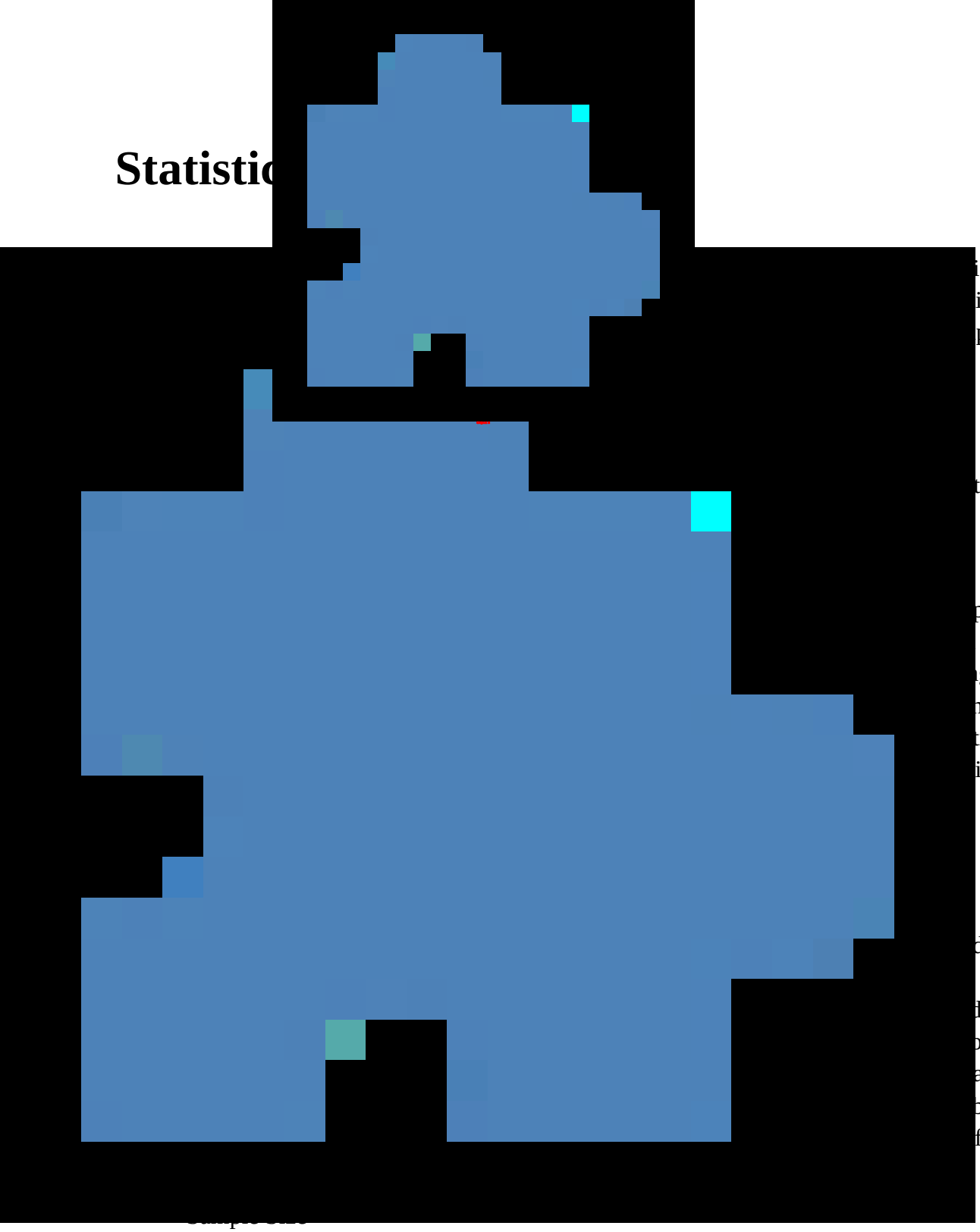
Statistic



ion, and the
itative data.
ple about a

tics are:

p you make
allows you
g inference
n. To take a
ticle, one of
ing to learn

data subject
confidence
d of making
o assess the
al inference
ble range of
for making

- Variability in the sample
- Size of the observed differences

Types of Statistical Inference

There are different types of statistical inferences that are extensively used for making conclusions. They are:

- One sample hypothesis testing
- Confidence Interval
- Pearson Correlation
- Bi-variate regression

- Multi-variate regression
- Chi-square statistics and contingency table
- ANOVA or T-test

Statistical Inference Procedure

The procedure involved in inferential statistics are:

- Begin with a theory
- Create a research hypothesis
- Operationalize the variables
- Recognize the population to which the study results should apply
- Formulate a null hypothesis for this population
- Accumulate a sample from the population and continue the study
- Conduct statistical tests to see if the collected sample properties are adequately different from what would be expected under the null hypothesis to be able to reject the null hypothesis

Statistical Inference Solution

Statistical inference solutions produce efficient use of statistical data relating to groups of individuals or trials. It deals with all characters, including the collection, investigation and analysis of data and organizing the collected data. By statistical inference solution, people can acquire knowledge after starting their work in diverse fields. Some statistical inference solution facts are:

- It is a common way to assume that the observed sample is of independent observations from a population type like Poisson or normal
- Statistical inference solution is used to evaluate the parameter(s) of the expected model like normal mean or binomial proportion

Importance of Statistical Inference

Inferential Statistics is important to examine the data properly. To make an accurate conclusion, proper data analysis is important to interpret the research results. It is majorly used in the future prediction for various observations in different fields. It helps us to make inference about the data. The statistical inference has a wide range of application in different fields, such as:

- Business Analysis
- Artificial Intelligence
- Financial Analysis
- Fraud Detection
- Machine Learning
- Share Market
- Pharmaceutical Sector

Statistical Inference Examples

An example of statistical inference is given below.

Question: From the shuffled pack of cards, a card is drawn. This trial is repeated for 400 times, and the suits are given below:

Suit	Spade	Clubs	Hearts	Diamonds
No.of times drawn	90	100	120	90

While a card is tried at random, then what is the probability of getting a

1. Diamond cards
2. Black cards
3. Except for spade

Solution:

By statistical inference solution,

Total number of events = 400

i.e.,90+100+120+90=400

(1) The probability of getting diamond cards:

Number of trials in which diamond card is drawn = 90

Therefore, P(diamond card) = 90/400 = 0.225

(2) The probability of getting black cards:

Number of trials in which black card showed up = 90+100 =190

Therefore, P(black card) = 190/400 = 0.475

(3) Except for spade

Number of trials other than spade showed up = 90+100+120 =310

Therefore, P(except spade) = 310/400 = 0.775

Stay tuned with BYJU'S – The Learning App for more Maths-related concepts and download the app for more personalized videos.

What Is The Statistical Inference?

Data scientists usually spend a large amount of time to gather and assess data. The data is then used to deduce conclusions using data analysis techniques.

Sometimes these conclusions are observed and the findings are easily described using charts and tables. This is known as descriptive statistics. Other times, we have to explore a measure that is unobserved. This is where the statistical inference comes in.

So far so good. Let’s now understand it

The descriptive statistical inference essentially describes the data to the users but it does not make any inferential from the data. Inferential statistics is the other branch of statistical inference. The sample is very unlikely to be an absolute true representation of the population and as a result, we always have a level of uncertainty when drawing conclusions about the population.

As an instance, the data scientists might aim to understand how a variable in their experiment behaves. Gathering all of the data (population) for that variable might be a humongous task. Data scientists, therefore, take a small sample of the population of their target variable to represent the population, and then they perform statistical inference on the small sample(s).

> The samples are used to estimate the population

The aim of the data scientists is to generalise from a sample to a population knowing there is a degree of uncertainty. Hence the analyses help them make propositions about the entire population of the data. Sometimes data scientists simulate the samples to understand how the population behaves and for that they make assumptions about the underlying probability distributions of the variable. This is one of the core reasons why the concept of probability is heavily recommended to the data scientists.

Subsequently, a number of hypotheses and claims are made about the properties of the population. Next, the statistical models are used to infer conclusions from the sample to deduce the properties of the population.

The article below provides a thorough understanding of what probability distributions are and I highly recommend everyone to read the article

2. Understanding Statistical Inference Process

This section will help us understand the process of statistical inference. Let’s assume that the data scientists want to learn about the behaviour of their target variables. They might be interested in understanding how a parameter of a population behaves.

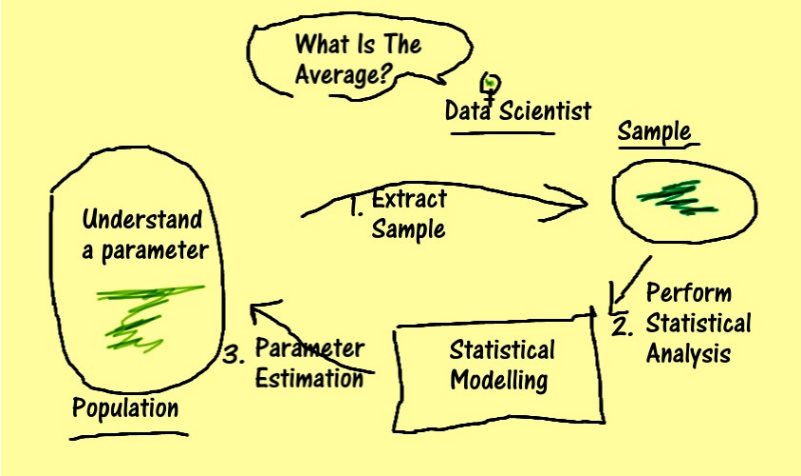
- As an instance, they might want to assess whether all overnight batch jobs across all of the departments in a bank complete within a particular time frame.
- Or, they might want to find the average height of a population in a country.
- Or, maybe they want to understand whether a business made the same profit and the users behaved differently before or after a specific event, such as after a new product was launched.
- Or they want to prove a particular claim about a population wrong.

Occasionally it is too difficult to gather all of the data of a population. Consequently, the data scientists prepare their sample set from the population.

For instance, the parameter the data scientists want to learn about could be the mean or variance of the population. They extract the sample from the population and they can then perform statistical

analysis to estimate the population parameter. Sometimes, they check whether the parameter meets a specific value which is believed to be true.

The diagram below illustrates the process:



> Notice how the statistical inference is as good as the sampling techniques

There is always some element of noise in the sample. **The standard deviation of the sample over the square root of the sample size is the sample standard error**. This is the noise/dispersion of the sample from the mean. This measure is based on the sample size. The formula indicates that the larger the sample size, the lower the impact of standard deviation, and the closer the sample value is to the population value. If we increase the sample size then eventually the sample will start resembling the population. I recommend reading this article that explains what convergence and CLT are. These are crucial topics for data scientists to understand:

2.1 Think Through This Example

As an instance, we can make a claim that our target variable follows a normal distribution and its mean is always zero, and variance is 1 and so on.

So, after gathering a small sample, we start plotting the values within 10 bins in a histogram. From the chart, we conclude that the data follows a normal distribution because of the bell-shaped curve that is produced. And from there, we can start estimating the mean and variance of the sample to draw inference about the population. We can also start producing more data from the believed probability distribution of the sample.

But how confident are we? How can we reject this claim? Or is there a way to calculate this claim so that we can be sure? Can we prove it via quantifiable measures? This is where statistical inference comes in!

> It’s vital to understand statistical inference because it can help us understand our statistical choices better

Victoria Bilsborough on Unsplash

A sample can be thought of as a random variable having its own probability distribution, patterns, and trends.

We can collect a number of samples and workout their means, standard deviation, and variances to gain better insight into the data.

The process of test statistics can be used to help us make calculated decisions. Before performing an experiment on the sample, scientists have an idea of what the expected results need to be. This is usually gathered by exploratory data analysis.

