outcomes (success/failure) where the probability of success for each trial is constant.

n **x** p**:** The mean of a binomial distribution is the total number of trials multiplied by the probability of success on each trial.



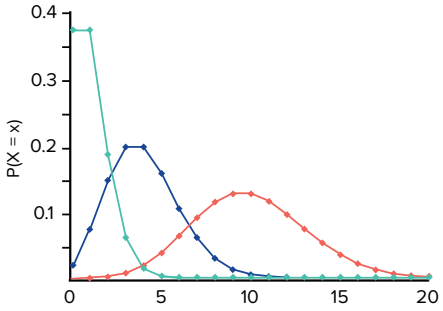
Binomial distribution with $n = 10$ trials and a $p = 70\%$ chance of success on each trial.

Poisson Distributions

Poisson distribution: A probability distribution family used to model the probability of events in a fixed interval that occur at a known average rate.

Lambda, λ : The Poisson distribution's only parameter—the average rate of events per fixed interval. Lambda is also the mean of a Poisson distribution.

The Poisson distribution has four conditions: your random variable is discrete and infinite, there is a known average success rate in the fixed interval, the successes are independent of each other, and it's not meaningful to talk about failure rate.



Poisson distributions with λ=1, λ=4, and λ=10.

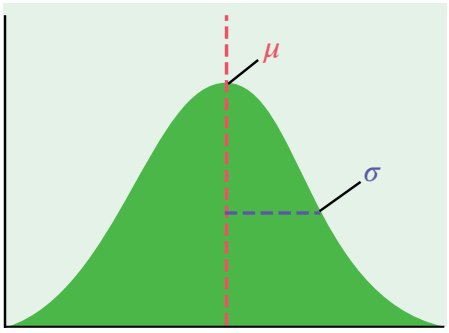
Introduction to Normal Distributions

Normal distribution: A commonly used continuous probability distribution known for its bell-shaped, unimodal, symmetric curve. Often used to model natural population phenomena like heights, lengths, and weights.

The **mean**, μ, and the **standard deviation**, σ, are the two parameters that define a normal curve.

Cumulative probability: The probability of a range of values, used when dealing with continuous distributions.

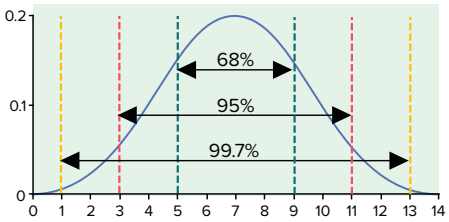
Central Limit Theorem (CLT): The distribution of the means of many samples from a population will be roughly normally distributed, even if the underlying population distribution is not normal.



The Empirical Rule

Empirical rule: If a data set is normally distributed, 68% of the data will fall within one standard deviation from the mean. Also called the **68-95-99.7** rule, since 95% of the data falls within two standard deviations of the mean, and 99.7% of the data falls within three.

You can use the Empirical rule in reverse to determine if a trend is normally distributed by measuring how much of the data is within one standard deviation of the mean.



Z-Scores and the Standard Normal Distribution

Z-score: In a normal distribution, this score tells us how many standard deviations a value is from the mean.

Z-score formula: To calculate the Z-score for a value of X, subtract the mean from X and then divide by the standard deviation.

$$Z = \frac{(X - \mu)}{\sigma}$$

Standard normal distribution: A normal distribution with a mean of 0 and a standard deviation of 1. Z-scores let us convert any normal distribution to the standard one.

Z-table: Allows us to find the cumulative probability associated with a Z-score. The first column shows the first two digits of a Z-score, and the top row shows its second decimal.

Cumulative Z-table: Gives the cumulative probability for values less than Z.

Cumulative from mean Z-table: Gives the cumulative probability for values between 0 and Z.

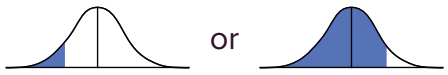
Complementary cumulative Z-table: Gives the cumulative probability for values greater than Z.

Cumulative Z-tables are the only ones that have values for both positive and negative Z-scores. Sometimes, you will have to be creative with which Z-score and Z-table you use to find a specific value.

