# Statistics

## Descriptive Statistics

- Methods of organizing, summarizing, and presenting data in an informative way.
- Americans spent an average of $143.56 on Valentine's Day–related gifts in 2018. About 15 percent of Americans purchased gifts cards for Valentine's Day. In addition, they spent an average of $5.50 on gifts for their pets.

## Inferential Statistics

- The process of sampling from a population with the objective of estimating properties of a population is called inferential statistics**.**
- Television networks constantly monitor the popularity of their programs by hiring Nielsen and other organizations to sample the preferences of TV viewers. During the week of December 3, 2018, *The Tonight Show Starring Jimmy Fallon* was viewed by 2.26 million people in the 18–49 age. The *Late Show with Stephen Colbert* led the age group with 3.23 million viewers.These program ratings are used to make decisions about advertising rates and whether to continue or cancel a program.

# Qualitative Data-Qualitative or Categorical Data is data that can't be measured or counted in the form of numbers. These types of data are sorted by category, not by number. That's why it is also known as Categorical Data.

## Nominal Data

- Nominal Data is used to label variables without any order or quantitative value. The color of hair can be considered nominal data, as one color can't be compared with another color.
- With the help of nominal data, we can't do any numerical tasks or can't give any order to sort the data. These data don't have any meaningful order; their values are distributed into distinct categories.

**Examples of Nominal Data :**

- Colour of hair (Blonde, red, Brown, Black, etc.)
- Marital status (Single, Widowed, Married)
- Nationality (Indian, German, American)
- Gender (Male, Female, Others)
- Eye Color (Black, Brown, etc.)

## Ordinal Data

- Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them.
- Ordinal data is qualitative data for which their values have some kind of relative position.

**Examples of Ordinal Data :**

- When companies ask for feedback, experience, or satisfaction on a scale of 1 to 10
- Letter grades in the exam (A, B, C, D, etc.)
- Ranking of people in a competition (First, Second, Third, etc.)
- Economic Status (High, Medium, and Low)
- Education Level (Higher, Secondary, Primary)

Quantitative Data-Quantitative data can be expressed in numerical values, making it countable and including statistical data analysis. These kinds of data are also known as Numerical data.

## Discrete

- The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers. The total number of students in a class is an example of discrete data. These data can't be broken into decimal or fraction values. The discrete data are countable and have finite values

**Examples of Discrete Data :**

- Total numbers of students present in a class
- Numbers of employees in a company
- The total number of players who participated in a competition
- Days in a week

## Continuous

- Continuous data are in the form of fractional numbers. It can be the version of an android phone, the height of a person, the length of an object, etc. Continuous data represents information that can be divided into smaller levels. The continuous variable can take any value within a range.

**Examples of Continuous Data :**

- Height of a person
- Speed of a vehicle
- "Time-taken" to finish the work
- Wi-Fi Frequency
- Market share price

# Range, Variance and Standard Deviation

- Range, variance, and standard deviation all measure the spread or variability of a data set in different ways. **The range** is easy to calculate—it's the difference between the largest and smallest data points in a set. **Standard deviation** is the square root of the variance. Standard deviation is a measure of how spread out the data is from its mean.
- **Population variance** is a measure of how spread out a group of data points is. Specifically, it quantifies the average squared deviation from the mean. So, if all data points are very close to the mean, the variance will be small; if data points are spread out over a wide range, the variance will be larger.
- **Population standard deviation** is a measure of how much variation there is among individual data points in a population. It's a way of quantifying how spread out the data is from its mean. A small standard deviation means that the data points are generally close to the mean, while a large standard deviation means that the data is more dispersed.

# Co-efficient of Variation (CV) vs. Standard Deviation

- The standard deviation is a statistic that measures the dispersion of a data set relative to its mean. It is used to determine the spread of values in a single data set rather than to compare different units.
- When we want to compare two or more data sets, the co-efficient of variation is used. The CV is the ratio of the standard deviation to the mean. And because it’s independent of the unit in which the measurement was taken, it can be used to compare data sets with different units or widely different means.

# Coefficient of Skewness

Coefficient of skewness is one way to measure the skewness of a distribution. The coefficient of skewness can be defined as a measure that is used to determine the strength and direction of the skewness of a sample distribution by using descriptive statistics such as the mean, median, or mode. If the curve of a normal distribution is distorted towards the left or right then it is known as a skewed distribution. The most important measure of skewness is the coefficient of skewness that was given by Karl Pearson. It is also known as Pearson's coefficient of skewness.

**Coefficient of Skewness Interpretation**

Depending upon the value of the coefficient of skewness, the following inferences can be drawn about a distribution.

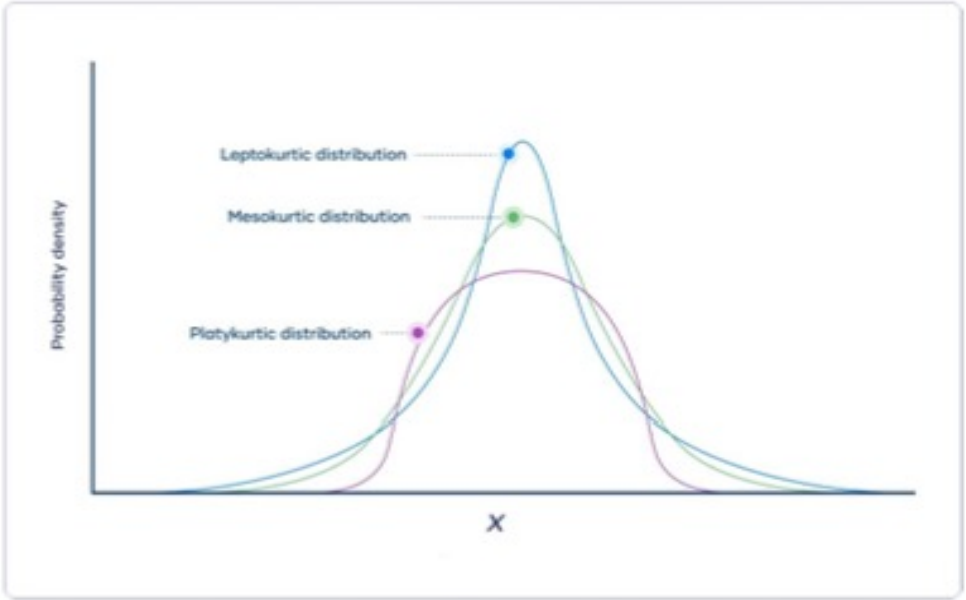
- If the mean exceeds the mode and median (Mode < Median < Mean) then the distribution is positively skewed. In other words, if the coefficient of skewness is positive then the distribution is skewed to the right.
- If the mode exceeds the median and mean (Mean < Median < Mode) then the distribution is negatively skewed. Thus, the coefficient of skewness will be negative and the distribution will be skewed to the left.
- If the value of the mean, median, and mode are equal then the distribution is a normal distribution and the coefficient of skewness will be 0.

# Kurtosis

Kurtosis is a statistic that measures the extent to which a distribution contains outliers. It assesses the propensity of a distribution to have extreme values within its tails. There are three kinds of kurtosis: leptokurtic, platykurtic, and mesokurtic. Statisticians define these types relative to the normal distribution. Higher kurtosis values indicate that the distribution has more outliers falling relatively far from the mean. Distributions with smaller values have a lower tendency for producing extreme values.

For instance, statisticians describe leptokurtic distributions as having higher kurtosis than the normal distribution. These distributions have “heavy tails,” indicating that they have relatively long tails that contain more outliers. Conversely, platykurtic distributions have “light tails” that are shorter and include fewer extreme values. Below, I’ll graph all three types for comparison.

| Value | Thickness of Tails | Interpretation |
|---|---|---|
| Kurtosis < 0 | Thin | Distribution is platykurtic |
| Kurtosis = 0 | Normal | Distribution is mesokurtic |
| Kurtosis > 0 | Thick | Distribution is leptokurtic |



# Covariance and Correlation

- Covariance and Correlation are very helpful in understanding the relationship between two continuous variables.
- Covariance tells whether both variables vary in the same direction (positive covariance) or in the opposite direction (negative covariance). There is no meaning of covariance numerical value only sign is useful.
- Whereas Correlation explains the change in one variable leads how much proportion change in the second variable. Correlation varies between -1 to +1. If the correlation value is 0 then it means there is no Linear Relationship between variables however other functional relationship may exist.

**COVARIANCE VS. CORRELATION**

Covariance reveals how two variables change together while correlation determines how closely two variables are related to each other.

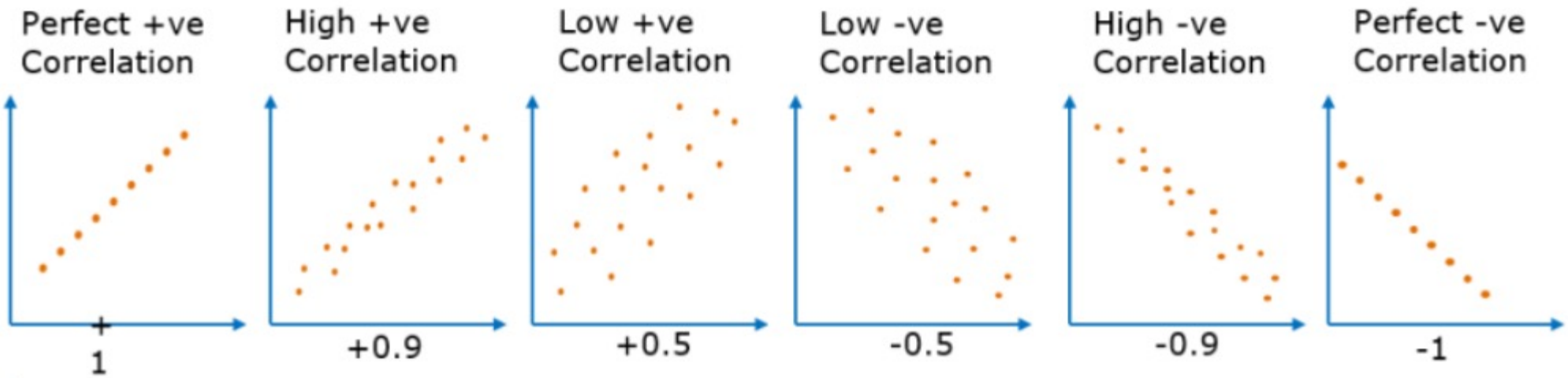
- Both covariance and correlation measure the relationship and the dependency between two variables.
- Covariance indicates the direction of the linear relationship between variables.
- Correlation measures both the strength and direction of the linear relationship between two variables.
- Correlation values are standardized.
- Covariance values are not standardized.