3. Test Statistics — Bigger Picture With An Example

Now, from the theory, let's review how statistical inference works.

Let's imagine that we are working on a data science project and our client is a large financial institution. Let me refer to our client as Bank A. They want us to understand whether their overnight batch systems on average take the same amount of average time as other banks in the industry. The client wants to use our results to understand whether they need to invest differently in technology this year.

Let's also assume that it is widely believed in the industry that the overnight batch job times in all of the banks on average take ≤ 6 hours to complete.

This is a big data problem. We can't go around every bank and get hold of all of their batch job data. This will be an extremely time-consuming and costly job. We can imagine how much data our population holds. Therefore, the first step would be to gather a good sample set across different banks.

For our example, we can categorise the banks into groups based on the size of their transactions and take one-year worth of the batch job timings from a member of each group.

> The sample is chosen such that it is the best representation of the population of data under test. Always spend time thinking through the sampling techniques
>
> Success of hypothesis analysis is based on the quality of the chosen sample.

We then make a claim that on average the overnight batch jobs for Bank A take longer than 6 hours to complete. Now, this is our hypothesis under test. Let's prove it right.

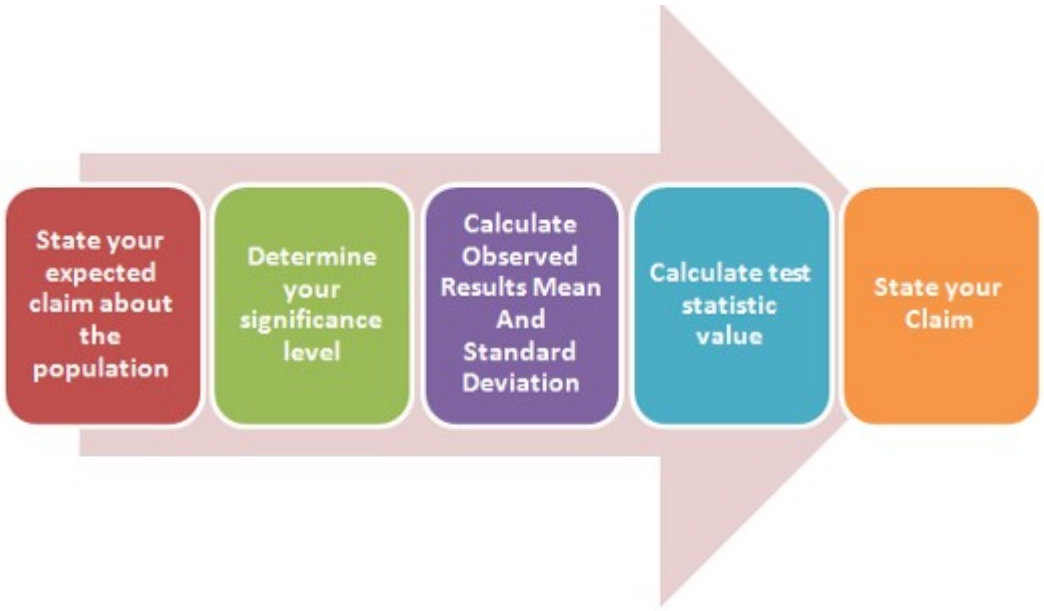
Let's hold the thought while I cover the foundation and then I will get back to it.

This brings me to the second sub-topic of this article — Test Statistics

We perform the following 5 steps to prove a claim:

1. Start by stating the widely believed claim, which is known as the Null Hypothesis.

2. Outline the minimum significance level/confidence level before we can reject the claim. This could be 5%. It means that we believe that around 5% of the time, our model will produce inaccurate results.
3. Calculate our sample results mean and standard deviation
4. Calculate the test statistics
5. Finally, based on the outcome, the chosen result is stated.



> *The chosen test statistic is dependent on the distribution of the sample along with the sample size.*

The next section will explain each of the steps in detail.

4. Hypothesis Testing

1. State Your Claim

There are two hypotheses in any test:

1. **Null Hypothesis—**what is known as the accepted truth and what we want to test. This is what we want to prove wrong. *For our case, it is that the average (mean) of batch jobs take less than or equal to 6 hours to complete.*
2. **Alternate Hypothesis** — what we need to accept if the Null Hypothesis is not true. This is what we believe is true. This is our hypothesis about the bank. *Our hypothesis is that the mean of the jobs is greater than 6 hours*.

Note, the null and alternative hypothesis cannot be true at the same time.

Side note*: This is a one-tail test. One Tail Alternative Hypothesis is a uni-directional test. The two tail alternative hypothesis tests are bidirectional tests whereby a statistician is interested in checking equality of data e.g. Whether a value is within a range of values.*

2. Determine your significance level:

Your significance level indicates how confident you are about the sample and your methodology to support your claim. The significance level is known as Alpha. The usual value for alpha is 1% or 5%. Lower alpha implies that you are very certain about the results. The chosen confidence level forms the foundation of risk management credit metrics.

> **Alpha** is the level of significance in Hypothesis analysis.
> To elaborate, Alpha is the range of values that can be **accepted** before the Null Hypothesis is rejected.
>
> *It is the lower threshold.*

3. Calculating Test Statistics

High-level Test Statistics Overview

We can choose T, Z, or F statistics amongst others. I will briefly explain them.

Once a sample is chosen to represent a population, its mean and standard deviation are calculated. And then we perform the test statistic.

There are a number of test statistics e.g. T, Z, F, etc.

I highly recommend reading this article as it explains the three most important statistics in-depth:

T statistics is used for testing the equality of means of two small populations. The sample follows Student T Distribution and the sample size is around 30 observations. The population standard deviation is unknown.

- **Example**: You have a sample of 10 cars and you want to measure the average fuel consumption of all cars in the town. Your hypothesized claim is that on average, cars consume 10 liters of fuel per day. Let's also consider that you are 99% confident in the methodology. You can then compare the hypothesized mean with the sample mean and work out if you need to reject the Null Hypothesis based on the T distribution table at 99%.

Z statistics is used for testing means of two large populations. The sample follows Normal Distribution and the sample size is usually greater than 30. The population standard deviation is known.

- **Example**: Assume you have collected a sample of 500 individuals to estimate the average number of people wearing blue shirts on a daily basis. Let's also consider that you are 95% confident in your model. You can then compare the hypothesized mean with the sample mean and work out if you need to reject the Null Hypothesis based on the Z distribution table at 95%.

F statistics is used for comparing the variances of two populations. Variation is the sum of the squared deviations of each observation from its group mean divided by the error degree of freedom.

- **Example**: You can use an F test to compare the variability of software bugs in two IT systems in your company.

Each of these test-statistics has its own straightforward formulae. For the sake of simplicity, I am not covering these formulae in this article.

When the results are calculated, they are checked against the value from the distribution table. Note, the results are dependent on the sample size, standard deviation, and the mean of the sample.

We can perform test statistics using Python. As an instance, to perform T-Test, we can do:

```
from scipy import stats
stats.ttest_ind(collection_one, collection_two)
```

From the results of the calculation, we can state the result.

5. State Your Result:

Let’s assume we computed 0.50 from the test statistics calculation. We can now look up the value in the probability distribution table for our sample for our chosen alpha of 95%. The Z-distribution table gives us a value of 1.96.
As $0.50 \leq 1.96$, we have to accept the Null Hypothesis.

This means that the batch jobs do complete within the claimed 6 hours time and our claim was wrong. That’s how statistical inference works.

5. Types Of Errors:

Lastly, I wanted to cover the are two types of errors. It is important to understand how and where the experiments can go wrong.

Type 1 And Type 2

There are two types of errors which can occur while we state the results:

1. **Type 1 error:** Null Hypothesis was correct but the analysis proved it wrong.
2. **Type 2 error:** Null Hypothesis was wrong but the analysis couldn’t prove that it was wrong

Statistics is a collection of tools that you can use to get answers to important questions about data.

You can use descriptive statistical methods to transform raw observations into information that you can understand and share. You can use inferential statistical methods to reason from small samples of data to whole domains.

Statistics is Required Prerequisite

Machine learning and statistics are two tightly related fields of study. So much so that statisticians refer to machine learning as "*applied statistics*" or "*statistical learning*" rather than the computer-science-centric name.

Machine learning is almost universally presented to beginners assuming that the reader has some background in statistics

Why Learn Statistics?

Raw observations alone are data, but they are not information or knowledge.

Data raises questions, such as:

- What is the most common or expected observation?
- What are the limits on the observations?
- What does the data look like?

Although they appear simple, these questions must be answered in order to turn raw observations into information that we can use and share.

Beyond raw data, we may design experiments in order to collect observations. From these experimental results we may have more sophisticated questions, such as:

- What variables are most relevant?
- What is the difference in an outcome between two experiments?
- Are the differences real or the result of noise in the data?

Questions of this type are important. The results matter to the project, to stakeholders, and to effective decision making.

Statistical methods are required to find answers to the questions that we have about data.

We can see that in order to both understand the data used to train a machine learning model and to interpret the results of testing different machine learning models, that statistical methods are required.

What is Statistics?

Statistics is a subfield of mathematics.

It refers to a collection of methods for working with data and using data to answer questions.

> Statistics is the art of making numerical conjectures about puzzling questions. […] The methods were developed over several hundred years by people who were looking for answers to their questions.

(— Page xiii, Statistics, Fourth Edition, 2007.)

It is because the field is comprised of a grab bag of methods for working with data that it can seem large and amorphous to beginners. It can be hard to see the line between methods that belong to

statistics and methods that belong to other fields of study. Often a technique can be both a classical method from statistics and a modern algorithm used for feature selection or modeling.

Although a working knowledge of statistics does not require deep theoretical knowledge, some important and easy-to-digest theorems from the relationship between statistics and probability can provide a valuable foundation.

Two examples include the law of large numbers and the central limit theorem; the first aids in understanding why bigger samples are often better and the second provides a foundation for how we can compare the expected values between samples (e.g mean values).

When it comes to the statistical tools that we use in practice, it can be helpful to divide the field of statistics into two large groups of methods: descriptive statistics for summarizing data and inferential statistics for drawing conclusions from samples of data.

> Statistics allow researchers to collect information, or data, from a large number of people and then summarize their typical experience. […] Statistics are also used to reach conclusions about general differences between groups. […] Statistics can also be used to see if scores on two variables are related and to make predictions.

(Pages ix-x, Statistics in Plain English, Third Edition, 2010.)

What is Statistics?

Statistics is a subfield of mathematics.

It refers to a collection of methods for working with data and using data to answer questions.

> Statistics is the art of making numerical conjectures about puzzling questions. […] The methods were developed over several hundred years by people who were looking for answers to their questions.

(Page xiii, Statistics, Fourth Edition, 2007.)

It is because the field is comprised of a grab bag of methods for working with data that it can seem large and amorphous to beginners. It can be hard to see the line between methods that belong to statistics and methods that belong to other fields of study. Often a technique can be both a classical method from statistics and a modern algorithm used for feature selection or modeling.

Although a working knowledge of statistics does not require deep theoretical knowledge, some important and easy-to-digest theorems from the relationship between statistics and probability can provide a valuable foundation.

Two examples include the law of large numbers and the central limit theorem; the first aids in understanding why bigger samples are often better and the second provides a foundation for how we can compare the expected values between samples (e.g mean values).

When it comes to the statistical tools that we use in practice, it can be helpful to divide the field of statistics into two large groups of methods: descriptive statistics for summarizing data and inferential statistics for drawing conclusions from samples of data.