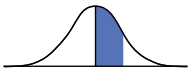
z	0	1	2	3	4	···
⋮						
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	
⋮						

Finding the cumulative probability of values greater than $Z = 1.50$.

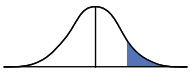
Cumulative



Cumulative from mean



Complementary cumulative



TWO-VARIABLE STATISTICS

Correlation and Causation

Correlation: A relationship that exists when one event can be used to predict a second event; if A correlates to B, B correlates to A.

Positive correlation: As A increases, so does B.

Negative correlation: As A increases, B decreases and vice versa.

Causation: Occurs when one of two correlated events functions as a cause of the other; if A causes B, B cannot cause A.

Correlation does not imply causation, but causation implies correlation.

Experiment: A study in which researchers directly influence the subjects. Experiments can determine causation.

Control group: In an experiment, the test subjects that do not receive the treatment under investigation.

Treatment group: The test subjects in an experiment who do receive the treatment under investigation.

In a **blind experiment**, fake treatments are given to the control group and experimental treatments to the treatment group. All test subjects are unaware of which treatment they receive.

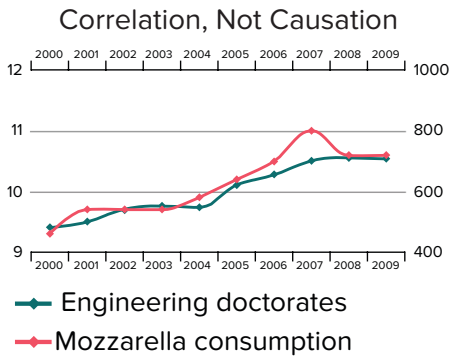
When neither the subjects nor the researchers know what group a subject belongs to, this is called a **double-blind experiment**.

Confounding variables: Outside factors that may alter an outcome.

Independent variable: The variable that is manipulated in the experiment. In a perfect experiment, the independent variable is the only difference between the treatment group and the control group.

Dependent variable: The variable being measured in an experiment—specifically, how it is affected or not affected by a change in the independent variable.

The **placebo effect** is a confounding variable that occurs when the thought of being tested or treated causes a reaction from the subject.

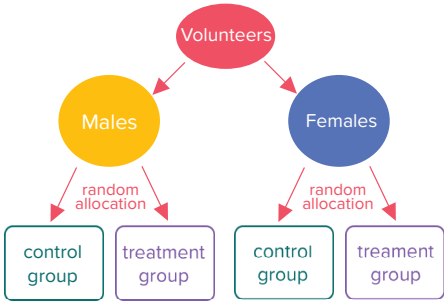


Observational study: A study in which researchers collect data through observation. These studies only demonstrate correlation and can never be used to prove causation.

Retrospective study: A study in which a researcher looks backward at an outcome.

Prospective study: A study in which a researcher looks forward to an outcome that has yet to occur.

Blocking: Dividing subjects into blocks according to a confounding variable and randomly assigning members of each block into the treatment and control groups.



Scatter Plots

Scatter plot: A graph that describes the relationship between two sets of quantitative data.

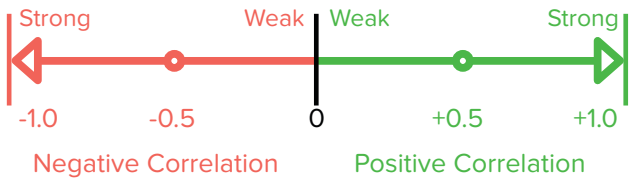
A **linear relationship** exists when the points on a scatter plot form a line.

A **direct relationship** or **positive correlation:** as x increases, so does y.

An **inverse relationship** or **negative correlation:** as the x increases, y decreases.

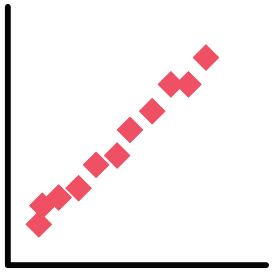
Zero correlation: There is no correlation between the data sets—no relationship exists between the x and y variables.