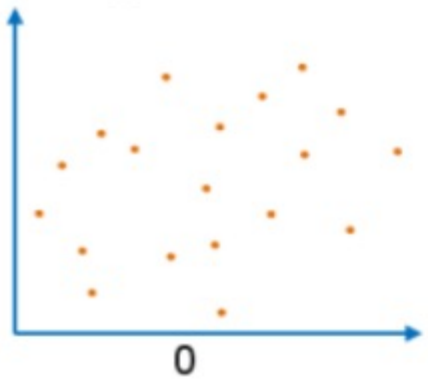
Correlation coefficient r is number between -1 to +1 and tells us how well a regression line fits the data and defined by

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where,

- $s_{xy}$ is the covariance between $x$ and $y$
- $s_x$ and $s_y$ are the standard deviations of $x$ and $y$ respectively.




1. Correlation coefficient r=0 mean there is no linear relationship between x and y however other functional relationship may exist.
2. One point to note here is if there is no relationship at all between x and y then r will certainly be 0 but not vice versa (refer point 1)

# What are Exhaustive Events?

- All possible outcomes of an experiment constitute exhaustive events as one of them will definitely occur. Now, exhaustive events may or may not be equally likely events, i.e., it is not necessary for events to have equal probability to be exhaustive.
- Let us consider an example of exhaustive events. There are six possible outcomes when rolling a die which is {1, 2, 3, 4, 5, 6}. Now, if we roll a die, one of these six outcomes will definitely occur. Hence, all these six outcomes are exhaustive events. Therefore, we can say that the union of the exhaustive events gives the entire sample space.
- Let us change the events for rolling a die and verify if the events are exhaustive and constitute the sample space. When rolling a die, let A be the event of getting a prime number, let B be the event of getting a composite number and C be the event of getting the number 1 (as 1 is neither prime nor composite and it is one of the possible outcomes). Now, we have A = {2, 3, 5}, B = {4, 6}, C = {1}. Now, when we roll a die, one of the six numbers - 1, 2, 3, 4, 5, 6 will occur which implies one of the events A, B, C will occur. Hence, these are exhaustive events. Also, A U B U C = {1, 2, 3, 4, 5, 6} = Sample Space.

# Bayes' Theorem

**Bayes' Theorem** is a way of finding a probability when we know certain other probabilities.

The formula is:

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

| | |
|---|---|
| Which tells us: | how often A happens *given that B happens*, written **P(A\|B)**, |
| When we know: | how often B happens *given that A happens*, written **P(B\|A)** |
| | and how likely A is on its own, written **P(A)** |
| | and how likely B is on its own, written **P(B)** |

# Bayes' Theorem

Let us say P(Fire) means how often there is fire, and P(Smoke) means how often we see smoke, then:

P(Fire|Smoke) means how often there is fire when we can see smoke
P(Smoke|Fire) means how often we can see smoke when there is fire

So the formula kind of tells us "forwards" P(Fire|Smoke) when we know "backwards" P(Smoke|Fire)

Example:

- dangerous fires are rare (1%)
- but smoke is fairly common (10%) due to barbecues,
- and 90% of dangerous fires make smoke

We can then discover the **probability of dangerous Fire when there is Smoke**:

$$P(\text{Fire}|\text{Smoke}) = \frac{P(\text{Fire})\,P(\text{Smoke}|\text{Fire})}{P(\text{Smoke})}$$

$$= \frac{1\% \times 90\%}{10\%}$$

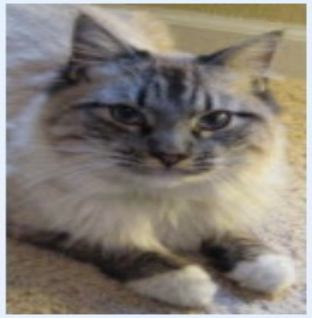
$$= 9\%$$

# Bayes' Theorem-With 2 A Cases-Cat Allergy-we have two possible cases for "A", such as **Pass**/**Fail** (or Yes/No etc)

## Example: Allergy or Not?

Hunter says she is itchy. There is a test for Allergy to Cats, but this test is not always right:

- For people that **really do** have the allergy, the test says "Yes" **80%** of the time
- For people that **do not** have the allergy, the test says "Yes" **10%** of the time ("false positive")



If 1% of the population have the allergy, and **Hunter's test says "Yes"**, what are the chances that Hunter really has the allergy?

We want to know the chance of having the allergy when test says "Yes", written **P(Allergy|Yes)**

Let's get our formula:

$$P(\text{Allergy}|\text{Yes}) = \frac{P(\text{Allergy})\ P(\text{Yes}|\text{Allergy})}{P(\text{Yes})}$$

- P(Allergy) is Probability of Allergy = 1%
- P(Yes|Allergy) is Probability of test saying "Yes" for people with allergy = 80%
- P(Yes) is Probability of test saying "Yes" (to anyone) = ??%

# Bayes' Theorem-With 2 A Cases-Cat Allergy

Oh no! We **don't know** what the **general** chance of the test saying "Yes" is ...

... but we can calculate it by adding up those **with**, and those **without** the allergy:

- 1% have the allergy, and the test says "Yes" to 80% of them
- 99% do **not** have the allergy and the test says "Yes" to 10% of them

Let's add that up:

$$P(\text{Yes}) = 1\% \times 80\% + 99\% \times 10\% = 10.7\%$$

Which means that about 10.7% of the population will get a "Yes" result.

So now we can complete our formula:

$$P(\text{Allergy}|\text{Yes}) = \frac{1\% \times 80\%}{10.7\%} = 7.48\%$$

$$P(\text{Allergy}|\text{Yes}) = \text{about } \mathbf{7\%}$$

In fact we can write a special version of the Bayes' formula just for things like this:

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\text{not } A)P(B|\text{not } A)}$$

# Bayes' Theorem-With 3 or More A Cases

We just saw "A" with two cases (A and not A), which we took care of in the bottom line.

When "A" has 3 or more cases we include them all in the bottom line:

$$P(A1|B) = \frac{P(A1)P(B|A1)}{P(A1)P(B|A1) + P(A2)P(B|A2) + P(A3)P(B|A3) + \ldots etc}$$

Example: The Art Competition has entries from three painters: Pam, Pia and Pablo

- Pam put in 15 paintings, 4% of her works have won First Prize.
- Pia put in 5 paintings, 6% of her works have won First Prize.
- Pablo put in 10 paintings, 3% of his works have won First Prize.

What is the chance that Pam will win First Prize?

# Bayes' Theorem-With 3 or More A Cases

$$P(\text{Pam}|\text{First}) = \frac{P(\text{Pam})P(\text{First}|\text{Pam})}{P(\text{Pam})P(\text{First}|\text{Pam}) + P(\text{Pia})P(\text{First}|\text{Pia}) + P(\text{Pablo})P(\text{First}|\text{Pablo})}$$

Put in the values:

$$P(\text{Pam}|\text{First}) = \frac{(15/30) \times 4\%}{(15/30) \times 4\% + (5/30) \times 6\% + (10/30) \times 3\%}$$

Multiply all by 30 (makes calculation easier):

$$P(\text{Pam}|\text{First}) = \frac{15 \times 4\%}{15 \times 4\% + 5 \times 6\% + 10 \times 3\%}$$

$$= \frac{0.6}{0.6 + 0.3 + 0.3}$$

$$= 50\%$$

# Difference between Mean and Expected Value

The mean and the expected value are both measures of central tendency, but they are used in different contexts. The mean is the arithmetic average of a set of numbers. It is calculated by adding up all the numbers in the set and dividing by the number of items in the set. For example, the mean of the set {1, 2, 3, 4} is (1 + 2 + 3 + 4)/4 = 2.5.The expected value, also known as mathematical expectation or mean value, is a concept in probability theory. It represents the long-term average of a random variable. For example, if you roll a fair die, the expected value of the roll is 3.5 (since each face has a value of 1 to 6, and the average of those numbers is 3.5).In summary, the mean is a measure of central tendency for a set of data, while the expected value is a measure of central tendency for a random variable in probability theory.

$$\text{expected value} = E(X) = \sum_{i=1}^{n} x_i p_i$$

$$\text{mean value} = E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

# Random Variable

- A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes. Random variables are often designated by letters and can be classified as discrete, which are variables that have specific values, or continuous, which are variables that can have any values within a continuous range.
- In probability and statistics, random variables are used to quantify outcomes of a random occurrence, and therefore, can take on many values. Random variables are required to be measurable and are typically real numbers. For example, the letter X may be designated to represent the sum of the resulting numbers after three dice are rolled. In this case, X could be 3 (1 + 1+ 1), 18 (6 + 6 + 6), or somewhere between 3 and 18, since the highest number of a die is 6 and the lowest number is 1.
- In the corporate world, random variables can be assigned to properties such as the average price of an asset over a given time period, the return on investment after a specified number of years, the estimated turnover rate at a company within the following six months, etc.

# Example of a Random Variable

A typical example of a random variable is the outcome of a coin toss. Consider a probability distribution in which the outcomes of a random event are not equally likely to happen. If the random variable Y is the number of heads we get from tossing two coins, then Y could be 0, 1, or 2. This means that we could have no heads, one head, or both heads on a two-coin toss.