> Statistics allow researchers to collect information, or data, from a large number of people and then summarize their typical experience. […] Statistics are also used to

> reach conclusions about general differences between groups. […] Statistics can also be used to see if scores on two variables are related and to make predictions.

(Pages ix-x, Statistics in Plain English, Third Edition, 2010.)

Measures of Dispersion

Suppose you are given a data series. Someone asks you to tell some interesting facts about this data series. How can you do so? You can say you can find the mean, the median or the mode of this data series and tell about its distribution. But is it the only thing you can do? Are the central tendencies the only way by which we can get to know about the concentration of the observation? In this section, we will learn about another measure to know more about the data. Here, we are going to know about the measure of dispersion.

As the name suggests, the measure of dispersion shows the scatterings of the data. It tells the variation of the data from one another and gives a clear idea about the distribution of the data. The measure of dispersion shows the homogeneity or the heterogeneity of the distribution of the observations.

Suppose you have four datasets of the same size and the mean is also same, say, m. In all the cases the sum of the observations will be the same. Here, the measure of central tendency is not giving a clear and complete idea about the distribution for the four given sets.
Can we get an idea about the distribution if we get to know about the dispersion of the observations from one another within and between the datasets? The main idea about the measure of dispersion is to get to know how the data are spread. It shows how much the data vary from their average value.

Characteristics of Measures of Dispersion

- A measure of dispersion should be rigidly defined
- It must be easy to calculate and understand
- Not affected much by the fluctuations of observations
- Based on all observations

Classification of Measures of Dispersion

The measure of dispersion is categorized as:

(i) An absolute measure of dispersion:

- The measures which express the scattering of observation in terms of distances i.e., range, quartile deviation.
- The measure which expresses the variations in terms of the average of deviations of observations like mean deviation and standard deviation.

(ii) A relative measure of dispersion:

We use a relative measure of dispersion for comparing distributions of two or more data set and for unit free comparison. They are the coefficient of range, the coefficient of mean deviation, the

coefficient of quartile deviation, the coefficient of variation, and the coefficient of standard deviation.

Range

A range is the most common and easily understandable measure of dispersion. It is the difference between two extreme observations of the data set. If X_{max} and X_{min} are the two extreme observations then

$$\text{Range} = X_{max} - X_{min}$$

Merits of Range

- It is the simplest of the measure of dispersion
- Easy to calculate
- Easy to understand
- Independent of change of origin

Demerits of Range

- It is based on two extreme observations. Hence, get affected by fluctuations
- A range is not a reliable measure of dispersion
- Dependent on change of scale

Quartile Deviation

The quartiles divide a data set into quarters. The first quartile, (Q_1) is the middle number between the smallest number and the median of the data. The second quartile, (Q_2) is the median of the data set. The third quartile, (Q_3) is the middle number between the median and the largest number.

Quartile deviation or semi-inter-quartile deviation is

$$Q = ½ \times (Q_3 - Q1)$$

Merits of Quartile Deviation

- All the drawbacks of Range are overcome by quartile deviation
- It uses half of the data
- Independent of change of origin
- The best measure of dispersion for open-end classification

Demerits of Quartile Deviation

- It ignores 50% of the data
- Dependent on change of scale
- Not a reliable measure of dispersion

Mean Deviation

Mean deviation is the arithmetic mean of the absolute deviations of the observations from a measure of central tendency. If $x_1, x_2, \dots, x_n$ are the set of observation, then the mean deviation of x about the average A (mean, median, or mode) is

Mean deviation from average $A = 1/n\ [\sum_i |x_i - A|]$

For a grouped frequency, it is calculated as:

Mean deviation from average $A = 1/N\ [\sum_i\ f_i\ |x_i - A|]$, $N = \sum f_i$

Here, x_i and f_i are respectively the mid value and the frequency of the i^{th} class interval.

Merits of Mean Deviation

- Based on all observations
- It provides a minimum value when the deviations are taken from the median
- Independent of change of origin

Demerits of Mean Deviation

- Not easily understandable
- Its calculation is not easy and time-consuming
- Dependent on the change of scale
- Ignorance of negative sign creates artificiality and becomes useless for further mathematical treatment

Standard Deviation

A standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by a Greek letter sigma, σ. It is also referred to as root mean square deviation. The standard deviation is given as

$$\sigma = [(\Sigma_i\ (y_i - \bar{y})/n]^{½} = [(\Sigma_i\ y_i^{\,2}/n) - \bar{y}^{\,2}]^{½}$$

For a grouped frequency distribution, it is

$$\sigma = [(\Sigma_i\ f_i\ (y_i - \bar{y})/N]^{½} = [(\Sigma_i\ f_i\ y_i^{\,2}/n) - \bar{y}^{\,2}]^{½}$$

The square of the standard deviation is the **variance**. It is also a measure of dispersion.

$$\sigma^2 = [(\Sigma_i\ (y_i - \bar{y}\ )\ /\ n]^{½} = [(\Sigma_i\ y_i^{\,2}/n) - \bar{y}^{\,2}]$$

For a grouped frequency distribution, it is

$$\sigma^2 = [(\Sigma_i\ f_i\ (y_i - \bar{y}\ )/N]^{½} = [(\Sigma_i\ f_i\ x_i^{\,2}/n) - \bar{y}^{\,2}].$$

If instead of a mean, we choose any other arbitrary number, say A, the standard deviation becomes the root mean deviation.

Variance of the Combined Series

If σ_1, σ_2 are two standard deviations of two series of sizes n_1 and n_2 with means $\bar{y}_1$ and $\bar{y}_2$. The variance of the two series of sizes $n_1 + n_2$ is:

$\sigma^2 = (1/ n_1 + n_2) \div [n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2)]$

where, $d_1 = \bar{y}_1 - \bar{y}$, $d_2 = \bar{y}_2 - \bar{y}$, and $\bar{y} = (n_1 \bar{y}_1 + n_2 \bar{y}_2) \div (n_1 + n_2)$.

Merits of Standard Deviation

- Squaring the deviations overcomes the drawback of ignoring signs in mean deviations
- Suitable for further mathematical treatment
- Least affected by the fluctuation of the observations
- The standard deviation is zero if all the observations are constant
- Independent of change of origin

Demerits of Standard Deviation

- Not easy to calculate
- Difficult to understand for a layman
- Dependent on the change of scale

Coefficient of Dispersion

Whenever we want to compare the variability of the two series which differ widely in their averages. Also, when the unit of measurement is different. We need to calculate the coefficients of dispersion along with the measure of dispersion. The coefficients of dispersion (C.D.) based on different measures of dispersion are

- Based on Range = $(X_{max} – X_{min})/(X_{max} + X_{min})$.
- C.D. based on quartile deviation = $(Q_3 – Q1)/(Q_3 + Q1)$.
- Based on mean deviation = Mean deviation/average from which it is calculated.
- For Standard deviation = S.D./Mean

Coefficient of Variation

100 times the coefficient of dispersion based on standard deviation is the coefficient of variation (C.V.).

C.V. = 100 × (S.D. / Mean) = $(\sigma/\bar{y}) \times 100$.

Solved Example on Measures of Dispersion

Problem: Below is the table showing the values of the results for two companies A, and B.

	Company A	Company B
Number of employees	900	1000
Average daily wage	Rs. 250	Rs. 220
Variance in the distribution of wages	100	144

1. Which of the company has a larger wage bill?
2. Calculate the coefficients of variations for both of the companies.
3. Calculate the average daily wage and the variance of the distribution of wages of all the employees in the firms A and B taken together.

Solution:

For Company A

No. of employees = $n_1 = 900$, and average daily wages = $\bar{y}_1 = \text{Rs. } 250$

We know, average daily wage = Total wages ⁄ Total number of employees

or, Total wages = Total employees × average daily wage = 900 × 250 = Rs. 225000 … (i)

For Company B

No. of employees = $n_2 = 1000$, and average daily wages = $\bar{y}_2 = \text{Rs. } 220$

So, Total wages = Total employees × average daily wage = 1000 × 220 = Rs. 220000 … (ii)

Comparing (i), and (ii), we see that Company A has a larger wage bill.

For Company A

Variance of distribution of wages = $\sigma_1^2 = 100$

C.V. of distribution of wages = 100 x standard deviation of distribution of wages/ average daily wages

Or, C.V. $_A = 100 \times \sqrt{100} / 250 = 100 \times 10 / 250 = 4$ … (i)

For Company B

Variance of distribution of wages = $\sigma_2^2 = 144$

C.V. $_B = 100 \times \sqrt{144} / 220 = 100 \times 12 / 220 = 5.45$ … (ii)

Comparing (i), and (ii), we see that Company B has greater variability.

For Company A and B, taken together

The average daily wages for both the companies taken together

$\bar{y} = (n_1 \bar{y}_1 + n_2 \bar{y}_2) / (n_1 + n_2) = (900 \times 250 + 1000 \times 220) \div (900 + 1000) = 445000 / 1900 = \text{Rs. } 234.21$

The combined variance, $\sigma^2 = (1/ n_1 + n_2) \div [n_1 (\sigma_1{}^2 + d_1{}^2) + n_2 (\sigma_2{}^2 + d_2{}^2)]$

Here, $d_1 = \bar{y}_1 - \bar{y} = 250 - 234.21 = 15.79$, $d_2 = \bar{y}_2 - \bar{y} = 220 - 234.21 = -14.21$.

Hence, $\sigma^2 = [900 \times (100 + 15.79^2) + 1000 \times (144 + -14.21^2)] / (900 + 1000)$

or, $\sigma^2 = (314391.69 + 345924.10) / 1900 = 347.53$.

Measures of Central Tendency: Mean, Median, and Mode

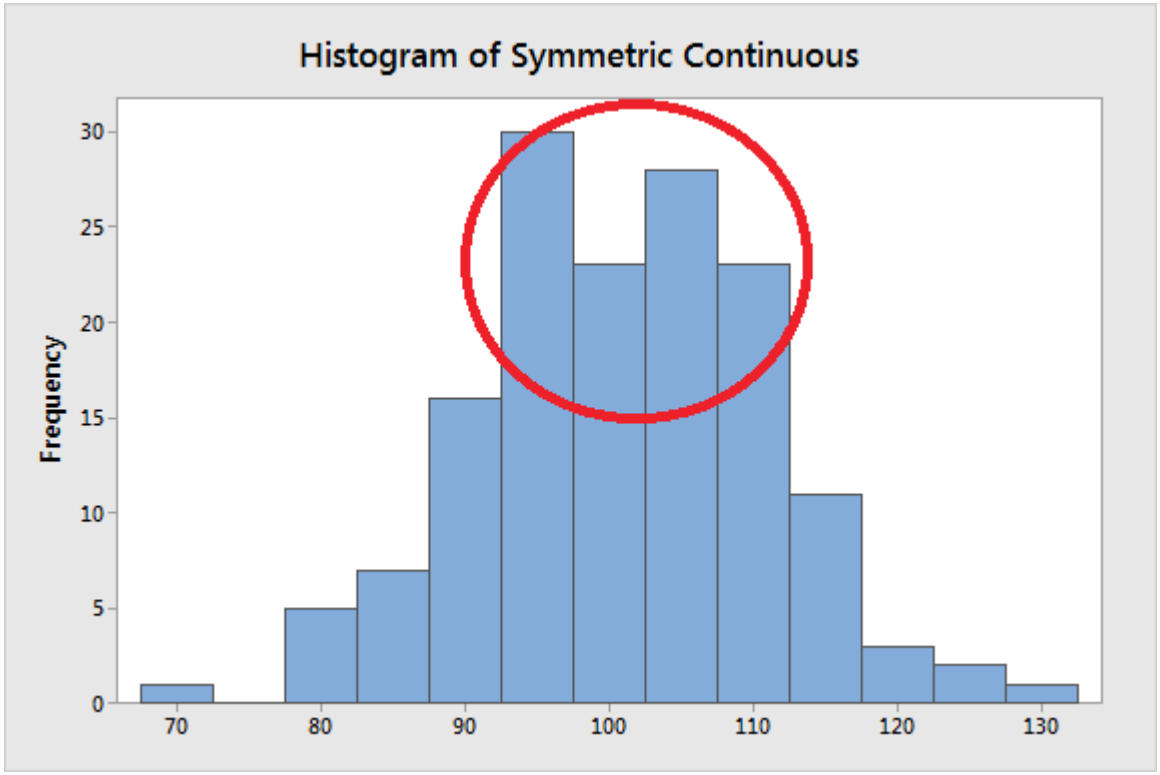
A measure of central tendency is a summary statistic that represents the center point or typical value of a dataset. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution. You can think of it as the tendency of data to cluster around a middle value. In statistics, the three most common measures of central tendency are the mean, median, and mode. Each of these measures calculates the location of the central point using a different method.