The **correlation coefficient (r)** indicates both the strength and the direction of the relationship between the variables of a scatter plot; r will always be a number between -1 and +1.

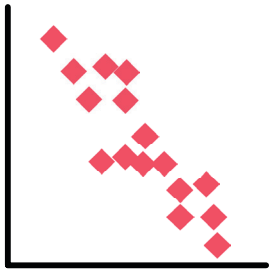


A negative value of r indicates an inverse relationship; a positive value of r indicates a direct relationship. Values of r close to zero mean little to no correlation exists.

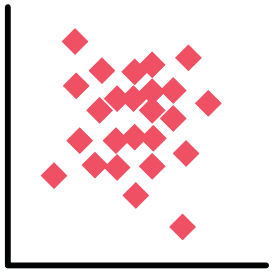
Covariance: A measure of how two quantities change together.



strong positive correlation



moderate negative correlation



no correlation

APPLICATIONS OF REGRESSION ANALYSIS

Variable Selection

Backward elimination: A variable selection method that starts with all of the variables in consideration, then removes them one by one.

Forward selection: A variable selection method that starts with no variables and adds them one by one.

Adjusted R^2: A value similar to R^2 that assesses a multivariate model's accuracy, but punishes a model for having too many variables relative to their benefit.

p-value: The probability that a variable is not significant in a model. A good variable usually has a p-value less than or equal to 0.05.

Common mistakes to avoid:

Overfitting: A model has so many input variables that it describes randomness rather than real relationships.

Collinearity: Two of the independent variables in a model are correlated, which muddles a model's accuracy.

Lurking variable: A variable that isn't included in a model, but affects both independent and dependent variables.

$$y = 2.8 + 0.26x_1 + 0.3x_2 - 0.043x_3 + 0.59x_4 - 0.024x_5 + 0.007x_6$$

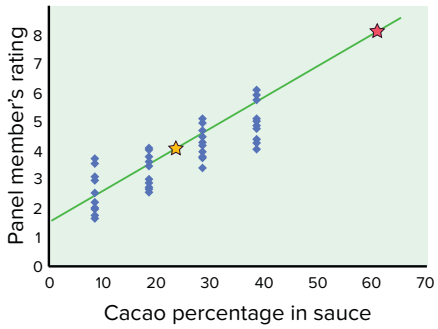
Variable	p-value
x_1 = number of years open	0.32
x_2 = average price of an entrée	0.45
x_3 = noise level (in decibels)	0.04
x_4 = lunch service (Y/N)	0.09
x_5 = total rating on *Revue*	0.07
x_6 = restaurant size (m^2)	0.50

Adj. R^2 = 0.696

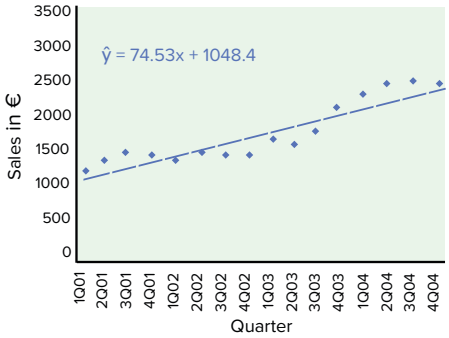
Interpolation and Extrapolation

Interpolation: A prediction made using an input that lies within the set of observed values. These are usually reliable predictions.

Extrapolation: A prediction made using an input that lies outside the set of observed values. These predictions are less reliable and should be used with caution.



Time Series Analysis



Time series data: A set of observations taken at evenly spaced, consecutive time intervals.

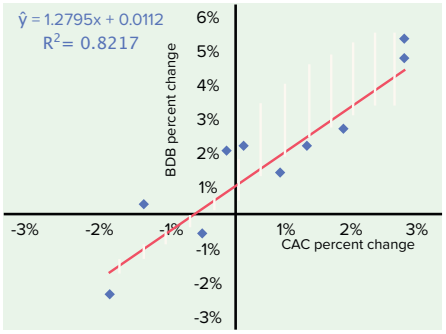
Autocorrelation: Errors show a sequential pattern since future outcomes tend to rely on the outcomes just before them.