However, the two coins land in four different ways: TT, HT, TH, and HH. Therefore, the P(Y=0) = 1/4 since we have one chance of getting no heads (i.e., two tails [TT] when the coins are tossed). Similarly, the probability of getting two heads (HH) is also 1/4. Notice that getting one head has a likelihood of occurring twice: in HT and TH. In this case, P (Y=1) = 2/4 = 1/2.

Upper case letters such as X or Y denote a random variable. Lower case letters like x or y denote the value of a random variable.

# Discrete and Continuous Random Variable

***Discrete Random Variables***

Discrete random variables take on a countable number of distinct values. Consider an experiment where a coin is tossed three times. If X represents the number of times that the coin comes up heads, then X is a discrete random variable that can only have the values 0, 1, 2, or 3 (from no heads in three successive coin tosses to all heads). No other value is possible for X.

***Continuous Random Variables***

Continuous random variables can represent any value within a specified range or interval and can take on an infinite number of possible values. An example of a continuous random variable would be an experiment that involves measuring the amount of rainfall in a city over a year or the average height of a random group of 25 people.

Drawing on the latter, if Y represents the random variable for the average height of a random group of 25 people, you will find that the resulting outcome is a continuous figure since height may be 5 ft or 5.01 ft or 5.0001 ft. Clearly, there is an infinite number of possible values for height.

# Mean and Variance of Bernoulli Distribution

**Mean of Bernoulli Distribution Proof:**

We know that for X,

$P(X = 1) = p$

$P(X = 0) = q$

$E[X] = P(X = 1) . 1 + P(X = 0) . 0$

$E[X] = p . 1 + q . 0$

$E[X] = p$

Thus, the mean or expected value of a Bernoulli distribution is given by $E[X] = p$.

**Variance of Bernoulli Distribution Proof:**

The variance can be defined as the difference of the mean of $X^2$ and the square of the mean of X. Mathematically this statement can be written as follows:

$Var[X] = E[X^2] - (E[X])^2$

Using the properties of $E[X^2]$, we get,

$E[X^2] = \sum x^2 P(X=x)$

$E[X^2] = 1^2 . p + 0^2 . q = p$

Substituting this value in $Var[X] = E[X^2] - (E[X])^2$ we have

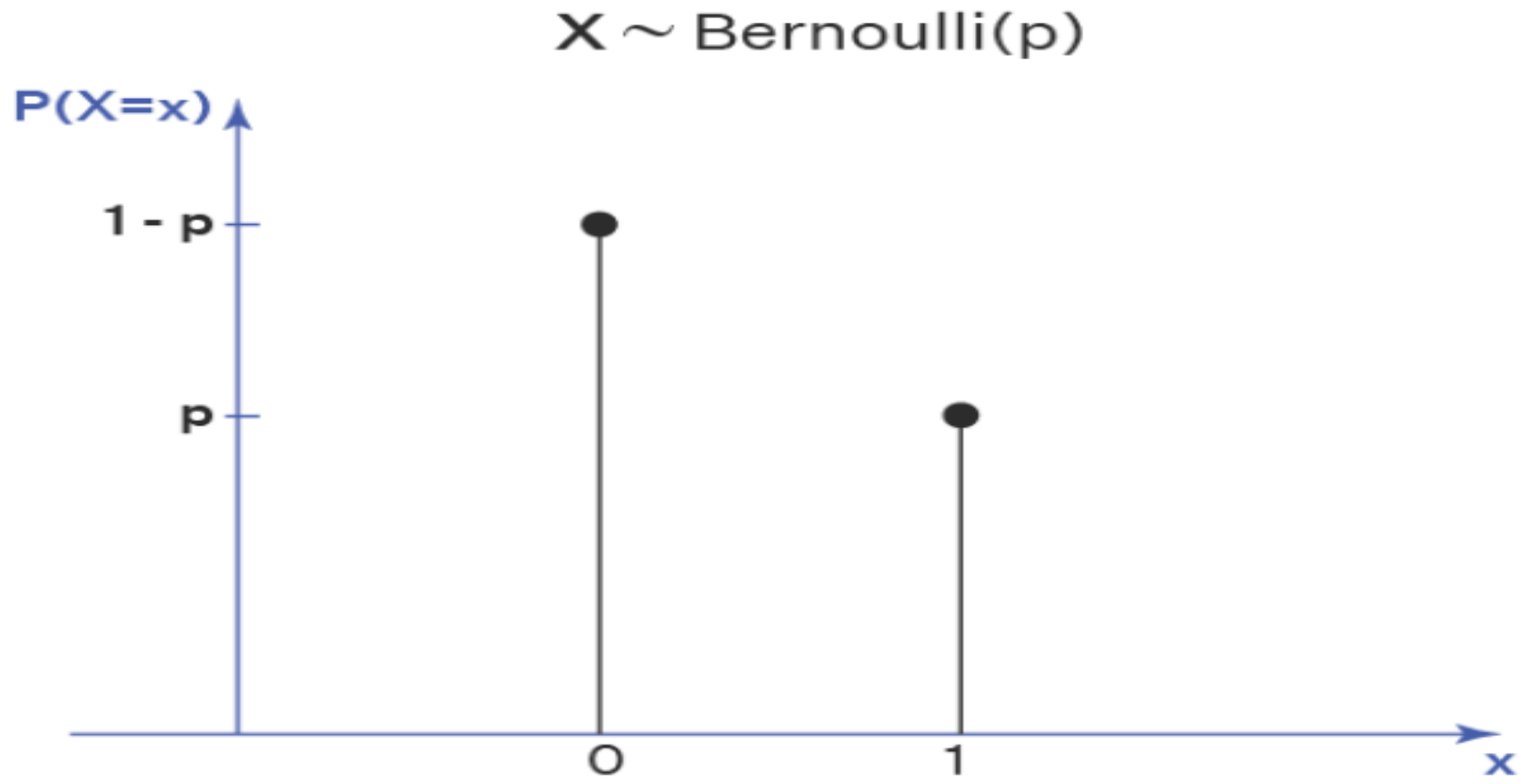
$Var[X] = p - p^2$

$= p(1 - p)$

$= p . q$

Hence, the variance of a Bernoulli distribution is $Var[X] = p(1 - p) = p . q$

# Bernoulli Distribution Graph




The graph shows that the probability of success is p when X = 1 and the probability of failure of X is (1 - p) or q if X = 0.

# Binomial Probability

**BINOMIAL PROBABILITY EXPERIMENT**

1. An outcome on each trial of an experiment is classified into one of two mutually exclusive categories—a success or a failure.
2. The random variable is the number of successes in a fixed number of trials.
3. The probability of success is the same for each trial.
4. The trials are independent, meaning that the outcome of one trial does not affect the outcome of any other trial.

## How Is a Binomial Probability Computed?

To construct a particular binomial probability, we use (1) the number of trials and (2) the probability of success on each trial. For example, if the Hannah Landscaping Company plants 10 Norfolk pine trees today knowing that 90% of these trees survive, we can compute the binomial probability that exactly 8 trees survive. In this case the number of trials is the 10 trees, the probability of success is .90, and the number of successes is eight. In fact, we can compute a binomial probability for any number of successes from 0 to 10 surviving trees.

A binomial probability is computed by the formula:

**BINOMIAL PROBABILITY FORMULA**

$$P(x) = {}_nC_x \pi^x (1 - \pi)^{n-x} \qquad \textbf{(6–3)}$$

where:

$C$ denotes a combination.
$n$ is the number of trials.
$x$ is the random variable defined as the number of successes.
$\pi$ is the probability of a success on each trial.

We use the Greek letter π (pi) to denote a binomial population parameter. Do not confuse it with the mathematical constant 3.1416.

# Binomial Probability Example

Debit and credit cards are widely used to make purchases. Recently, www.creditcards.com reported 28% of purchases at coffee shops were made with a debit card. For 10 randomly selected purchases at the Starbucks on the corner of 12th Street and Main, what is the probability exactly one of the purchases was made with a debit card? What is the probability distribution for the random variable, number of purchases made with a debit card? What is the probability that six or more purchases out of 10 are made with a debit card? What is the probability that five or fewer purchases out of 10 are made with a debit card?

## SOLUTION

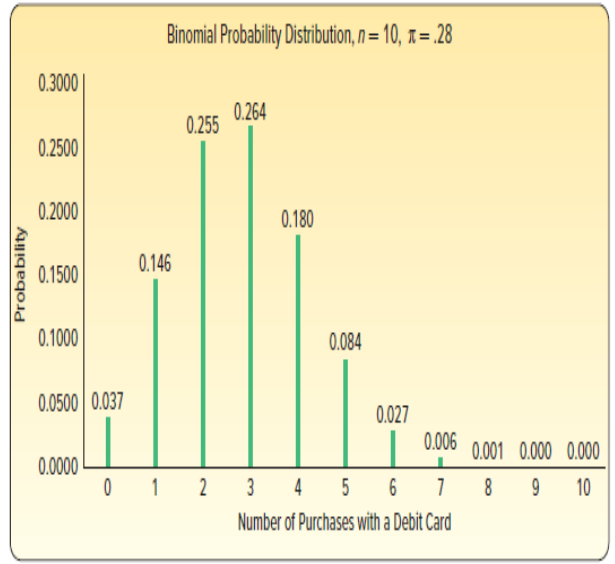
This example fits all the requirements for a binomial distribution. The probability of success, a purchase made with a debit card, is .28, so let $\pi$ =.28. We determined the number of purchases to be 10, so the number of trials is 10 and $n$ = 10. The trials are independent, and the probability of success is the same for each trial. The random variable, $x$, is the number purchases with a debit card in 10 trials. The random variable, $x$, can be equal to 0, no purchases made with a debit card, 1, one purchase made with a debit card, or 2, 3, 4, or 10 purchases made with a debit card. To calculate the probability for each value of the random variable, apply formula 6–3. The probability that no purchases in 10 trials are made with a debit card is:

$$P(0) = {}_nC_x(\pi)^x(1-\pi)^{n-x} = {}_{10}C_0(.28)^0(1-.28)^{10-0} = (1)(1)(.0374) = .0374$$

The probability that exactly one of the 10 purchases is made with a debit card is .1456, found by:

$$P(1) = {}_nC_x(\pi)^x(1-\pi)^{n-x} = {}_{10}C_1(.28)^1(1-.28)^{10-1} = (10)(.28)(.0520) = .1456$$

Using statistical software, the entire binomial probability distribution with $\pi$ =.28 and $n$ = 10 is shown in the following bar chart and table.