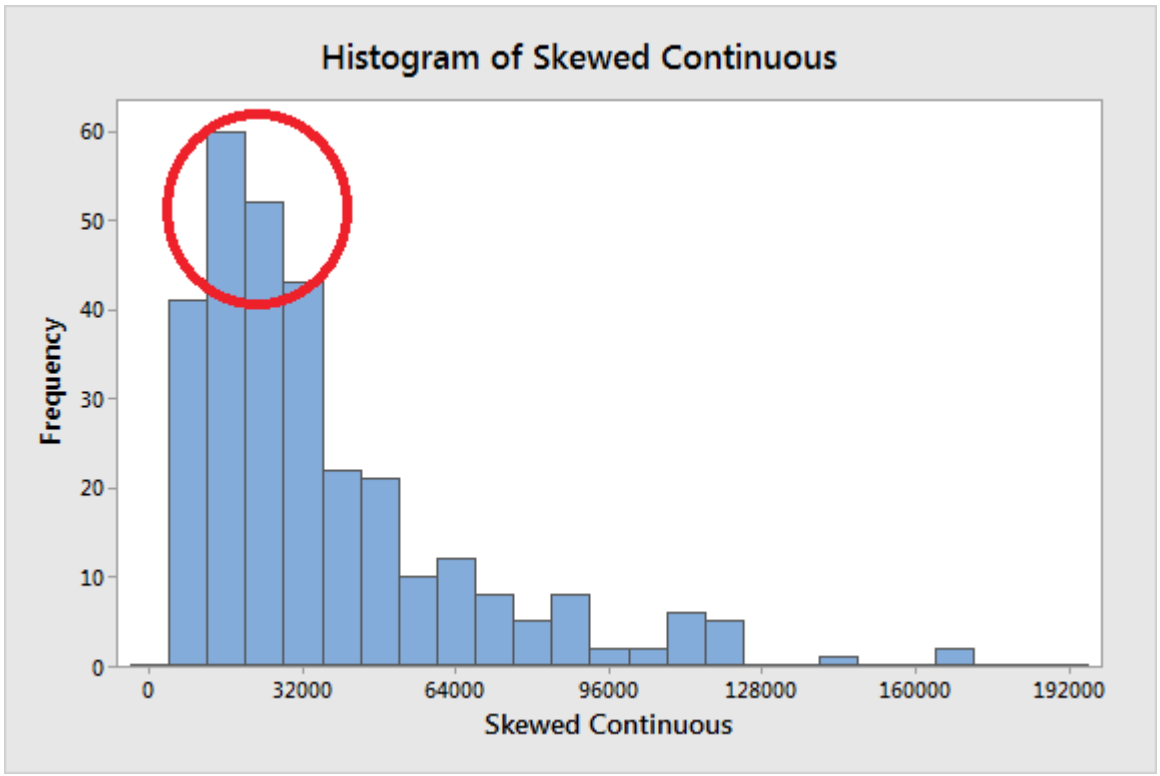
Choosing the best measure of central tendency depends on the type of data you have. In this post, I explore these measures of central tendency, show you how to calculate them, and how to determine which one is best for your data.

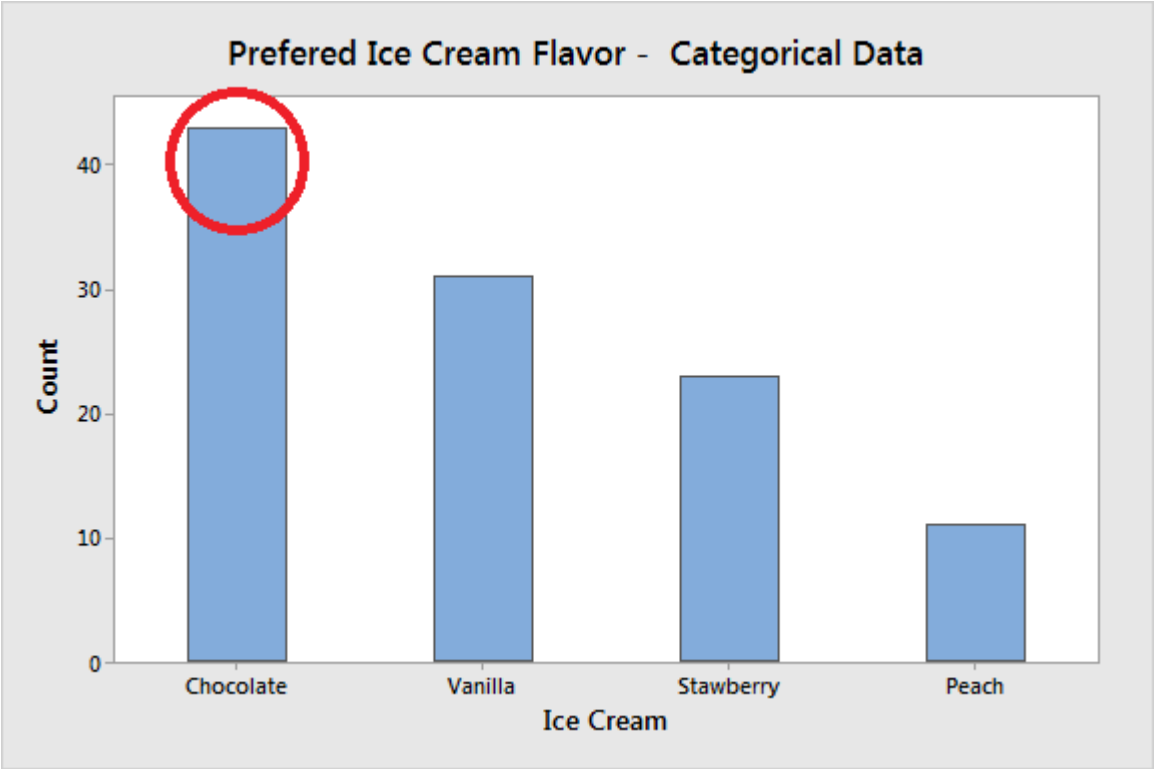
Locating the Center of Your Data

Most articles that you'll read about the mean, median, and mode focus on how you calculate each one. I'm going to take a slightly different approach to start out. My philosophy throughout my blog is to help you intuitively grasp statistics by focusing on concepts. Consequently, I'm going to start by illustrating the central point of several datasets graphically—so you understand the goal. Then, we'll move on to choosing the best measure of central tendency for your data and the calculations.

The three distributions below represent different data conditions. In each distribution, look for the region where the most common values fall. Even though the shapes and type of data are different, you can find that central location. That's the area in the distribution where the most common values are located.

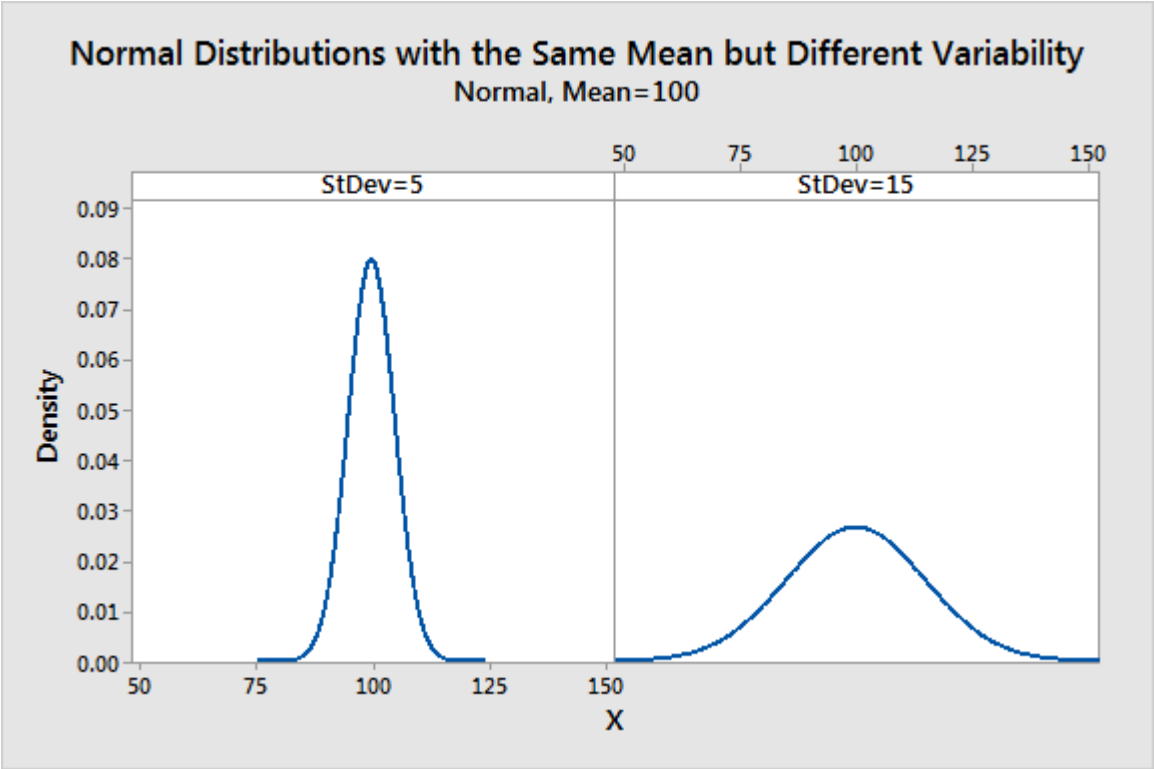

Histogram of Symmetric Continuous
Frequency
0
5
10
15
20
25
30
70
80
90
100
110
120
130


Histogram of Skewed Continuous
Frequency
0
10
20
30
40
50
60
0
32000
64000
96000
128000
160000
192000
Skewed Continuous



As the graphs highlight, you can see where most values tend to occur. That's the concept. Measures of central tendency represent this idea with a value. Coming up, you'll learn that as the distribution and kind of data changes, so does the best measure of central tendency. Consequently, you need to know the type of data you have, and graph it, before choosing a measure of central tendency!

The central tendency of a distribution represents one characteristic of a distribution. Another aspect is the variability around that central value. While measures of variability is the topic of a different article (link below), this property describes how far away the data points tend to fall from the center. The graph below shows how distributions with the same central tendency (mean = 100) can actually be quite different. The panel on the left displays a distribution that is tightly clustered around the mean, while the distribution on the right is more spread out. It is crucial to understand that the central tendency summarizes only one aspect of a distribution and that it provides an incomplete picture by itself.