Time series analysis: A specific type of regression analysis that takes typical time-based trends into account to build a model for time series data.

- **Secular trend:** A high-level, overarching trend in a time series data set. Typically, it's as simple as increasing, decreasing, or staying the same.
- **Cyclic trend:** The movement of time series data as it responds to the broad and unpredictable economic cycle of recession and growth.
- **Seasonal trend:** The predictable, repetitive rise and fall of timebased outcomes.

Autoregressive model: A time series analysis method in which the independent variable is a previous outcome, rather than time.

Linear Regression and Finance



Beta: The percent change in a stock for every 1% change in the broader market or any other index used for comparison.

Volatility: How much a stock jumps up and down compared to the broader market. A high beta means high volatility. Generally speaking:

$$\beta < 1: \text{low volatility}$$

$$\beta > 1: \text{high volatility}$$

R^2 describes how strong the relationship between a stock and the broader market is, no matter the volatility of that relationship.

REGRESSION THEORY

Introduction to Regression Analysis

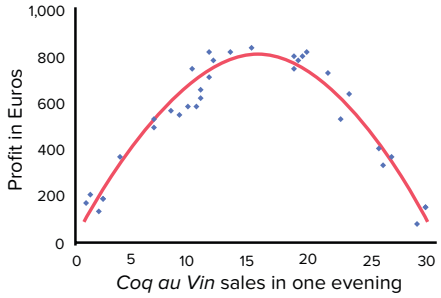
Regression analysis: A statistical method for estimating an idealized relationship between one or more independent variables (predictors) and a dependent variable (outcome).

Model: A type of graph or relationship between two variables. Examples include linear, quadratic, exponential, and polynomial.

Predicted values: The values that a regression model predicts for each input.

Observed values: The actual observed data points that are used to create a regression model.

Error / Residual: The difference between observed and predicted values.



Linear Regression

Simple linear regression: Creating an idealized linear relationship between a predictor (or predictors) and an outcome.

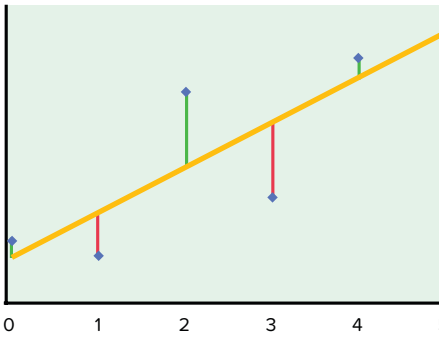
Least squares linear regression ensures that a regression line is as close to all the data points as possible by minimizing the sum of the squared values of all residuals.

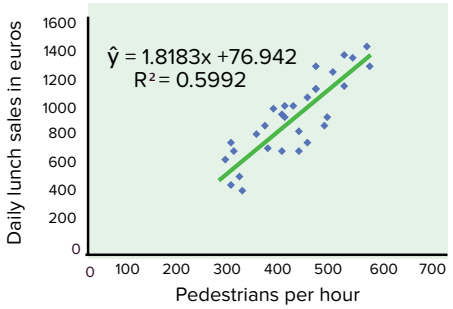
Slope: The steepness of a linear model. Think rise over run.

Y-intercept: The point where a linear model crosses the y-axis.

Trend line: The linear model that's closest to all data points, represented with the equation $\hat{y} = \alpha + \beta x$, where α is the y-intercept, β is the slope, x is the input value, and $\hat{y}$ is the predicted outcome.

R^2 / Coefficient of determination: A value that tells us what percent of change in the output is due to changes in the input, i.e., how accurate the model is.



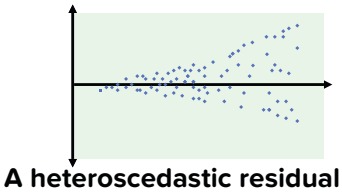


Residual and Outlier Analysis

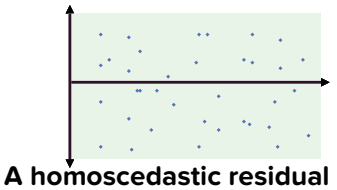
Residual analysis: Investigating error values and outliers in a model to assess the model's accuracy.