# Binomial Probability Mean and variance

The mean (μ) and the variance ($\sigma^2$) of a binomial distribution are computed in a "shortcut" fashion by:

**MEAN OF A BINOMIAL DISTRIBUTION**

$$\mu = n\pi$$

**VARIANCE OF A BINOMIAL DISTRIBUTION**

$$\sigma^2 = n\pi(1 - \pi)$$

For the example regarding the number of debit purchases in the sample of five customers, recall that $\pi$ =.28 and $n$ = 10. Hence:

$$\mu = n\pi = (10)(.28) = 2.8$$
$$\sigma^2 = n\pi(1 - \pi) = 10\ (.28)\ (1 - .28) = 2.016$$
$$\sigma = 1.420$$

# Poisson Probability Distribution

The Poisson distribution is a probability distribution that describes the number of events that occur within a fixed interval of time or space, given a certain average rate of occurrence and under certain assumptions.

The Poisson distribution is often used in scenarios where events occur randomly and independently, and the events are rare within the given interval. Some common examples of situations that can be modeled using the Poisson distribution include:

1. **Phone Calls:** The number of phone calls received by a call center in an hour.
2. **Accidents:** The number of accidents occurring at a specific intersection in a day.
3. **Typographical Errors:** The number of typographical errors on a page of a book.
4. **Radioactive Decay:** The number of particles decaying in a certain substance over a fixed period.

# Poisson Probability Distribution

The Poisson distribution is characterized by a single parameter, often denoted as λ (lambda), which represents the average rate of events occurring in the interval. The probability mass function (PMF) of the Poisson distribution is given by the formula:

$P(X = k) = (e^{(-\lambda)} * \lambda^k) / k!$

Where:

- $P(X = k)$ is the probability of observing exactly k events in the interval.
- e is the base of the natural logarithm (approximately 2.71828).
- λ is the average rate of events.
- k is the number of events observed.
- k! represents the factorial of k ($k! = k * (k - 1) * (k - 2) * ... * 2 * 1$).

It's important to note that the Poisson distribution is most suitable for situations where events are rare and independent. If events become more frequent or are not independent, other distributions like the binomial distribution or the normal distribution might be more appropriate.

In practice, the Poisson distribution can be useful for predicting the likelihood of certain events occurring within a given time or space interval based on historical data or assumptions about the average rate of occurrence.

# Poisson Probability Distribution-Example

Example: Customer Arrivals at a Café Suppose you run a small café and you're interested in modeling the number of customers who arrive during the morning rush hour (from 8:00 AM to 10:00 AM). On average, you've observed that around 5 customers arrive during this time period.
You can use the Poisson distribution to answer questions like: What's the probability of having exactly 3 customers during this time frame? Or what's the probability of having more than 7 customers?
Using the Poisson distribution formula: $P(X = k) = (e^{(-\lambda)} * \lambda^k) / k!$
Where:
λ (lambda) = 5 (average number of customer arrivals during the morning rush hour) k = the number of customers we're interested in

Let's calculate a few probabilities:

1. Probability of having exactly 3 customers (k = 3): $P(X = 3) = (e^{(-5)} * 5^3) / 3! \approx 0.1404$
2. Probability of having more than 7 customers: $P(X > 7) = 1 - P(X \leq 7)$ Calculate $P(X \leq 7)$ by adding up probabilities for k = 0 to 7: $P(X \leq 7) = P(X = 0) + P(X = 1) + P(X = 2) + ... + P(X = 7)$ Then, subtract from 1 to get $P(X > 7)$.

This example demonstrates how the Poisson distribution can be used to estimate the likelihood of specific event counts occurring based on an average rate. Remember that the Poisson distribution is appropriate when events are rare and independent, which might not be the case if the arrival rate becomes very high or if customer arrivals are influenced by external factors.

# Poisson Probability Distribution-Mean and Variance

For a Poisson distribution with parameter $\lambda$ (lambda), the mean ($\mu$) and variance ($\sigma^2$) are given by the same parameter $\lambda$:

Mean ($\mu$) = $\lambda$ Variance ($\sigma^2$) = $\lambda$

In mathematical terms, this relationship is expressed as follows:

$\mu = \lambda$ $\sigma^2 = \lambda$

This means that the average number of events in a Poisson distribution is equal to its variance. This relationship is a unique characteristic of the Poisson distribution. It also makes sense intuitively, as the Poisson distribution is based on the assumption that events occur at a constant average rate, and both the mean and variance reflect this rate of occurrence.

So, in the example we discussed earlier with the café's customer arrivals, if the average number of customers arriving during the morning rush hour is $\lambda = 5$, then both the mean and the variance of customer arrivals during that time period would be 5.

# Continuous Probability Distribution

A continuous probability distribution usually results from measuring something, such as the distance from the dormitory to the classroom, the weight of an individual, or the amount of bonus earned by CEOs. As an example, at Dave's Inlet Fish Shack flounder is the featured, fresh-fish menu item. The distribution of the amount of flounder sold per day has a mean of 10.0 pounds per day and a standard deviation of 3.0 pounds per day. This distribution is continuous because Dave, the owner, "measures" the amount of flounder sold each day. It is important to realize that a continuous random variable has an infinite number of values within a particular range. So, for a continuous random variable, probability is for a range of values.

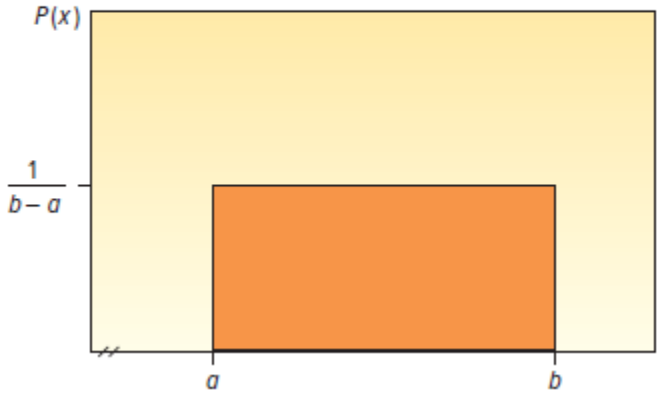
## *Uniform Probability Distributions:*

The uniform probability distribution is the simplest distribution for a continuous random variable. This distribution is rectangular in shape and is completely defined by its minimum and maximum values. Here are some examples that follow a uniform distribution.

- The sales of gasoline at the Kwik Fill in Medina, New York, follow a uniform distribution that varies between 2,000 and 5,000 gallons per day. The random variable is the number of gallons sold per day and is continuous within the interval between 2,000 gallons and 5,000 gallons.
- Volunteers at the Grand Strand Public Library prepare federal income tax forms. The time to prepare form 1040-EZ follows a uniform distribution over the interval between 10 minutes and 30 minutes. The random variable is the number of minutes to complete the form, and it can assume any value between 10 and 30.

# *Continuous Uniform Probability Distributions:*

A uniform distribution is shown in Chart 7–1. The distribution's shape is rectangular and has a minimum value of *a* and a maximum of *b*. Also notice in Chart 7–1 the height of the distribution is constant or uniform for all values between *a* and *b*.




The mean of a uniform distribution is located in the middle of the interval between the minimum and maximum values. It is computed as:

**MEAN OF THE UNIFORM DISTRIBUTION**

$$\mu = \frac{a + b}{2}$$

The standard deviation describes the dispersion of a distribution. In the uniform distribution, the standard deviation is also related to the interval between the maximum and minimum values.

**STANDARD DEVIATION OF THE UNIFORM DISTRIBUTION**

$$\sigma = \sqrt{\frac{(b - a)^2}{12}}$$

# *Uniform Probability Distributions:*

The continuous uniform distribution is a type of probability distribution in statistics and probability theory. It describes a random variable that takes on values within a certain range with equal probability. In other words, every value within the range is equally likely to occur.

The probability density function (PDF) of a continuous uniform distribution is defined as:

```
f(x | a, b) = 1 / (b - a) for a ≤ x ≤ b
```

Where:

- `a` is the lower bound of the distribution (minimum value)
- `b` is the upper bound of the distribution (maximum value)
- `x` is a specific value within the range `[a, b]`
- `f(x | a, b)` is the probability density function at `x` given the parameters `a` and `b`

The cumulative distribution function (CDF) of the continuous uniform distribution is given by:

```
F(x | a, b) = 0 for x < a
F(x | a, b) = (x - a) / (b - a) for a ≤ x ≤ b
F(x | a, b) = 1 for x > b
```

# *Properties of Uniform Probability Distributions*

Some properties of the continuous uniform distribution:

1. Mean: The mean (expected value) of a continuous uniform distribution with parameters `a` and `b` is `(a + b) / 2`.
2. Variance: The variance of the distribution is `(b - a)^2 / 12`.
3. Range: The range of the distribution is `[a, b]`.
4. Uniformity: All values within the range have equal probability density, making it a uniform distribution.

The continuous uniform distribution is commonly used in various applications, such as random number generation, modeling situations where each outcome is equally likely within a specific range, and as a building block for more complex distributions and simulations.

# *Uniform Probability Distributions: Example*

**Example: Random Time Arrival**

Imagine a situation where customers arrive at a coffee shop between 9:00 AM and 10:00 AM. The arrival times are modeled using a continuous uniform distribution. The coffee shop opens at 9:00 AM (`a = 9`) and closes at 10:00 AM (`b = 10`). The goal is to find the probability of a customer arriving between a specific time range.