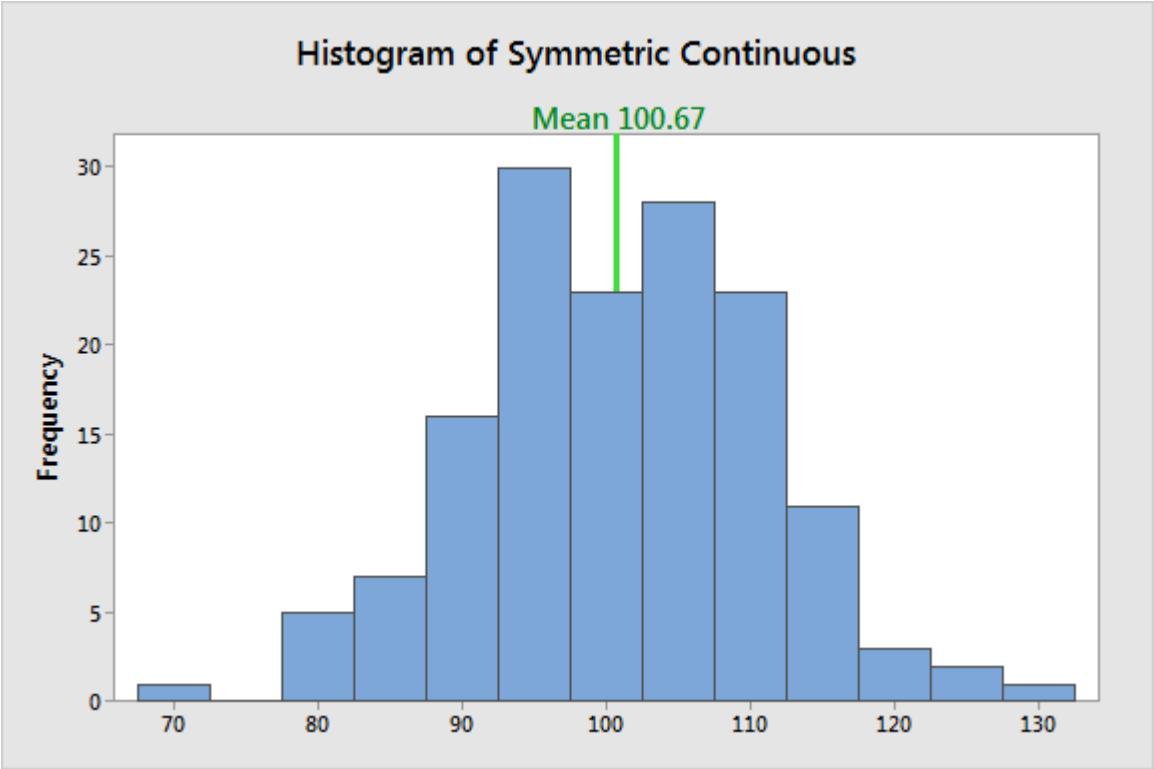


Mean

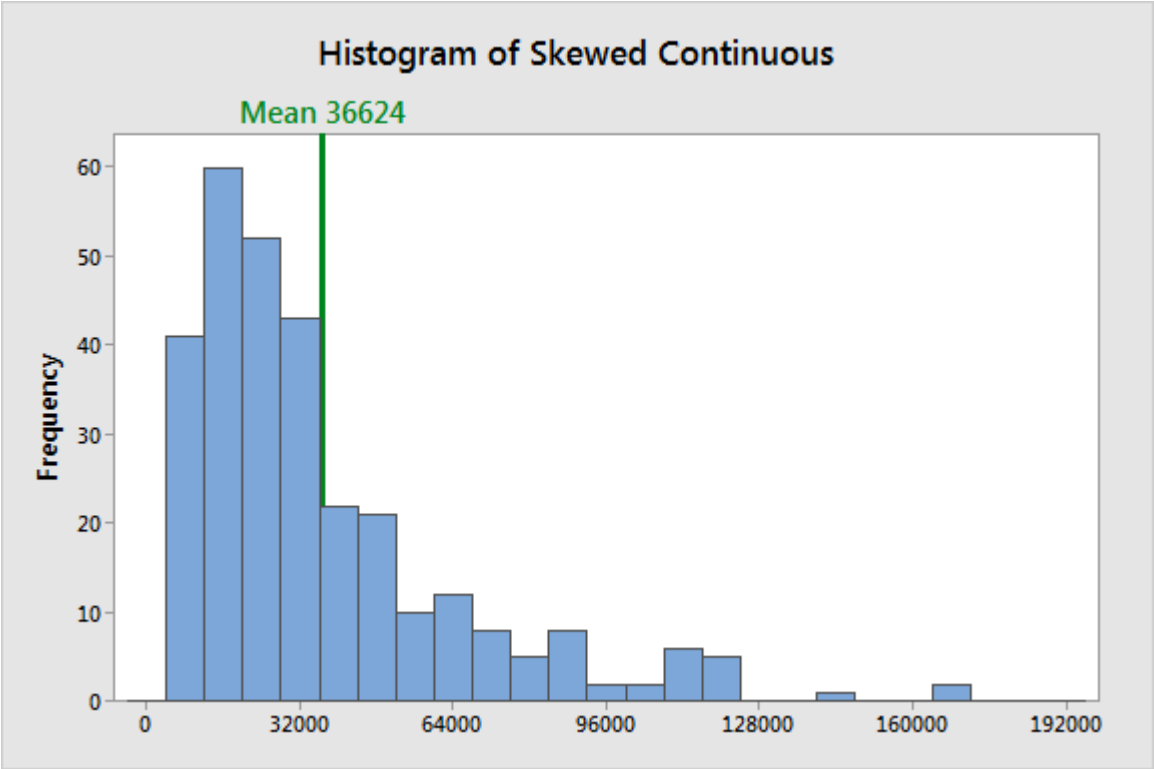
The mean is the arithmetic average, and it is probably the measure of central tendency that you are most familiar. Calculating the mean is very simple. You just add up all of the values and divide by the number of observations in your dataset.

$$\frac{x_1 + x_2 + \cdots + x_n}{n}$$

The calculation of the mean incorporates all values in the data. If you change any value, the mean changes. However, the mean doesn't always locate the center of the data accurately. Observe the histograms below where I display the mean in the distributions.



In a symmetric distribution, the mean locates the center accurately.



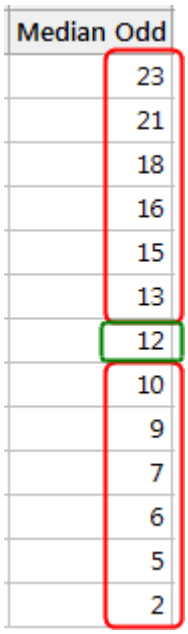
However, in a skewed distribution, the mean can miss the mark. In the histogram above, it is starting to fall outside the central area. This problem occurs because outliers have a substantial impact on the mean. Extreme values in an extended tail pull the mean away from the center. As the distribution becomes more skewed, the mean is drawn further away from the center. Consequently, it's best to use the mean as a measure of the central tendency when you have a symmetric distribution.

When to use the mean: Symmetric distribution, Continuous data

Median

The median is the middle value. It is the value that splits the dataset in half. To find the median, order your data from smallest to largest, and then find the data point that has an equal amount of values above it and below it. The method for locating the median varies slightly depending on whether your dataset has an even or odd number of values. I'll show you how to find the median for both cases. In the examples below, I use whole numbers for simplicity, but you can have decimal places.

In the dataset with the odd number of observations, notice how the number 12 has six values above it and six below it. Therefore, 12 is the median of this dataset.



Median Odd
23
21
18
16
15
13
12
10
9
7
6
5
2

When there is an even number of values, you count in to the two innermost values and then take the average. The average of 27 and 29 is 28. Consequently, 28 is the median of this dataset.

Median Even
40
38
35
33
32
30
28
29
27
26
24
23
22
19
17

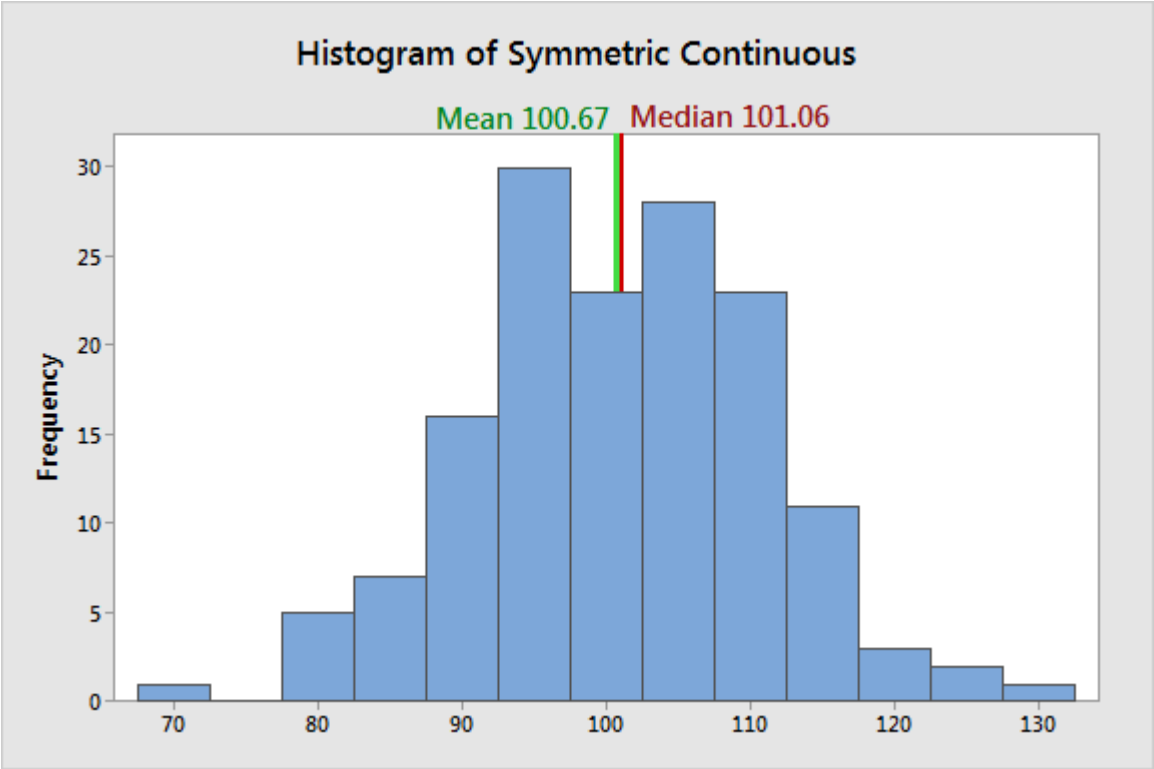
Outliers and skewed data have a smaller effect on the median. To understand why, imagine we have the Median dataset below and find that the median is 46. However, we discover data entry errors and need to change four values, which are shaded in the Median Fixed dataset. We'll make them all significantly higher so that we now have a skewed distribution with large outliers.