Residual plot: A graph with residual values on the y-axis and the independent variable or the model's predicted values on the x-axis. A good model should have a randomly scattered residual plot.

A **heteroscedastic** residual plot has inconsistent variance in error values, while a **homoscedastic** residual plot has errors with the same average size overall.



A heteroscedastic residual plot



A homoscedastic residual plot

Multiple Regression and Data Transformation

Multivariate model: A model with more than one input variable: $\hat{y} = \alpha + \beta x_1 + \beta x_2 + \beta x_3 \ldots$ and so on. We rely on residual analysis and adjusted R^2 values to assess their accuracy, since we can't visualize them with a graph.

Data transformation: Performing a mathematical operation on every data point in a data set that linearizes the relationship between this input and the output of a multivariate linear model.

Logarithmic growth: Growth starts fast, but slows and flattens out over time. Taking the log of each input value will linearize a logarithmic growth relationship.

Other kinds of data transformations:

- Take the inverse of each observation.
- Take the square root of each observation.
- Square each observation.

Dummy Variables and Logistic Regression

Quantitative variables deal with measurable, ordered numerical values, while **categorical variables** deal with unordered classifications like names or labels.

Dummy variable: A categorical variable that's represented with a **binary system** of 0s and 1s so that it can be included in a regression model. If a categorical variable has k classifications, k-1 dummy variables can represent it.

Dummy coding: The assigning of 0s and 1s to a categorical variable's categories. All but one category receives a variable that is equal to 1 if it's true and 0 if otherwise.

Logistic regression: A regression method used when the output of a model is categorical. A logistic model's accuracy can be assessed with pseudo R^2.

$$\hat{y} = 18.4 - 13.0x_1 + 16.0x_2 + 10.1x_3$$

Dip vs. No dip	Flavor
choco-dip: $x_1 = 1$	vanilla: $x_2 = 1, x_3 = 0$
no dip: $x_1 = 0$	chocolate: $x_2 = 0, x_3 = 1$
	pistachio: $x_2 = 0, x_3 = 0$

The one category in a categorical variable whose dummy coding is all 0s is called a **baseline** or **reference category**.

STATISTICAL SAMPLING

Sampling Basics

Population: A collective group being surveyed.

Census: A collection of data from every member in a population.

Sample: A subgroup taken from the population.

Sampling frame: A list or database of all members of a population.

Random sampling: Every element in a population has an equal chance of being selected for a sample.

Biased sampling: Every element in a population does not have an equal chance of being selected for a sample.

Simple and Stratified Random Sampling

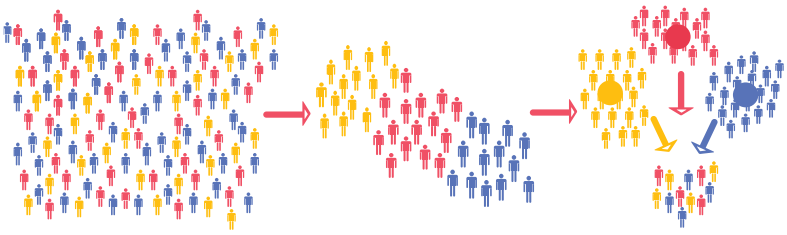
Simple random sampling:

1. Create a *sampling frame.*
2. Randomly select elments from the frame to be in the sample.

When sampling a large group, simple random sampling is essential.

Stratified random sampling:

1. Divide the population into **strata** (organized groups) with similar attributes.
2. Then apply simple random sampling to each stratum.



Stratified random sampling ensures that contrasting segments of a population are equally represented.

Cluster and Systematic Sampling

Cluster random sampling:

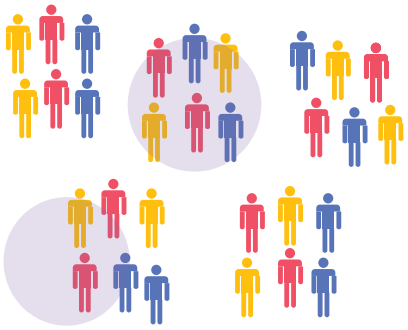
1. Divide a population into clusters: groups that are representative of the whole population.
2. Use random sampling to select a sample of clusters.
3. Survey each element within those clusters.