Let's say we want to find the probability that a customer arrives between 9:30 AM and 9:45 AM. We can use the cumulative distribution function (CDF) to calculate this:

$$F(9.45 \mid 9, 10) - F(9.30 \mid 9, 10) = (9.45 - 9) / (10 - 9) - (9.30 - 9) / (10 - 9) = 0.45 - 0.30 = 0.15$$

So, the probability that a customer arrives between 9:30 AM and 9:45 AM is 0.15 or 15%.

This example demonstrates how the continuous uniform distribution can be used to model the random arrival times of customers within a specified time range. Keep in mind that this is a simplified example, and in real-world scenarios, other factors might also influence arrival times.

# Sampling

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. This is called a **sampling method**. There are two primary types of sampling methods that you can use in your research:

- **Probability sampling** involves random selection, allowing you to make strong statistical inferences about the whole group.
- **Non-probability sampling** involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

# Probability Sampling

## Simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population. To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Example: You want to select a simple random sample of 1000 employees of a social media marketing company. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

## Stratified sampling

- Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.
- To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).
- Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Example: The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

# Probability Sampling

## Systematic sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example: All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

## Cluster sampling

- Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.
- If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.
- This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.

Example: The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

# Non-Probability Sampling Methods

## Convenience sampling

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results. Convenience samples are at risk for both sampling bias and selection bias.

Example: You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

## Voluntary response sampling

Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).

Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others, leading to self-selection bias.

Example: You send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can't be sure that their opinions are representative of all students.

# Non-Probability Sampling Methods

## Purposive sampling

This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

Example: You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.

## Snowball sampling

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to “snowballs” as you get in contact with more people. The downside here is also representativeness, as you have no way of knowing how representative your sample is due to the reliance on participants recruiting others. This can lead to sampling bias.

Example: You are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn’t possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.

## Quota sampling

Quota sampling relies on the non-random selection of a predetermined number or proportion of units. This is called a quota.

You first divide the population into mutually exclusive subgroups (called strata) and then recruit sample units until you reach your quota. These units share specific characteristics, determined by you prior to forming your strata. The aim of quota sampling is to control what or who makes up your sample.

Example: You want to gauge consumer interest in a new produce delivery service in Boston, focused on dietary preferences. You divide the population into meat eaters, vegetarians, and vegans, drawing a sample of 1000 people. Since the company wants to cater to all consumers, you set a quota of 200 people for each dietary group. In this way, all dietary preferences are equally represented in your research, and you can easily compare these groups.You continue recruiting until you reach the quota of 200 participants for each subgroup.

# Parameter vs Statistic

A **parameter** is a number describing a whole population (e.g., population mean), while a **statistic** is a number describing a sample (e.g., sample mean).

The goal of quantitative research is to understand characteristics of populations by finding parameters. In practice, it's often too difficult, time-consuming or unfeasible to collect data from every member of a population. Instead, data is collected from samples.

With inferential statistics, we can use sample statistics to make educated guesses about population parameters.

| Statistical Notation | Population Parameter | Sample Statistic |
|---|---|---|
| Mean | $\mu$ | $\bar{x}$ |
| Variance | $\sigma^2$ | $s^2$ |
| Standard Deviation | $\sigma$ | $s$ |
| Standard Error of mean | $\sigma_{\bar{x}}$ | $s_{\bar{x}}$ |
| Standard Error of proportion | $\sigma_p$ | $s_p$ |

# Standard Deviation vs Standard Error

- Standard deviation describes variability within a single sample, while standard error describes variability across multiple samples of a population.
- Standard deviation is a descriptive statistic that can be calculated from sample data, while standard error is an inferential statistic that can only be estimated.
- Standard deviation measures how much observations vary from one another, while standard error looks at how accurate the mean of a sample of data is compared to the true population mean.
- The formula for standard deviation calculates the square root of the variance, while the formula for standard error calculates the standard deviation divided by the square root of the sample size.

***When to use which one:***

**Standard deviation** is useful when you need to compare and describe different data values that are widely scattered within a single dataset. Because standard deviation measures how close each observation is to the mean, it can tell you how precise the measurements are. So, if you have a dataset forecasting air pollution for a certain city, a standard deviation of 0.89 (i.e. a low standard deviation) shows you that the data is precise.

**Standard error** is useful if you want to test a hypothesis, as it allows you to gauge how accurate and precise your sample data is in relation to drawing conclusions about the actual overall population. For example, if you want to investigate the spending habits of everyone over 50 in New York City, using a sample of 500 people, standard error can tell you how “powerful” or applicable your findings are.

# Simple and Composite Hypothesis

A simple hypothesis is one in which all parameters of the distribution are specified. For example, the heights of college students are normally distributed with $\sigma^2 = 4$, and the hypothesis that its mean $\mu$ is, say, $62$"; that is, $H_o : \mu = 62$. So we have stated a simple hypothesis, as the mean and variance together specify a normal distribution completely. A simple hypothesis, in general, states that $\theta = \theta_o$ where $\theta_o$ is the specified value of a parameter $\theta$, ($\theta$ may represent $\mu, p, \mu_1 - \mu_2$ etc).

A hypothesis which is not simple (i.e. in which not all of the parameters are specified) is called a composite hypothesis. For instance, if we hypothesize that $H_o : \mu > 62$ (and $\sigma^2 = 4$) or $H_o : \mu = 62$ and $\sigma^2 < 4$, the hypothesis becomes a composite hypothesis because we cannot know the exact distribution of the population in either case. Obviously, the parameters $\mu > 62$" and$\sigma^2 < 4$ have more than one value and no specified values are being assigned. The general form of a composite hypothesis is $\theta \leqslant \theta_o$ or $\theta \geqslant \theta_o$; that is, the parameter $\theta$ does not exceed or does not fall short of a specified value $\theta_o$. The concept of simple and composite hypotheses applies to both the null hypothesis and alternative hypothesis.

# Types of Errors-Type I and Type II Error

A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

In statistics, a **Type I error** is a false positive conclusion, while a **Type II error** is a false negative conclusion.

Making a statistical decision always involves uncertainties, so the risks of making these errors are unavoidable in hypothesis testing.

The probability of making a Type I error is the significance level, or alpha (α), while the probability of making a Type II error is beta (β). These risks can be minimized through careful planning in your study design.

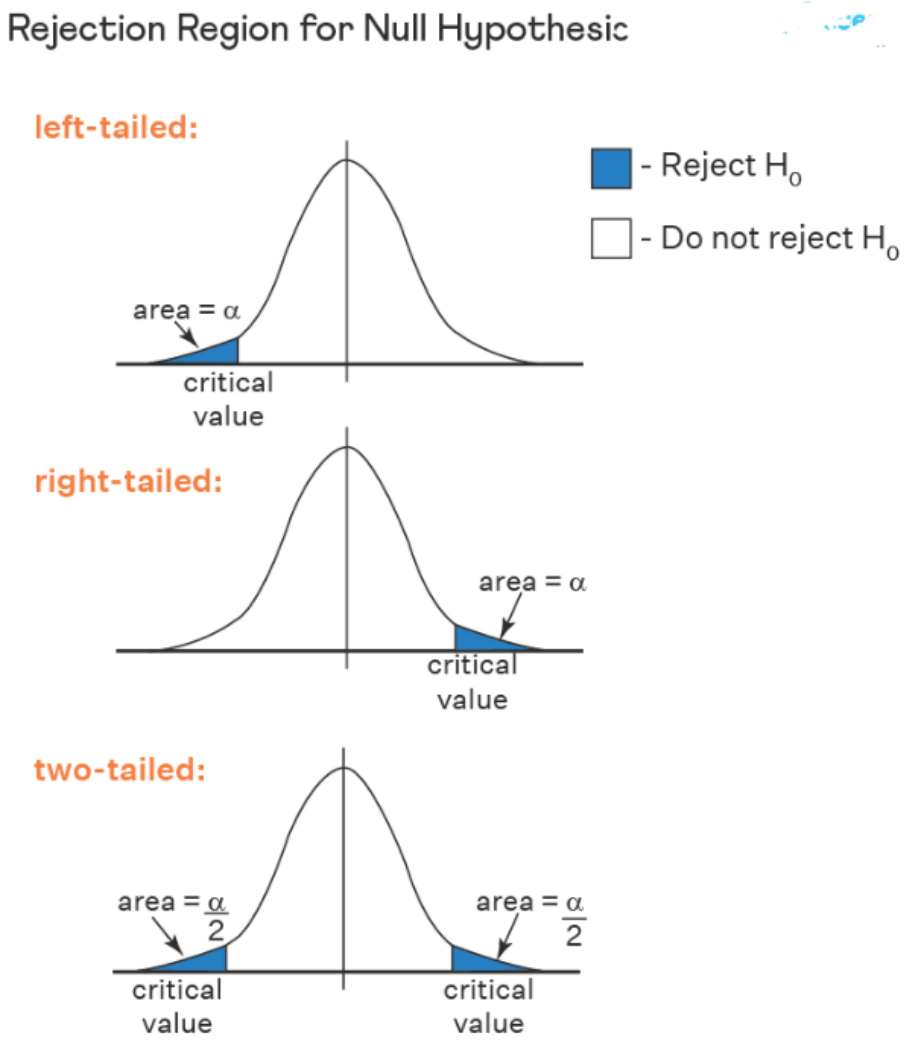
Example: Type I vs Type II error You decide to get tested for COVID-19 based on mild symptoms. There are two errors that could potentially occur:

**Type I error (false positive):** the test result says you have coronavirus, but you actually don't.

**Type II error (false negative):** the test result says you don't have coronavirus, but you actually do.

# Critical Region

A critical region, also known as the rejection region, is a set of values for the test statistic for which the null hypothesis is rejected. i.e. if the observed test statistic is in the critical region then we reject the null hypothesis and accept the alternative hypothesis.