Median	Median Fixed
69	112
56	93
54	89
52	82
47	47
46	46
46	46
45	45
43	43
36	36
35	35
34	34
31	31

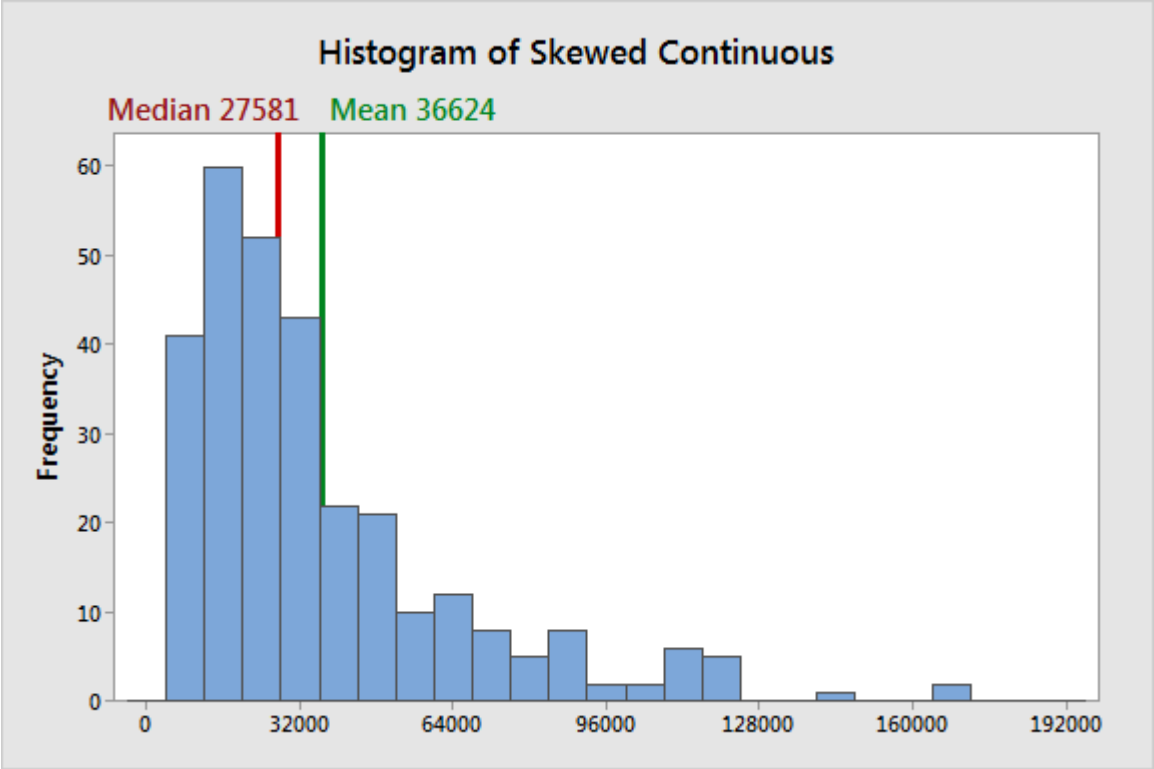
As you can see, the median doesn't change at all. It is still 46. Unlike the mean, the median value doesn't depend on all the values in the dataset. Consequently, when some of the values are more extreme, the effect on the median is smaller. Of course, with other types of changes, the median can change. When you have a skewed distribution, the median is a better measure of central tendency than the mean.

Comparing the mean and median

Now, let's test the median on the symmetrical and skewed distributions to see how it performs, and I'll include the mean on the histograms so we can make comparisons.



In a symmetric distribution, the mean and median both find the center accurately. They are approximately equal.



In a skewed distribution, the outliers in the tail pull the mean away from the center towards the longer tail. For this example, the mean and median differ by over 9000, and the median better represents the central tendency for the distribution.

These data are based on the U.S. household income for 2006. Income is the classic example of when to use the median because it tends to be skewed. The median indicates that half of all incomes

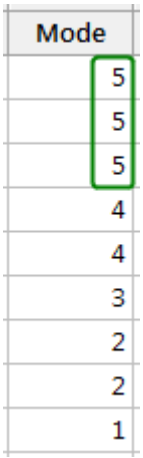
fall below 27581, and half are above it. For these data, the mean overestimates where most household incomes fall.

When to use the median: Skewed distribution, Continuous data, Ordinal data

Mode

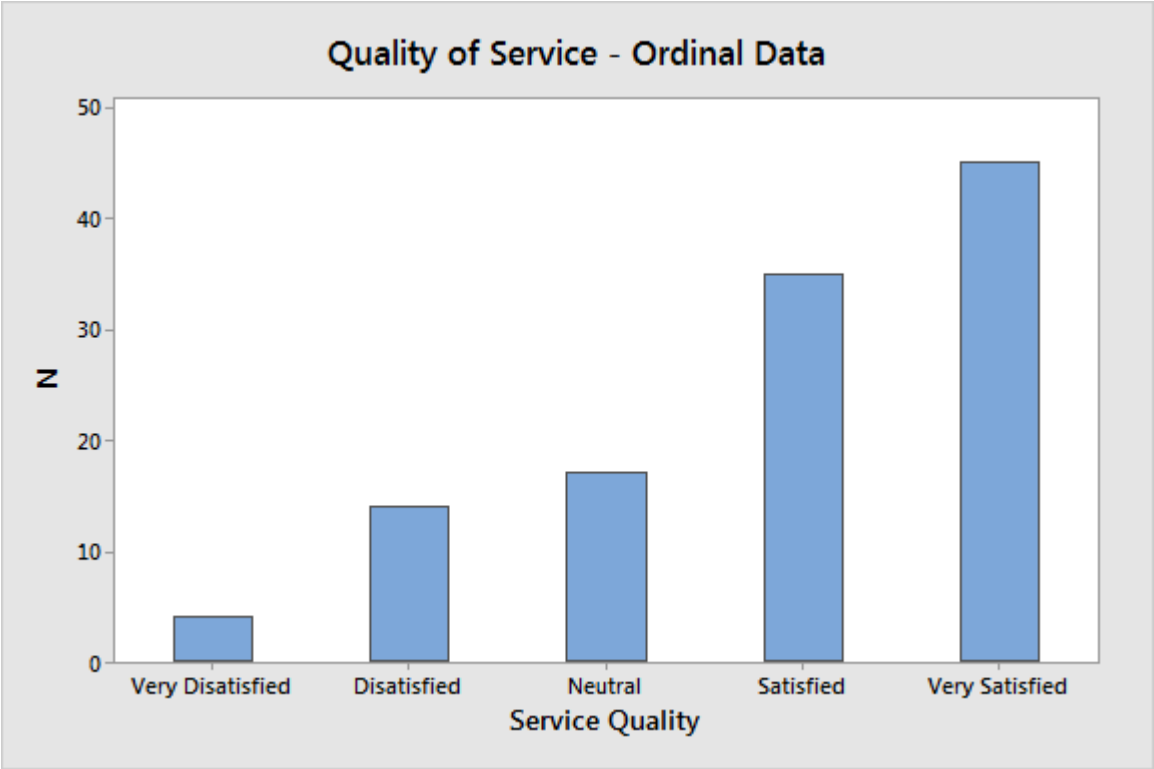
The mode is the value that occurs the most frequently in your data set. On a bar chart, the mode is the highest bar. If the data have multiple values that are tied for occurring the most frequently, you have a multimodal distribution. If no value repeats, the data do not have a mode.

In the dataset below, the value 5 occurs most frequently, which makes it the mode. These data might represent a 5-point Likert scale.



Mode
5
5
5
4
4
3
2
2
1

Typically, you use the mode with categorical, ordinal, and discrete data. In fact, the mode is the only measure of central tendency that you can use with categorical data—such as the most preferred flavor of ice cream. However, with categorical data, there isn't a central value because you can't order the groups. With ordinal and discrete data, the mode can be a value that is not in the center. Again, the mode represents the most common value.



In the graph of service quality, Very Satisfied is the mode of this distribution because it is the most common value in the data. Notice how it is at the extreme end of the distribution. I'm sure the service providers are pleased with these results!

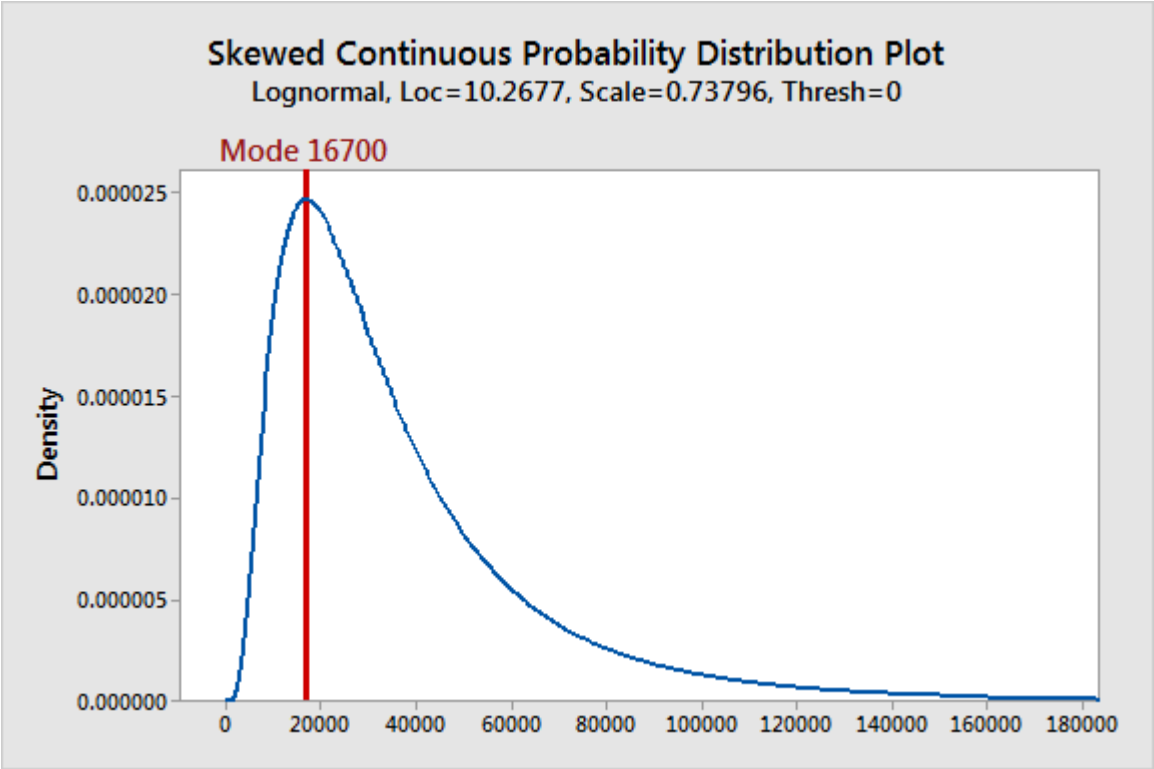
Finding the mode for continuous data

In the continuous data below, no values repeat, which means there is no mode. With continuous data, it is unlikely that two or more values will be exactly equal because there are an infinite number of values between any two values.

No Mode
122.275
109.085
103.079
102.691
98.228
96.221
94.724
92.619
89.483
75.762

When you are working with the raw continuous data, don't be surprised if there is no mode. However, you can find the mode for continuous data by locating the maximum value on a probability distribution plot. If you can identify a probability distribution that fits your data, find the peak value and use it as the mode.

The probability distribution plot displays a lognormal distribution that has a mode of 16700. This distribution corresponds to the U.S. household income example in the median section.



When to use the mode: Categorical data, Ordinal data, Count data, Probability Distributions

Which is Best—the Mean, Median, or Mode?

When you have a symmetrical distribution for continuous data, the mean, median, and mode are equal. In this case, analysts tend to use the mean because it includes all of the data in the calculations. However, if you have a skewed distribution, the median is often the best measure of central tendency.

When you have ordinal data, the median or mode is usually the best choice. For categorical data, you have to use the mode.

In cases where you are deciding between the mean and median as the better measure of central tendency, you are also determining which types of statistical hypothesis tests are appropriate for your data—if that is your ultimate goal. I have written an article that discusses when to use parametric (mean) and nonparametric (median) hypothesis tests along with the advantages and disadvantages of each type.