Systematic random sampling:

1. Divide the size of the frame (N) by the desired sample size (n) to get the index number (k).
2. Starting with a randomly selected number, choose every k^{th} element in the frame.

For example: If N = 500 and n = 50, we would choose every 10^{th} item from the frame.

Cluster random sampling saves money.



The patterns in a population may skew the sampling results in systematic random sampling.

Avoiding Biased Sampling Techniques

Convenience sampling: A sampling of the most readily available elements of the population.

Quota sampling: A population is divided into groups and then uses convenience sampling to fill a quota of required samples.

Snowball sampling: When an interviewer asks if a subject can provide friends to be interviewed.

Conducting a Survey

Survey: A method of data collection where individuals are asked questions in order to measure ideas, opinions, and experiences.

Voluntary response bias: People with strong opinions are more likely to respond to a poll.

Nonresponse bias: When a survey receives little to no response.

Response bias: When people respond with a knowingly false answer.

CONFIDENCE INTERVALS

Working with Confidence Intervals

Parameters: Numerical facts that describe an entire population.

Statistics: Values gleaned from a random sample of a population.

Inferential statistics: The process of using data from a sample to draw conclusions about an entire population.

Confidence interval: A range of values computed from a sample statistic that is highly likely to contain the population parameter. A confidence interval is centered around a point estimate (a single sample statistic), and the width is called the margin of error.

Confidence level: A percentage that specifies how confident one is that a confidence interval contains the parameter of interest. A confidence level of 95% says that, for every 100 samples taken, 95 of the resulting confidence intervals will likely contain the parameter.

In a **confidence interval for means**, the point estimate is the sample mean, x, which is used to estimate the population mean, μ.

A full description of a confidence interval includes the confidence level, a description of the population parameter in question, and the interval's lower and upper limits (with units).

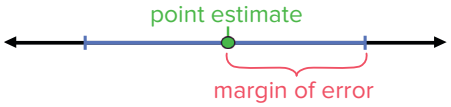
Sampling distribution for the sample mean: A hypothetical plot of sample means vs. their frequency in an infinite number of repeat samples of the same size. This plot is centered around the true population mean.

We calculate the margin of error using the **critical value:** a Z-score that corresponds to our confidence level and maps to the sampling distribution. For 95% confidence, Z = 1.96.

In a **confidence interval for proportions**, the point estimate is p, the sample proportion, which is used to estimate the population proportion, p.

A population proportion is the ratio of the people or things in a population with a certain characteristic (often called a **success**) to the total size of the population.

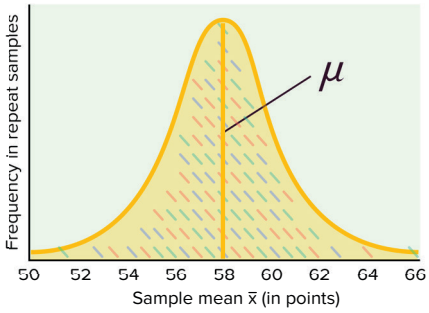
Sampling distribution of the sample proportion: A plot that represents the p values that would be calculated from an infinite number of repeat samples of the same size. This plot is centered around the true population proportion.



$$\left(\bar{x} - Z\frac{\sigma}{\sqrt{n}}, \bar{x} + Z\frac{\sigma}{\sqrt{n}}\right)$$

$\bar{x}$ = sample population mean
Z = critical value
σ = population standard deviation
n = sample size

Often in a means problem, the population standard deviation, σ, isn't known. If the sample is sufficiently large and random, we can substitute s, the sample standard deviation, for sigma in our confidence interval for means formula.



$$\left(\hat{p} \pm Z\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right)$$

$\hat{p}$ = point estimate
Z = critical value
n = sample size

To build a confidence interval for means, the sample size must be greater than or equal to 30, and for confidence intervals for proportions, you must have at least 15 successes and 15 failures.