# Test Statistic

In statistics, a test statistic is a numerical summary of a sample that is used to make inferences about a population. It's a crucial component of hypothesis testing and helps you determine whether to accept or reject a null hypothesis based on the data you've collected.

Here's a general overview of how test statistics work in hypothesis testing:

1. **State the Hypotheses**: Start by stating your null hypothesis (H0) and alternative hypothesis (Ha). The null hypothesis usually represents the status quo or no effect, while the alternative hypothesis represents the effect you're trying to investigate.
2. **Choose a Test**: Depending on your data and research question, you'll choose a specific statistical test. Common examples include t-tests, z-tests, chi-squared tests, ANOVA, and regression analysis.
3. **Collect and Analyze Data**: Gather your sample data and perform the necessary calculations to obtain a test statistic. The formula for the test statistic varies based on the chosen test.
4. **Determine the Critical Region**: Define a significance level (often denoted by α) that represents the threshold for considering results as statistically significant. This helps you determine the critical region in the distribution of the test statistic.
5. **Compare Test Statistic and Critical Value**: Compare the calculated test statistic to the critical value(s) associated with the chosen significance level. If the test statistic falls into the critical region, you reject the null hypothesis in favor of the alternative hypothesis.
6. **Calculate P-value (Optional)**: In many cases, you can also calculate a p-value associated with the test statistic. The p-value represents the probability of observing a test statistic as extreme as, or more extreme than, the one calculated, assuming that the null hypothesis is true. A low p-value suggests that the null hypothesis is unlikely to be true.
7. **Make a Decision**: Based on the comparison of the test statistic and the critical value (or p-value), you make a decision regarding the null hypothesis. If the test statistic is beyond the critical value or the p-value is below the significance level, you reject the null hypothesis. Otherwise, you fail to reject it.

It's important to note that different tests have different test statistics and formulas for calculating them. The choice of test depends on factors like the nature of your data, the research question, and the assumptions underlying the test.

Remember that statistical hypothesis testing involves some complexity and requires careful consideration of assumptions and limitations. It's recommended to have a good understanding of statistical concepts and data analysis techniques before conducting hypothesis tests.

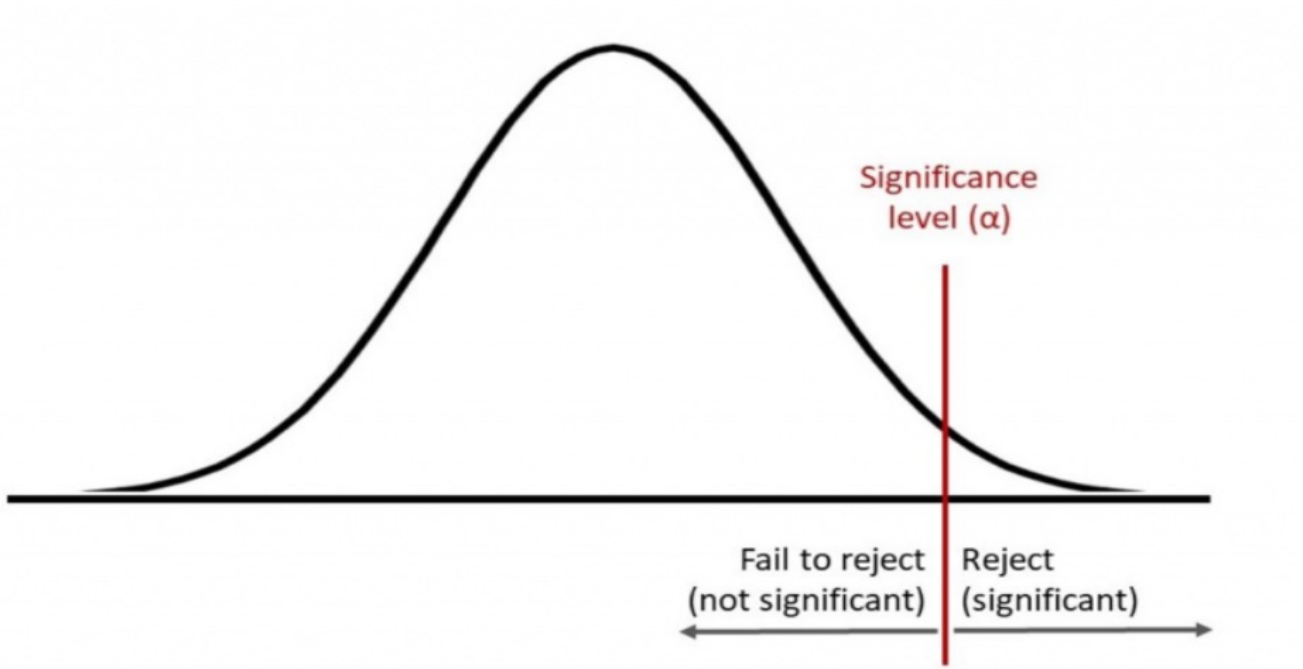
# Level of Significance

The level of significance, often denoted as "α" (alpha), is a critical value used in hypothesis testing to determine whether there is enough evidence to reject a null hypothesis.

In hypothesis testing, you have two hypotheses:

1. Null Hypothesis (H0): This is the default or initial assumption that there is no effect, no difference, or no relationship in the population. It's often stated as the hypothesis to be tested.
2. Alternative Hypothesis (Ha or H1): This is the hypothesis that contradicts the null hypothesis. It suggests that there is an effect, a difference, or a relationship in the population.

The level of significance represents the threshold for making a decision about the null hypothesis. It helps in determining when to reject the null hypothesis in favor of the alternative hypothesis. Common levels of significance include 0.05 (5%), 0.01 (1%), and 0.10 (10%).



# Level of Significance

Here's how it works:

1. Choose a significance level (α) before conducting the test. This level represents the maximum probability of making a Type I error, which is the incorrect rejection of a true null hypothesis (false positive).
2. Collect and analyze the sample data.
3. Calculate the test statistic (such as a t-statistic or z-score) based on the sample data and the statistical test being used.
4. Compare the calculated test statistic with the critical value(s) from the appropriate statistical distribution (such as the t-distribution or normal distribution). The critical value(s) are determined based on the chosen significance level and the degrees of freedom.
5. If the calculated test statistic is more extreme than the critical value(s), you would reject the null hypothesis. If it's not as extreme, you would fail to reject the null hypothesis.
6. Make a decision and draw a conclusion based on the comparison. If you reject the null hypothesis, you would typically support the alternative hypothesis. If you fail to reject the null hypothesis, you do not have enough evidence to support the alternative hypothesis.

Choosing an appropriate significance level is important. A lower significance level reduces the likelihood of making a Type I error but increases the likelihood of making a Type II error (failing to reject a false null hypothesis). The choice of significance level depends on the consequences of making each type of error and the specific context of the problem.

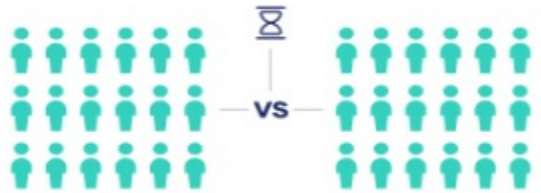
# Testing for Mean

A *t* test can only be used when comparing the means of two groups (a.k.a. pairwise comparison). If you want to compare more than two groups, or if you want to do multiple pairwise comparisons, use an ANOVA test or a post-hoc test.

## What type of *t* test should I use?

When choosing a *t* test, you will need to consider two things: whether the groups being compared come from a single population or two different populations, and whether you want to test the difference in a specific direction.

### Paired-samples t test

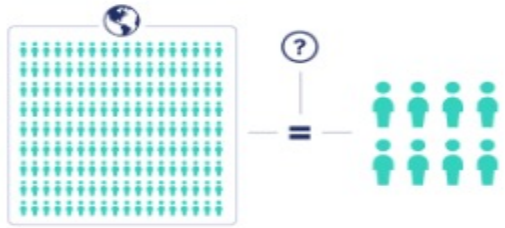



Investigate whether there's a difference within a group between two points in time (within-subjects).

### Independent-samples t test


Investigate whether there's a difference between two groups (between-subjects).

### One-sample t test




Investigate whether there's a difference between a group and a standard value or whether a subgroup belongs to a population.

# Testing for Mean

## One-sample, two-sample, or paired *t* test?

- If the groups come from a single population (e.g., measuring before and after an experimental treatment), perform a **paired *t* test**. This is a within-subjects design.
- If the groups come from two different populations (e.g., two different species, or people from two separate cities), perform a **two-sample *t* test** (a.k.a. **independent *t* test**). This is a between-subjects design.
- If there is one group being compared against a standard value (e.g., comparing the acidity of a liquid to a neutral pH of 7), perform a **one-sample *t* test**.

## One-tailed or two-tailed *t* test?

- If you only care whether the two populations are different from one another, perform a **two-tailed *t* test**.
- If you want to know whether one population mean is greater than or less than the other, perform a **one-tailed *t* test.**

## t test example: In your test of whether petal length differs by species:

- Your observations come from two separate populations (separate species), so you perform a two-sample *t* test.
- You don't care about the direction of the difference, only whether there is a difference, so you choose to use a two-tailed *t* test.

**One-Sample Testing:**

In one-sample testing for means, the most commonly used test statistic is the **t-statistic**.

The formula for the t-statistic in this case is:

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

Where:

- $\bar{x}$ is the sample mean.
- $\mu$ is the hypothesized population mean.
- $s$ is the sample standard deviation.
- $n$ is the sample size.

This t-statistic follows a t-distribution with $n - 1$ degrees of freedom. You then compare the calculated t-statistic to critical values from the t-distribution or use p-values to determine whether the observed difference is statistically significant.

**Two-Sample Testing:**

For independent two-sample testing, you can use either the **pooled t-test** (assuming equal variances) or **Welch's t-test** (allowing for unequal variances). The formula for both tests' statistics is similar, with the main difference lying in the calculation of the standard error.