ANOVA, Regression, and Chi-Square

A variety of statistical procedures exist. The appropriate statistical procedure depends on the research question(s) we are asking and the type of data we collected. While EPSY 5601 is not intended to be a statistics class, some familiarity with different statistical procedures is warranted.

Parametric Data Analysis

Investigating Differences

One Independent Variable (With Two Levels) and One Dependent Variable

When we wish to know whether the means of two groups (one independent variable (e.g., gender) with two levels (e.g., males and females) differ, a *t* test is appropriate. In order to calculate a *t* test, we need to know the mean, standard deviation, and number of subjects in each of the two groups. An example of a *t* test research question is "*Is there a significant difference between the reading scores of boys and girls in sixth grade?*" A sample answer might be, "Boys (*M*=5.67, *SD*=.45) and girls (*M*=5.76, *SD*=.50) score similarly in reading, *t*(23)=.54, *p*>.05." [Note: The (23) is the degrees of freedom for a *t* test. It is the number of subjects minus the number of groups (always 2 groups with a t-test). In this example, there were 25 subjects and 2 groups so the degrees of freedom is 25-2=23.] Remember, a *t* test can only compare the means of two groups (independent variable, e.g., gender) on a single dependent variable (e.g., reading score). You may wish to review the instructor notes for *t* tests.

One Independent Variable (With More Than Two Levels) and One Dependent Variable

If the independent variable (e.g., political party affiliation) has more than two levels (e.g., Democrats, Republicans, and Independents) to compare and we wish to know if they differ on a dependent variable (e.g., attitude about a tax cut), we need to do an ANOVA (**AN**alysis **O**f **VA**riance). In other words, if we have one independent variable (with three or more groups/levels) and one dependent variable, we do a one-way ANOVA. A sample research question is, "*Do Democrats, Republicans, and Independents differ on their option about a tax cut?*" A sample answer is, "Democrats (*M*=3.56, *SD*=.56) are less likely to favor a tax cut than Republicans (*M*=5.67, *SD*=.60) or Independents (*M*=5.34, *SD*=.45), *F*(2,120)=5.67, *p*<.05." [Note: The (2,120) are the degrees of freedom for an ANOVA. The first number is the number of groups minus 1. Because we had three political parties it is 2, 3-1=2. The second number is the total number of subjects minus the number of groups. Because we had 123 subject and 3 groups, it is 120 (123-3)]. The one-way ANOVA has one independent variable (political party) with more than two groups/levels (Democrat, Republican, and Independent) and one dependent variable (attitude about a tax cut).

More Than One Independent Variable (With Two or More Levels Each) and One Dependent Variable

ANOVAs can have more than one independent variable. A two-way ANOVA has two independent variable (e.g. political party and gender), a three-way ANOVA has three independent variables (e.g., political party, gender, and education status), etc. These ANOVA still only have one dependent variable (e.g., attitude about a tax cut). A two-way ANOVA has three research questions: One for each of the two independent variables and one for the interaction of the two independent variables.

Sample Research Questions for a Two-Way ANOVA:
Do Democrats, Republicans, and Independents differ on their opinion about a tax cut?
Do males and females differ on their opinion about a tax cut?
Is there an interaction between gender and political party affiliation regarding opinions about a tax cut?

A two-way ANOVA has three null hypotheses, three alternative hypotheses and three answers to the research question. The answers to the research questions are similar to the answer provided for the one-way ANOVA, only there are three of them.

One or More Independent Variables (With Two or More Levels Each) and More Than One Dependent Variable

Sometimes we have several independent variables and several dependent variables. In this case we do a MANOVA (**M**ultiple **AN**alysis **O**f **VA**riance). Suffices to say, multivariate statistics (of which MANOVA is a member) can be rather complicated.

Investigating Relationships

Simple Correlation

Sometimes we wish to know if there is a relationship between two variables. A simple correlation measures the relationship between two variables. The variables have equal status and are not considered independent variables or dependent variables. In our class we used Pearson's *r* which measures a linear relationship between two continuous variables. While other types of relationships with other types of variables exist, we will not cover them in this class. A sample research question for a simple correlation is, "*What is the relationship between height and arm span*?" A sample answer is, "There is a relationship between height and arm span, $r(34)=.87$, $p<.05$." You may wish to review the instructor notes for correlations. A canonical correlation measures the relationship between sets of multiple variables (this is multivariate statistic and is beyond the scope of this discussion).

Regression

An extension of the simple correlation is regression. In regression, one or more variables (predictors) are used to predict an outcome (criterion). One may wish to predict a college student's GPA by using his or her high school GPA, SAT scores, and college major. Data for several hundred students would be fed into a regression statistics program and the statistics program would determine how well the predictor variables (high school GPA, SAT scores, and college major) were related to the criterion variable (college GPA). Based on the information, the program would create a mathematical formula for predicting the criterion variable (college GPA) using those predictor variables (high school GPA, SAT scores, and/or college major) that are significant. Not all of the variables entered may be significant predictors. A sample research question might be, "*What is the individual and combined power of high school GPA, SAT scores, and college major in predicting graduating college GPA?*" The output of a regression analysis contains a variety of information. R^2 tells how much of the variation in the criterion (e.g., final college GPA) can be accounted for by the predictors (e.g., high school GPA, SAT scores, and college major (dummy coded 0 for Education Major and 1 for Non-Education Major). A research report might note that "High school GPA, SAT scores, and college major are significant predictors of final college GPA, $R^2=.56$." In this example, 56% of an individual's college GPA can be predicted with his or her high school GPA, SAT scores, and college major). The regression equation for such a study might look like the following: Y'= .15 + (HS GPA * .75) + (SAT * .001) + (Major * -.75). By inserting an individual's high school GPA, SAT score, and

college major (0 for Education Major and 1 for Non-Education Major) into the formula, we could predict what someone's final college GPA will be (well…at least 56% of it). For example, someone with a high school GPA of 4.0, SAT score of 800, and an education major (0), would have a predicted GPA of 3.95 (.15 + (4.0 * .75) + (800 * .001) + (0 * -.75)). Universities often use regression when selecting students for enrollment.

I have created a sample SPSS regression printout with interpretation if you wish to explore this topic further. You will not be responsible for reading or interpreting the SPSS printout.

Non Parametric Data Analysis

Chi-Square

We might count the incidents of something and compare what our actual data showed with what we would expect. Suppose we surveyed 27 people regarding whether they preferred red, blue, or yellow as a color. If there were no preference, we would expect that 9 would select red, 9 would select blue, and 9 would select yellow. We use a chi-square to compare what we observe (actual) with what we expect. If our sample indicated that 2 liked red, 20 liked blue, and 5 liked yellow, we might be rather confident that more people prefer blue. If our sample indicated that 8 liked read, 10 liked blue, and 9 liked yellow, we might not be very confident that blue is generally favored. Chi-square helps us make decisions about whether the observed outcome differs significantly from the expected outcome. A sample research question is, "*Is there a preference for the red, blue, and yellow color?*" A sample answer is "There was not equal preference for the colors red, blue, or yellow. More people preferred blue than red or yellow, X^2 (2) = 12.54, $p < .05$". Just as t-tests tell us how confident we can be about saying that there are differences between the means of two groups, the chi-square tells us how confident we can be about saying that our observed results differ from expected results.

Correlation testing via t test

As in Sampling Distributions, we can consider the distribution of r over repeated samples of x and y. The following theorem is analogous to the Central Limit Theorem, but for r instead of $\bar{x}$. This time we require that x and y have a joint bivariate normal distribution or that samples are sufficiently large. You can think of a bivariate normal distribution as the three-dimensional version of the normal distribution, in which any vertical slice through the surface which graphs the distribution results in an ordinary bell curve.

The sampling distribution of r is only symmetric when $\rho = 0$ (i.e. when x and y are independent). If $\rho \neq 0$, then the sampling distribution is asymmetric and so the following theorem does not apply, and other methods of inference must be used.

Theorem 1: Suppose $\rho = 0$. If x and y have a bivariate normal distribution or if the sample size n is sufficiently large, then r has a normal distribution with mean 0, and $t = r/s_r \sim T(n-2)$ where

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

Here the numerator r of the random variable t is the estimate of $\rho = 0$ and s_r is the standard error of t.

Observation: If we solve the equation in Theorem 1 for r, we get

$$r^2 = \frac{t^2}{t^2+df}$$

Observation: The theorem can be used to test the hypothesis that population random variables x and y are independent i.e. $\rho = 0$.

Example 1: A study is designed to check the relationship between smoking and longevity. A sample of 15 men 50 years and older was taken and the average number of cigarettes smoked per day and the age at death was recorded, as summarized in the table in Figure 1. Can we conclude from the sample that longevity is independent of smoking?

Cigarettes	5	23	25	48	17	8	4	26	11	19	14	35	29	4	23
Longevity	80	78	60	53	85	84	73	79	81	75	68	72	58	92	65

Figure 1 – Data for Example 1

The scatter diagram for this data is as follows. We have also included the linear trend line that seems to best match the data. We will study this further in Linear Regression.

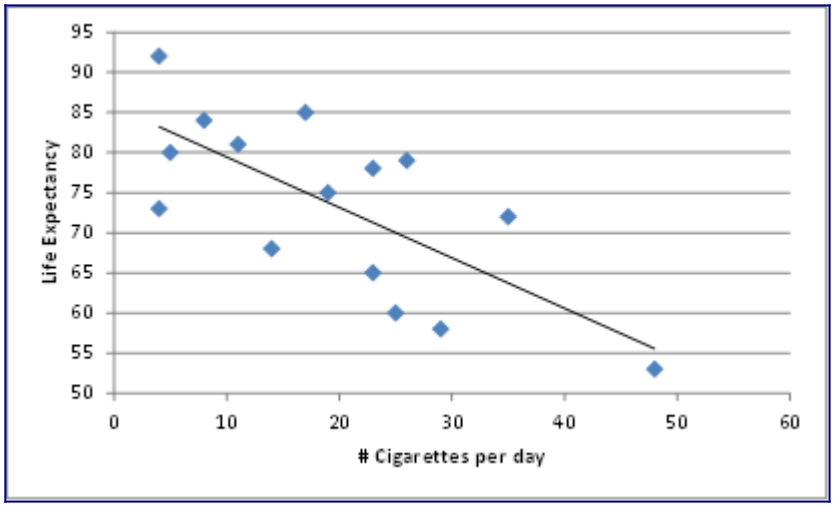


Figure 2 – Scatter diagram for Example 1

Next we calculate the correlation coefficient of the sample using the CORREL function:

r = CORREL(R1, R2) = -.713

From the scatter diagram and the correlation coefficient, it is clear that the population correlation is likely to be negative. The absolute value of the correlation coefficient looks high, but is it high enough? To determine this, we establish the following null hypothesis:

H_0: $\rho = 0$

Recall that $\rho = 0$ would mean that the two population variables are independent. We use $t = r/s_r$ as the test statistic where s_r is as in Theorem 1. Based on the null hypothesis, $\rho = 0$, we can apply Theorem 1, provided x and y have a bivariate normal distribution. It is difficult to check for bivariate normality, but we can at least check to make sure that each variable is approximately normal via QQ plots.