HYPOTHESIS TESTING

Setting Up The Hypothesis

Hypothesis testing: A branch of inferential statistics that allows us to use samples to draw conclusions about a population based on a set of statistical hypotheses.

Statistical hypothesis: A testable assumption about a population parameter. The first step in hypothesis testing is to set the following two hypotheses:

Null hypothesis: A safe or commonly held assumption, written H_0. If there is sufficient evidence that H_0 is false, the researcher can **reject H_0**. If there is insufficient evidence that H_0 is false, the researcher will **fail to reject H_0**.

Alternative hypothesis: A proposal that's outside of what is safe or expected, written H_a. The alternative hypothesis states that the population parameter is somehow different from a given value.

The null hypothesis typically contains an equals sign.

H_0: $\mu = ?$ **or** $p = ?$

The alternative hypothesis typically contains $\neq$, $>$ or $<$.

H_a: $\mu \neq ?$, $\mu < ?$, or $\mu > ?$

H_a: $p \neq ?$, $p < ?$, or $p > ?$

Setting The Criteria For A Decision

After setting the hypotheses, a researcher specifies α (alpha), the **significance level**—the probability that a sample's average leads one to *reject* the null hypothesis when, in reality, it's *true*.

Assumption of Normality: In a hypothesis test, we assume that the sampling distribution is normal and centered around the true population parameter. This is usually a sound assumption with a large and sufficiently random sample.

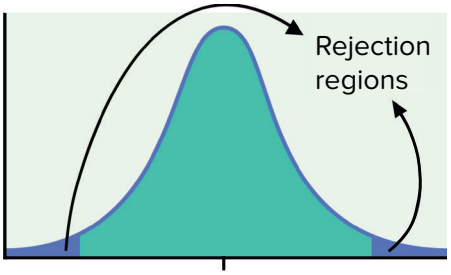
Rejection region: The tail(s) of a sampling distribution, the area under which corresponds with the significance level α. If a result falls in a rejection region, we have sufficient evidence to reject H_0.

The boundary (or boundaries) of the rejection region(s) on the standard normal distribution is known as the **critical value**. When the Assumption of Normality is met, this can be found on a Z-table.

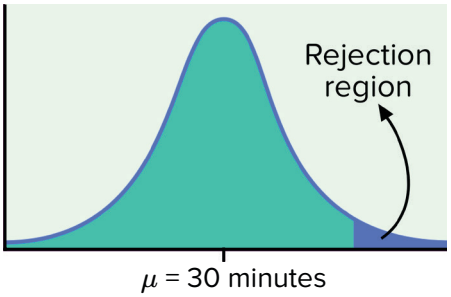
Two-tailed test: A test with rejection regions on both sides of the sampling distribution, typically characterized by a ≠ sign in the alternative hypothesis.

One-tailed test: A test with a rejection region on only one side of the sampling distribution, typically characterized by a < or > sign in the alternative hypothesis.

Two-tailed



One-tailed



Test Statistics and p-values

A **test statistic** converts a sample statistic into a location on a well-known distribution to help researchers interpret a sample's results.

z-statistic: Used when the sampling distribution can be modeled with the standard normal distribution (which is possible when the Assumption of Normality is met).

z-statistic for means

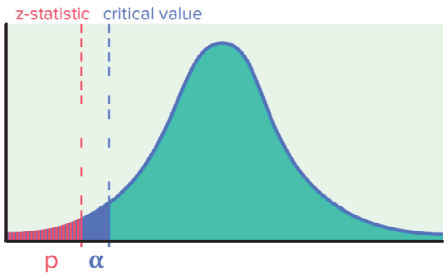
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$\bar{x}$ = sample mean
μ = population mean from H_0
σ = population standard deviation
n = sample size

z-statistic for proportions

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}}$$

$\hat{p}$ = sample proportion
p_0 = population proportion from H_0
n = sample size



If the test statistic falls in the rejection region (meaning it's more extreme in either direction than the critical value), the researcher can reject H_0.

p-value: Gives the exact probability of achieving a result that's just as extreme or more extreme than the result if H_0 is true. p-values help quantify the strength of a rejection of the null hypothesis.