For the pooled t-test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

For Welch's t-test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where the subscripts 1 and 2 refer to the two samples, and $s_1$, $s_2$, $n_1$, and $n_2$ represent the sample standard deviations and sizes for the respective samples. The t-statistic in both cases follows a t-distribution with degrees of freedom calculated using a more complex formula (Welch-Satterthwaite equation) that accounts for unequal variances.

**Paired Two-Sample Testing:**

For paired two-sample testing, where you compare two related samples, you calculate the **paired t-statistic**. The formula for the paired t-statistic is:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

Where:

$\bar{d}$ is the mean of the paired differences.
$s_d$ is the standard deviation of the paired differences.
$n$ is the number of paired observations.

The paired t-statistic follows a t-distribution with $n - 1$ degrees of freedom.

In all cases, once you calculate the respective t-statistic, you can compare it to critical values from the t-distribution or use p-values to assess the statistical significance of the observed differences. The smaller the p-value (typically below a chosen significance level, often 0.05), the stronger the evidence against the null hypothesis, indicating a significant difference.

# Why does t-distribution have (n-1) degree of freedom?

Imagine you have 4 numbers and the mean of them is 5.

a , b , c , d  mean is  5. so you must have 4 numbers that the sum of them is equal to 20.

Now I want to suggest these 4 numbers freely. for the first one I say 5

$$5 + b + c + d = 20$$

for next number i suggest 2

$$5 + 2 + c + d = 20$$

for the next number i suggest 0

$$5 + 2 + 0 + d = 20$$

now for the fourth number (d) I have not the freedom to suggest a number anymore, because the fourth one (d) must be 13.

so you have freedom to choose 3 of them minus 1 of them.

so n-1 is the degree of freedom for measuring the mean of a sample form a population.

# Testing for Mean

- **One-Sample Testing for Mean:** One-sample testing is used when you want to compare the mean of a single sample to a known or hypothesized population mean. The process involves comparing the sample mean to the hypothesized population mean and determining whether any observed differences are statistically significant. This is often done using a t-test.
- **Two-Sample Testing for Mean:** Two-sample testing is used to compare the means of two independent samples. The goal is to determine if the observed difference in means between the two samples is statistically significant. The two samples could come from different groups, treatments, or populations. A common example is comparing the exam scores of two different classes to see if there's a significant difference in their average scores.
- **Paired Two-Sample Testing for Mean:** Paired two-sample testing (also known as paired t-test) is used when you want to compare the means of two related or paired samples. The samples are not independent but rather represent measurements taken under different conditions on the same subjects or items. For instance, you might compare the before-and-after weights of individuals who underwent a specific treatment.

In summary:

- One-sample testing compares a sample mean to a known or hypothesized population mean.
- Two-sample testing compares the means of two independent samples.
- Paired two-sample testing compares the means of two related samples.

# Chi-Square test of independence

A **chi-square ($X^2$) test of independence** is a nonparametric hypothesis test. They're used to determine whether your data are significantly different from what you expected. You can use it to test whether two categorical variables are related to each other. If two variables are related, the probability of one variable having a certain value is dependent on the value of the other variable.

- The chi-square test of independence calculations are based on the observed frequencies, which are the numbers of observations in each combined group.
- The test compares the observed frequencies to the frequencies you would expect if the two variables are unrelated. When the variables are unrelated, the observed and expected frequencies will be similar.

**Example:** Imagine a city wants to encourage more of its residents to recycle their household waste.

The city decides to test two interventions: an educational flyer (pamphlet) or a phone call. They randomly select 300 households and randomly assign them to the flyer, phone call, or control group (no intervention). They'll use the results of their experiment to decide which intervention to use for the whole city.

The city plans to use a chi-square test of independence to test whether the proportion of households who recycle differs between the interventions.

# Chi-Square test of independence

## Contingency tables

When you want to perform a chi-square test of independence, the best way to organize your data is a type of **frequency distribution table** called a **contingency table**.

A contingency table, also known as a cross tabulation or crosstab, shows the number of observations in each combination of groups. It also usually includes row and column totals.

Example: Contingency table

Six months after the intervention, the city looks at the outcomes for the 300 households (only four households are shown here):

| Household address | Intervention | Outcome |
|---|---|---|
| 25 Elm Street | Flyer | Recycles |
| 100 Cedar Street | Control | Recycles |
| 3 Maple Street | Control | Does not recycle |
| 123 Oak Street | Phone call | Recycles |
| ... | ... | ... |

They reorganize the data into a contingency table:

| Intervention | Recycles | Does not recycle | Row totals |
|---|---|---|---|
| Flyer (pamphlet) | 89 | 9 | 98 |
| Phone call | 84 | 8 | 92 |
| Control | 86 | 24 | 110 |
| Column totals | 259 | 41 | $N = 300$ |

# Chi-Square test of independence

## Chi-square test of independence hypotheses

The chi-square test of independence is an inferential statistical test, meaning that it allows you to draw conclusions about a **population** based on a **sample**. Specifically, it allows you to conclude whether two variables are related in the population.

Like all hypothesis tests, the chi-square test of independence evaluates a null and alternative hypothesis. The hypotheses are two competing answers to the question "Are variable 1 and variable 2 related?"

- Null hypothesis ($H_0$): Variable 1 and variable 2 are **not related** in the population; The proportions of variable 1 are **the same** for different values of variable 2.
- Alternative hypothesis ($H_a$): Variable 1 and variable 2 are **related** in the population; The proportions of variable 1 are **not the same** for different values of variable 2.

The population is all households in the city.

Null hypothesis ($H_0$): Whether a household recycles and the type of intervention they receive are **not related** in the population; The proportion of households that recycle is **the same** for all interventions.

Alternative hypothesis ($H_a$): Whether a household recycles and the type of intervention they receive are **related** in the population; The proportion of households that recycle is **not the same** for all interventions.

# Chi-Square test of independence

## Expected values

- A chi-square test of independence works by comparing the observed and the expected frequencies. The expected frequencies are such that the proportions of one variable are the same for all values of the other variable.
- You can calculate the expected frequencies using the contingency table. The expected frequency for row *r* and column *c* is:

$$\frac{(\text{Row } r \text{ total} \times \text{Column } c \text{ total})}{N}$$

Example: Expected values

The city calculates the expected frequencies using the contingency table.

**Observed and expected frequencies (observed above, expected below)**

| Intervention | Recycles | Does not recycle | Row totals |
|---|---|---|---|
| Flyer (pamphlet) | 89 | 9 | 98 |
| | $\frac{(98 \times 259)}{300} = 84.61$ | $\frac{(98 \times 41)}{300} = 13.39$ | |
| Phone call | 84 | 8 | 92 |
| | $\frac{(92 \times 259)}{300} = 79.43$ | $\frac{(92 \times 41)}{300} = 12.57$ | |
| Control | 86 | 24 | 110 |
| | $\frac{(110 \times 259)}{300} = 94.97$ | $\frac{(110 \times 41)}{300} = 15.03$ | |
| Column totals | 259 | 41 | $N = 300$ |

# Chi-Square test of independence

## How to calculate the test statistic (formula)

Pearson's chi-square ($X^2$) is the test statistic for the chi-square test of independence:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

Where

- $X^2$ is the chi-square test statistic
- $\Sigma$ is the summation operator (it means "take the sum of")
- $O$ is the observed frequency
- $E$ is the expected frequency

The chi-square test statistic measures how much your observed frequencies differ from the frequencies you would expect if the two variables are unrelated. It is large when there's a big difference between the observed and expected frequencies ($O$ - $E$ in the equation).

Follow these five steps to calculate the test statistic:

### Step 1: Create a table

Create a table with the observed and expected frequencies in two columns.

Example: Step 1

| Intervention | Outcome | Observed | Expected |
|---|---|---|---|
| Flyer | Recycles | 89 | 84.61 |
| | Does not recycle | 9 | 13.39 |
| Phone call | Recycles | 84 | 79.43 |
| | Does not recycle | 8 | 12.57 |
| Control | Recycles | 86 | 94.97 |
| | Does not recycle | 24 | 15.03 |

# Chi-Square test of independence

## Step 2: Calculate $O - E$

In a new column called "$O - E$", subtract the expected frequencies from the observed frequencies.

Example: Step 2

| Intervention | Outcome | Observed | Expected | $O - E$ |
|---|---|---|---|---|
| Flyer | Recycles | 89 | 84.61 | 4.39 |
| | Does not recycle | 9 | 13.39 | -4.39 |
| Phone call | Recycles | 84 | 79.43 | 4.57 |
| | Does not recycle | 8 | 12.57 | -4.57 |
| Control | Recycles | 86 | 94.97 | -8.97 |
| | Does not recycle | 24 | 15.03 | 8.97 |

## Step 3: Calculate $(O - E)^2$

In a new column called "$(O - E)^2$", square the values in the previous column.

Example: Step 3

| Intervention | Outcome | Observed | Expected | $O - E$ | $(O - E)^2$ |
|---|---|---|---|---|---|
| Flyer | Recycles | 89 | 84.61 | 4.39 | 19.27 |
| | Does not recycle | 9 | 13.39 | -4.39 | 19.27 |
| Phone call | Recycles | 84 | 79.43 | 4.57 | 20.88 |
| | Does not recycle | 8 | 12.57 | -4.57 | 20.88 |
| Control | Recycles | 86 | 94.97 | -8.97 | 80.46 |
| | Does not recycle | 24 | 15.03 | 8.97 | 80.46 |

# Chi-Square test of independence

## Step 4: Calculate $(O - E)^2 / E$

In a final column called "$(O - E)^2 / E$", divide the previous column by the expected frequencies.