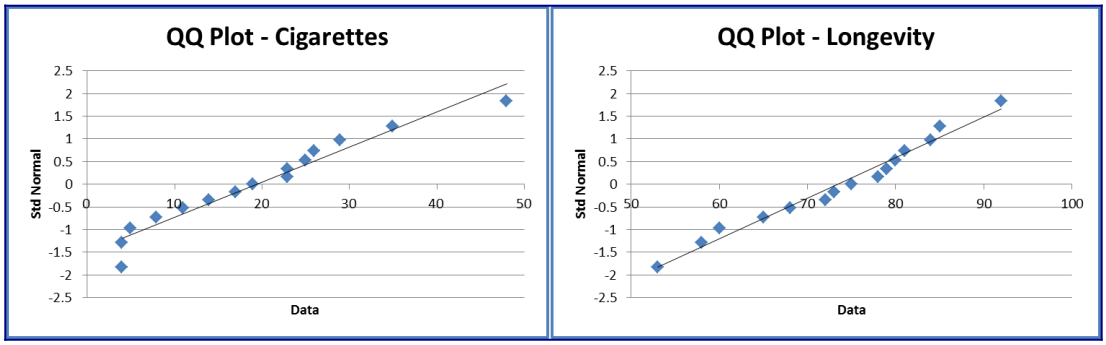


Figure 3 – Testing for normality

Both samples appear to normal, and so by Theorem 1, we know that t has approximately a t distribution with $n - 2 = 13$ degrees of freedom. We now calculate

$s_r = \sqrt{(1-r^2)/(n-2)} = \sqrt{(1-.713^2)/(15-2)} = 0.194$ $t = r/s_r = -.713/0.194 = -3.67$

Finally, we perform either one of the following tests:

p-value = TDIST(ABS(-3.67), 13, 2) = .00282 < .05 = α (two-tail)

t_{crit} = TINV(.05, 13) = 2.16 < 3.67 = $|t_{obs}|$

And so we reject the null hypothesis and conclude there is a non-zero correlation between smoking and longevity. In fact, it appears from the data that increased levels of smoking reduce longevity.

Correlation and Chi-square Test for Independence

In Independence Testing we used the chi-square test to determine whether two variables were independent. We now look at the same problem using dichotomous variables.

Example 1: Calculate the point-biserial correlation coefficient for the data in Example 2 of Independence Testing (repeated in Figure 1) using dichotomous variables (repeated in Figure 1).

	A	B	C	D
3		Therapy 1	Therapy 2	Total
4	Cured	31	57	88
5	Not Cured	11	51	62
6	Total	42	108	150

Figure 1 – Contingency table for data in Example 1

This time let $x = 1$ if the patient is cured and $x = 0$ if the patient is not cured, and let y = 1 if therapy 1 is used and y = 0 if therapy 2 is used. Thus for 31 patients $x = 1$ and y = 1, for 11 patients $x = 0$ and y = 1, for 57 patients $x = 1$ and y = 0 and for 51 patients $x = 0$ and y = 0.

If we list all 150 pairs of x and y as shown in range P3:Q152 of Figure 2 (only the first 6 data rows are displayed) we can calculate the correlation coefficient using the CORREL function to get = .192.

	O	P	Q	R	S	T	U
2		x	y				
3		1	1	r	0.191767		=CORREL(P3:P152,Q3:Q152)
4		1	1				
5		1	1				
6		1	1				
7		1	1				
8		1	1				

Figure 2 – Calculation of the point-biserial correlation coefficient

Observation: Instead of listing all the n pairs of samples values and using the CORREL function, we can calculate the correlation coefficient using Property 3 of Relationship between Correlation and t Test, which is especially useful for large values of n. This is shown in Figure 3.

	F	G	H	I
2	Point biserial correlation			
3				
4	m1	0.7381		=B4/B6
5	m0	0.5278		=C4/C6
6	n1	42		=B6
7	n0	108		=C6
8	n	150		=G6+G7
9	s	0.4941		=SQRT(D4*D5/(G8*(G8-1)))
10	r	0.1918		=(G4-G5)/G9*SQRT(G6*G7/(G8*(G8-1)))

Figure 3 – Alternative approach

Actually, based on a little algebra it is easy to see that the correlation coefficient can also be calculated using the formula =(B4*C6-C4*B6)/SQRT(B6*C6*D4*D5).

Property 1: For problems such as those in Example 1, if $\rho = 0$ (the null hypothesis), then $nr^2 \sim \chi^2$ (1).

Observation: Property 1 provides an alternative method for carrying out chi-square tests such as the one we did in Example 2 of Independence Testing.

Example 2: Using Property 1, determine whether there is a significant difference in the two therapies for curing patients of cocaine dependence based on the data in Figure 1.

	K	L	M	N
2				
3	r	0.191767		=S3
4	n	150		=D6
5	$\chi 2 = nr^2$	5.516208		=L4*L3^2
6	p-value	0.018841		=CHISQ.DIST.RT(L5,1)

Figure 4 – Chi-square test for Example 2

Note that the chi-square value of 5.67 is the same as we saw in Example 2 of Chi-square Test of Independence. Since the p-value = CHITEST(5.67,1) = 0.017 < .05 = α, we again reject the null hypothesis and conclude there is a significant difference between the two therapies.

Observation: If we calculate the value of χ^2 for independence as in Independence Testing, from the previous observation we conclude that $r = \sqrt{\chi^2/n}$. This gives us a way to measure the effect of the chi-square test of independence, namely $\varphi = \sqrt{\chi^2/n}$.

Care should be taken with the use of φ since even relatively small values can indicate an important effect. E.g. in the previous example, there is clearly an important

difference between the two therapies (not just a significant difference), but if you look at *r* we see that only 4.3% of the variance is explained by the choice of therapy.

Observation: In Example 1 we calculated the correlation coefficient of *x* with y by listing all 132 values and then using Excel's correlation function CORREL. The following is an alternative approach for calculating *r*, which is especially useful if *n* is very large.

	F	G	H	I	J	K	L	M	N	O
3	x	y	count		x - x̄	y - ȳ	(x - x̄)(y - ȳ)	(x - x̄)²	(y - ȳ)²	count
4	1	1	31		0.413333	0.72	0.2976	0.17084444	0.5184	31
5	0	1	11		-0.58667	0.72	-0.4224	0.34417778	0.5184	11
6	1	0	57		0.413333	-0.28	-0.11573333	0.17084444	0.0784	57
7	0	0	51		-0.58667	-0.28	0.16426667	0.34417778	0.0784	51
8			150			Σ	6.36	36.3733333	30.24	
9	x̄	ȳ								
10	0.5867	0.28								

Figure 5 – Calculation of *r* for data in Example 1

First, we repeat the data from Figure 1 using the dummy variables *x* and y (in range F4:H7). Essentially this is a frequency table. We then calculate the mean of *x* and y. E.g. the mean of *x* (in cell F10) is calculated by the formula =SUMPRODUCT(F4:F7,H4:H7)/H8.

Next we calculate $\sum(x_i - \bar{x})(y_i - \bar{y})$, $\sum(x_i - \bar{x})^2$ and $\sum(y_i - \bar{y})^2$ (in cells L8, M8 and N8). E.g. the first of these terms is calculated by the formula =SUMPRODUCT(L4:L7,O4:O7). Now the **point-serial correlation coefficient** is the first of these terms divided by the square root of the product of the other two, i.e. *r* = L8/SQRT(M8*N8).

What Is Hypothesis Testing?

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process. The word "population" will be used for both of these cases in the following descriptions.

How Hypothesis Testing Works

In hypothesis testing, an analyst tests a statistical sample, with the goal of providing evidence on the plausibility of the null hypothesis.

Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed. All analysts use a random population sample to test two different hypotheses: the null hypothesis and the alternative hypothesis.

The null hypothesis is usually a hypothesis of equality between population parameters; e.g., a null hypothesis may state that the population mean return is equal to zero. The alternative hypothesis is effectively the opposite of a null hypothesis (e.g., the population mean return is not equal to zero). Thus, they are mutually exclusive, and only one can be true. However, one of the two hypotheses will always be true.

4 Steps of Hypothesis Testing

All hypotheses are tested using a four-step process:

1. The first step is for the analyst to state the two hypotheses so that only one can be right.
2. The next step is to formulate an analysis plan, which outlines how the data will be evaluated.
3. The third step is to carry out the plan and physically analyze the sample data.
4. The fourth and final step is to analyze the results and either reject the null hypothesis, or state that the null hypothesis is plausible, given the data.

Real-World Example of Hypothesis Testing

If, for example, a person wants to test that a penny has exactly a 50% chance of landing on heads, the null hypothesis would be that 50% is correct, and the alternative hypothesis would be that 50% is not correct.

Mathematically, the null hypothesis would be represented as Ho: P = 0.5. The alternative hypothesis would be denoted as "Ha" and be identical to the null hypothesis, except with the equal sign struck-through, meaning that it does not equal 50%.

A random sample of 100 coin flips is taken, and the null hypothesis is then tested. If it is found that the 100 coin flips were distributed as 40 heads and 60 tails, the analyst would assume that a penny does not have a 50% chance of landing on heads and would reject the null hypothesis and accept the alternative hypothesis.

If, on the other hand, there were 48 heads and 52 tails, then it is plausible that the coin could be fair and still produce such a result. In cases such as this where the null hypothesis is "accepted," the analyst states that the difference between the expected results (50 heads and 50 tails) and the observed results (48 heads and 52 tails) is "explainable by chance alone."

The Difference Between Bivariate & Multivariate Analyses

Bivariate and multivariate analyses are statistical methods to investigate relationships between data samples. Bivariate analysis looks at two paired data sets, studying whether a relationship exists between them. Multivariate analysis uses two or more variables and analyzes which, if any, are correlated with a specific outcome. The goal in the latter case is to determine which variables influence or cause the outcome.

Bivariate Analysis

Bivariate analysis investigates the relationship between two data sets, with a pair of observations taken from a single sample or individual. However, each sample is independent. You analyze the data using tools such as t-tests and chi-squared tests, to see if the two groups of data correlate with each other. If the variables are quantitative, you usually graph them on a scatterplot. Bivariate analysis also examines the strength of any correlation.

Bivariate Analysis Examples

One example of bivariate analysis is a research team recording the age of both husband and wife in a single marriage. This data is paired because both ages come from the same marriage, but independent because one person's age doesn't cause another person's age. You plot the data to showing a correlation: the older husbands have older wives. A second example is recording measurements of individuals' grip strength and arm strength. The data is paired because both measurements come from a single person, but independent because different muscles are used. You plot data from many individuals to show a correlation: people with higher grip strength have higher arm strength.

Multivariate Analysis

Multivariate analysis examines several variables to see if one or more of them are predictive of a certain outcome. The predictive variables are independent variables and the outcome is the dependent variable. The variables can be continuous, meaning they can have a range of values, or they can be dichotomous, meaning they represent the answer to a yes or no question. Multiple regression analysis is the most common method used in multivariate analysis to find correlations between data sets. Others include logistic regression and multivariate analysis of variance.

Multivariate Analysis Example

Multivariate analysis was used in by researchers in a 2009 Journal of Pediatrics study to investigate whether negative life events, family environment, family violence, media violence and depression are predictors of youth aggression and bullying. In this case, negative life events, family environment, family violence, media violence and depression were the independent predictor variables, and aggression and bullying were the dependent outcome variables. Over 600 subjects, with an average age of 12 years old, were given questionnaires to determine the predictor variables for each child. A survey also determined the outcome variables for each child. Multiple regression equations and structural equation modeling was used to study the data set. Negative life events and depression were found to be the strongest predictors of youth aggression.

Additional Links

1. Statistics versus machine learning
   https://www.nature.com/articles/nmeth.4642.pdf\

2. Statistical Hypothesis Analysis in Python with ANOVAs, Chi-Square, and Pearson Correlation
   https://stackabuse.com/statistical-hypothesis-analysis-in-python-with-anovas-chi-square-and-pearson-correlation/
3. Statistics Solutions
   https://www.statisticssolutions.com/wp-content/uploads/SPSS Manual.pdf
4. Bivariate &/vs. Multivariate
   https://psych.unl.edu/psycrs/350/unit4/biv_mult.pdf
5. Hypothesis Testing
   https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/