Interpreting and Reporting Results

When reporting hypothesis test results, a researcher should include the experimental design, hypotheses, sample size, significance level, p-value, and any other relevant information.

Small effect size is when a result is statistically significant but too small to be of any practical significance. It can influence the context of a reported result.

There are three common mistakes that must be avoided when reporting test results:

Cherry picking: A researcher uses only data that support their desired conclusion.

p-hacking: A researcher manipulates data collection or analysis until non-significant results become significant.

Significance chasing: A researcher reports insignificant results as if they're "almost" significant.

How do you choose between performing a hypothesis test or building a confidence interval?

Hypothesis tests are good for comparing values and give information about the strength of a rejection of H_0 in the form of a p-value.

Confidence intervals give a range of plausible values for μ. They tell you if you can reject H_0, but not how strong that rejection is.

MUCH ADO ABOUT ERRORS

Test Results vs. States of Nature

State of nature: The underlying truth in a hypothesis test, regardless of whether H_0 is rejected.

There are only two states of nature:

1. H_0 is true
2. H_0 is false

When the result of your hypothesis test matches the state of nature, the test result has led you to a **correct inference**.

There are two ways to reach a correct inference:

1. You reject a null hypothesis that's false.
2. You fail to reject a null hypothesis that's true.

Every hypothesis test is vulnerable to two errors:

- Rejecting H_0 when it's true.
- Failing to reject H_0 when it's false.

In both cases, your test results do not match the state of nature.

Balancing Type I and II Errors

Type I error: When H_0 is true in reality but a test says to reject it.

Type II error: When H_0 is false but your test results lead you to fail to reject it.

alpha (α): The probability of making a type I error (rejecting H_0 when it's true). This is the same as the "significance level."

beta (β): The probability of making a type II error (failing to reject H_0 when it's false).

There are two ways to lower β:

1. Increase α, which increases the chance of a type I error.
2. Increase n, the sample size, thereby decreasing the type II risk while maintaining the type I risk.

Power: The probability of rejecting H_0 when it's false.

Power is the complement of beta ($P = 1 - \beta$), meaning that as beta increases, power decreases.

The **table of error types** pairs states of nature with test results and shows whether each possibility is a correct inference or an error.

		State of nature	
		H_0 is true	H_0 is false
Hypothesis test result	FTR H_0	Correct inference	Type II Error
	Reject H_0	Type I Error	Correct inference

DEALING WITH SMALL SAMPLES

t-Distributions

t-distribution: A set of probability distributions similar to normal distributions, but tweaked to adjust for an unknown σ value.

Degrees of freedom (df): A metric that serves as a proxy for sample size (and occasionally other parameters specific to given samples) and determines the exact shape of a t-distribution.

In a t-test, $df = (n-1)$.

t-scores tell how many standard deviations a value is on the t-distribution from the value hypothesized by H_0, just like a z-score on the standard normal distribution.

Specific t-scores can be found using a **t-table**.

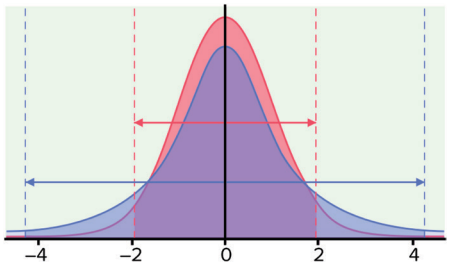
The confidence interval for means using a t-score is calculated with the following formula:

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$

The **t-statistic** is used when performing a **t-test**, or a hypothesis test with a t-distribution. It's calculated as follows:

$$t = \frac{x - \mu}{s / \sqrt{n}}$$

t-distributions are wider and have fatter tails than a normal distribution.



The confidence interval you'll get using a t-score is wider than what you'd get using a z-score with the same data.

Likewise, the rejection region on the t-distribution will be farther from the mean than it would be on the standard normal distribution.

Small Sample Decision Tree

This decision tree can help you determine which formulas are appropriate for a given situation.