Example: Step 4

| Intervention | Outcome | Observed | Expected | $O - E$ | $(O - E)^2$ | $(O - E)^2 / E$ |
|---|---|---|---|---|---|---|
| flyer | Recycles | 89 | 84.61 | 4.39 | 19.27 | 0.23 |
| | Does not recycle | 9 | 13.39 | -4.39 | 19.27 | 1.44 |
| Phone call | Recycles | 84 | 79.43 | 4.57 | 20.88 | 0.26 |
| | Does not recycle | 8 | 12.57 | -4.57 | 20.88 | 1.66 |
| Control | Recycles | 86 | 94.97 | -8.97 | 80.46 | 0.85 |
| | Does not recycle | 24 | 15.03 | 8.97 | 80.46 | 5.35 |

## Step 5: Calculate $X^2$

Finally, add up the values of the previous column to calculate the chi-square test statistic (X2).

Example: Step 5

$X^2 = 0.23 + 1.44 + 0.26 + 1.66 + 0.85 + 5.35$

$X^2 = 9.79$

# Chi-Square test of independence

**How to perform the chi-square test of independence**

If the test statistic is big enough then you should conclude that the observed frequencies are not what you'd expect if the variables are unrelated. But what counts as big enough?

We compare the test statistic to a critical value from a chi-square distribution to decide whether it's big enough to reject the null hypothesis that the two variables are unrelated. This procedure is called the chi-square test of independence.

Follow these steps to perform a chi-square test of independence (the first two steps have already been completed for the recycling example):

## Step 1: Calculate the expected frequencies

Use the contingency table to calculate the expected frequencies following the formula:

$$\frac{(\text{Row } r \text{ total} \times \text{Column } c \text{ total})}{\text{Grand total}}$$

## Step 2: Calculate chi-square

Use the Pearson's chi-square formula to calculate the test statistic:

$$X^2 = \sum \frac{(O-E)^2}{E}$$

# Chi-Square test of independence

## Step 3: Find the critical chi-square value

You can find the critical value in a chi-square critical value table or using statistical software.

You need to known two numbers to find the critical value:

- **The degrees of freedom (*df*):** For a chi-square test of independence, the *df* is (number of variable 1 groups − 1) * (number of variable 2 groups − 1).
- **Significance level (α):** By convention, the significance level is usually .05.

Example: Finding the critical chi-square value

Since there are three intervention groups (flyer, phone call, and control) and two outcome groups (recycle and does not recycle) there are (3 − 1) * (2 − 1) = 2 degrees of freedom.

For a test of significance at α = .05 and *df* = 2, the $X^2$ critical value is 5.99.

## Step 4: Compare the chi-square value to the critical value

Is the test statistic big enough to reject the null hypothesis? Compare it to the critical value to find out.

Example: Comparing the chi-square value to the critical value

$X^2 = 9.79$

Critical value = 5.99

The $X^2$ value is greater than the critical value.

# Chi-Square test of independence

### Step 5: Decide whether to reject the null hypothesis

- If the $X^2$ value is **greater** than the critical value, then the difference between the observed and expected distributions is statistically significant ($p < \alpha$).
  - The data allows you to reject the null hypothesis that the variables are unrelated and provides support for the alternative hypothesis that the variables are related.
- If the $X^2$ value is **less** than the critical value, then the difference between the observed and expected distributions is not statistically significant ($p > \alpha$).
  - The data doesn't allow you to reject the null hypothesis that the variables are unrelated and doesn't provide support for the alternative hypothesis that the variables are related.

#### Example: Deciding whether to reject the null hypothesis

The $X^2$ value is greater than the critical value. Therefore, the city **rejects** the null hypothesis that whether a household recycles and the type of intervention they receive are **unrelated**.

There is a **significant difference** between the observed frequencies and the frequencies expected if the two variables were unrelated ($p < .05$). This suggests that the proportion of households that recycle is **not the same** for all interventions.

The city concludes that their interventions have an effect on whether households choose to recycle.

# Chi-Square Goodness of Fit Test

A **chi-square ($X^2$) goodness of fit test** is a type of Pearson's chi-square test. You can use it to test whether the observed distribution of a categorical variable differs from your expectations.

Example: Chi-square goodness of fit test

You're hired by a dog food company to help them test three new dog food flavors.

You recruit a random sample of 75 dogs and offer each dog a choice between the three flavors by placing bowls in front of them. You expect that the flavors will be equally popular among the dogs, with about 25 dogs choosing each flavor.

Once you have your experimental results, you plan to use a chi-square goodness of fit test to figure out whether the distribution of the dogs' flavor choices is significantly different from your expectations.

## What is the chi-square goodness of fit test?

A chi-square ($X^2$) goodness of fit test is a **goodness of fit** test for a categorical variable. Goodness of fit is a measure of how well a statistical model fits a set of observations.

- When goodness of fit is **high**, the values expected based on the model are **close to** the observed values.
- When goodness of fit is **low**, the values expected based on the model are **far from** the observed values.

The statistical models that are analyzed by chi-square goodness of fit tests are **distributions**. They can be any distribution, from as simple as equal probability for all groups, to as complex as a probability distribution with many parameters.

# Chi-Square Goodness of Fit Test

## Hypothesis testing

The chi-square goodness of fit test is a **hypothesis test**. It allows you to draw conclusions about the distribution of a population based on a sample. Using the chi-square goodness of fit test, you can test whether the goodness of fit is "good enough" to conclude that the population follows the distribution.

With the chi-square goodness of fit test, you can ask questions such as: Was this sample drawn from a population that has...

- Equal proportions of male and female turtles?
- Equal proportions of red, blue, yellow, green, and purple jelly beans?
- 90% right-handed and 10% left-handed people?
- Offspring with an equal probability of inheriting all possible genotypic combinations (i.e., unlinked genes)?
- A Poisson distribution of floods per year?
- A normal distribution of bread prices?

### Example: Observed and expected frequencies

After weeks of hard work, your dog food experiment is complete and you compile your data in a table:

**Observed and expected frequencies of dogs' flavor choices**

| Flavor | Observed | Expected |
|---|---|---|
| Garlic Blast | 22 | 25 |
| Blueberry Delight | 30 | 25 |
| Minty Munch | 23 | 25 |

# Chi-Square Goodness of Fit Test

## Chi-square goodness of fit test hypotheses

Like all hypothesis tests, a chi-square goodness of fit test evaluates two hypotheses: the null and alternative hypotheses. They're two competing answers to the question "Was the sample drawn from a population that follows the specified distribution?"

- **Null hypothesis ($H_0$):** The population follows the specified distribution.
- **Alternative hypothesis ($H_a$):** The population does not follow the specified distribution.

These are general hypotheses that apply to all chi-square goodness of fit tests. You should make your hypotheses more specific by describing the "specified distribution." You can name the probability distribution (e.g., Poisson distribution) or give the expected proportions of each

### Example: Null and alternative hypothesis

- **Null hypothesis ($H_0$):** The dog population chooses the three flavors in equal proportions ($p_1 = p_2 = p_3$).
- **Alternative hypothesis ($H_a$):** The dog population does not choose the three flavors in equal proportions.

# Chi-Square Goodness of Fit Test

## When to use the chi-square goodness of fit test

The following conditions are necessary if you want to perform a chi-square goodness of fit test:

1. You want to test a hypothesis about the distribution of **one categorical variable**. If your variable is continuous, you can convert it to a categorical variable by separating the observations into intervals. This process is known as data binning.
2. The **sample was randomly selected** from the population.
3. There are a **minimum of five observations expected** in each group.

### Example: Chi-square goodness of fit test conditions

You can use a chi-square goodness of fit test to analyze the dog food data because all three conditions have been met:

1. You want to test a hypothesis about the distribution of one categorical variable. The categorical variable is the dog food flavors.
2. You recruited a random sample of 75 dogs.
3. There were a minimum of five observations expected in each group. For all three dog food flavors, you expected 25 observations of dogs choosing the flavor.

# Chi-Square Goodness of Fit Test

Find the critical chi-square value in a chi-square critical value table or using statistical software.

The critical value is calculated from a chi-square distribution. To find the critical chi-square value, you'll need to know two things:

- **The degrees of freedom (*df*):** For chi-square goodness of fit tests, the *df* is the number of groups minus one.
- **Significance level (α):** By convention, the significance level is usually .05.

Example: Finding the critical chi-square value

Since there are three groups (Garlic Blast, Blueberry Delight, and Minty Munch), there are two degrees of freedom.

For a test of significance at α = .05 and *df* = 2, the $X^2$ critical value is 5.99.

Add up the values of the previous column. This is the chi-square test statistic ($X^2$).

Example: Step 5

| Flavor | Observed | Expected | $O - E$ | $(O - E)^2$ | $(O - E)^2 / E$ |
|---|---|---|---|---|---|
| Garlic Blast | 22 | 25 | -3 | 9 | 9/25 = 0.36 |
| Blueberry Delight | 30 | 25 | 5 | 25 | 1 |
| Minty Munch | 23 | 25 | -2 | 4 | 0.16 |

$X^2 = 0.36 + 1 + 0.16 = 1.52$

# Chi-Square Goodness of Fit Test

- If the $X^2$ value is **greater** than the critical value, then the difference between the observed and expected distributions is statistically significant ($p < \alpha$).
  - The data allows you to reject the null hypothesis and provides support for the alternative hypothesis.
- If the $X^2$ value is **less** than the critical value, then the difference between the observed and expected distributions is not statistically significant ($p > \alpha$).
  - The data doesn't allow you to reject the null hypothesis and doesn't provide support for the alternative hypothesis.

**Example: Deciding whether to reject the null hypothesis**

The $X^2$ value is less than the critical value. Therefore, you **should not reject** the null hypothesis that the dog population chooses the three flavors in equal proportions. There is no significant difference between the observed and expected flavor choice distribution ($p > .05$). This suggests that the dog food flavors are equally popular in the dog population.

Compare the chi-square value to the critical value to determine which is larger.

**Example: Comparing the chi-square value to the critical value**

$X^2 = 1.52$

Critical value = 5.99

The $X^2$ value is less than the critical value.