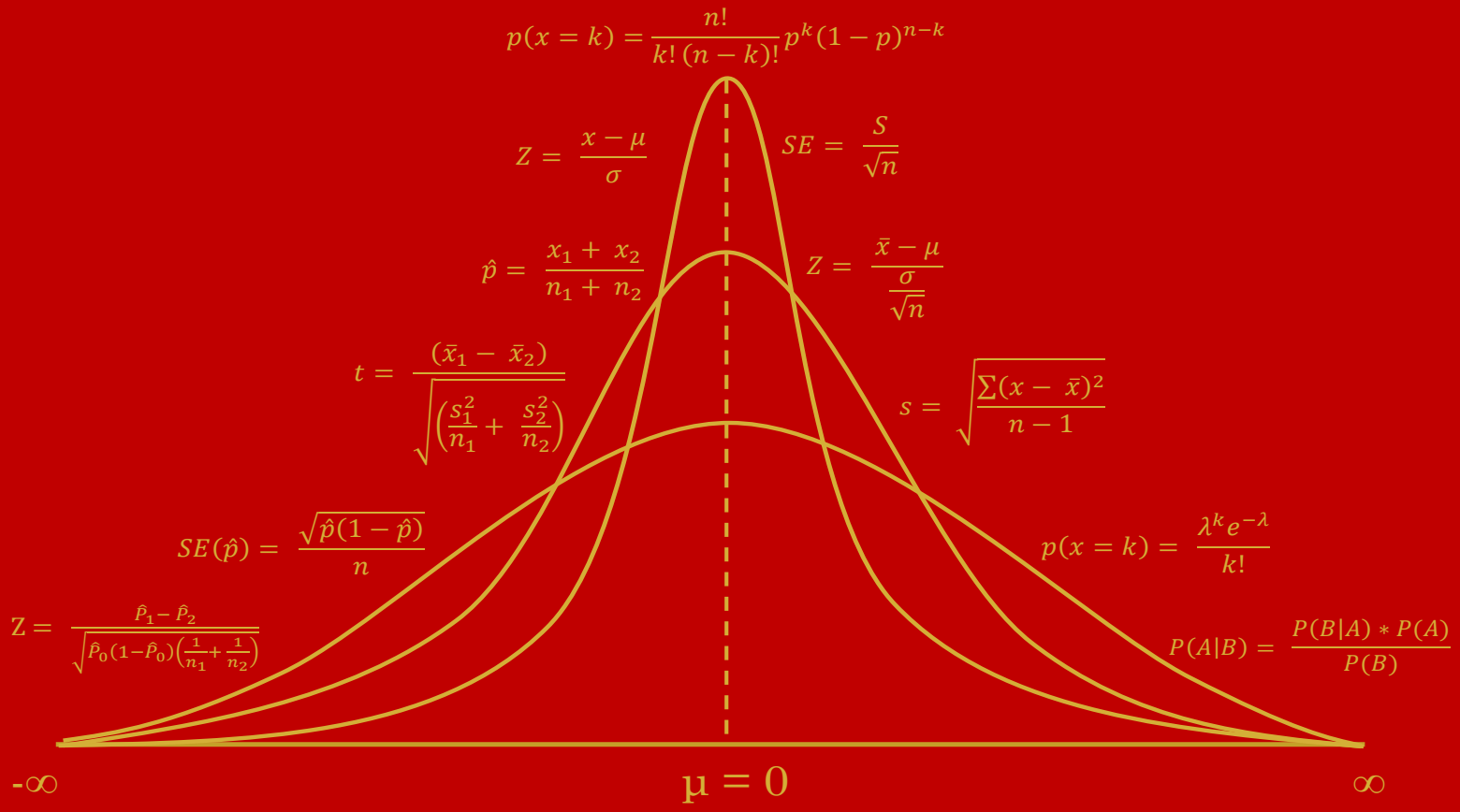
Applied Statistics for Data Science with Python



Alexander Thompson BSc (Hons)

Applied Statistics for Data Science with Python

Alexander Thompson BSc (Hons)

28th January 2026

Contents

Preface

This is the Statistics checklist for computing statistics in Python, and is sourced from the statistical information from the Google Advanced Data Analytics Professional Certificate. This book covers in depth statistical methodology and how to apply them to the data science field, covering the topics of probability and its distributions to sampling and the central limit theorem to confidence intervals and hypothesis tests. These concepts enable oneself to be able to extend their analytical outreach with the end result being an improved analysis of the data.

In regards to the data science field, these concepts play the role of a bridge, connecting exploratory data analysis to regression and machine learning modelling. This is due to the statistical concepts explaining everything from the basics, such as descriptive statistics and probabilities to the more advanced concepts such as hypothesis testing, with regression modelling building directly on top of this. With respect to machine learning, these statistical concepts are heavily applied, such as with the Naive Bayes machine learning technique using posterior probability and the Random Forest Technique using sampling with replacement. Therefore, these concepts bridge the gap between basic data analytics and high-level data science and is why this book is essential for covering these concepts.

These concepts are carried out using the programming language Python using the appropriate functions and mainly using the statsmodels library. Due to the updates made over time to the Python language the code in this book will have changed over time as the different versions of Python get released. Python is a great tool for statistical analysis and works excellently within the data analytics framework. Using Python allows the analyst to go straight from the data preparation to machine learning model development all in the same script or notebook.

This book is the full and complete guide to descriptive and inferential statistics and their applications via Python, and is intended for use as a handbook for revision and applied work. Whether it is revising statistical concepts for a project or teaching another analyst about statistics and how to apply them, this book is the definitive guide and will help any data scientist reach their goals. A majority of the concepts covered here in this book I am very familiar with due to my Bachelors of Science degree in Economics, where hypothesis testing and descriptive statistics were prevalent. Therefore, my added input into these concepts will aid in the understanding and application of these concepts and how they are applied into the real analytical world.

"Fill your mind, not to impress others but to endure yourself"

Introduction

Statistics underpins nearly every stage of modern data science, from exploratory data analysis to predictive modelling and evidence-based decision making under uncertainty. In applied contexts such as economics, finance and machine learning, statistical reasoning is essential not only for producing analytical outputs but also for interpreting the results correctly, assessing their ability and communicating insights with clarity and precision.

Within economics, statistical methods have long been used to analyse relationships between variables, test theoretical models and draw inferences about populations, markets and behaviours from observed data. These same principles now form the backbone of contemporary data science, where analysts work with increasingly large complex datasets to model outcomes, forecast trends and inform strategic or policy decisions. In both domains, the ability to reason statistically provides the crucial link between raw data and meaningful interpretation. At its core, statistics offers a structured framework for understanding variation. Real-world data are inherently imperfect and are shaped by randomness, measurement error, sampling processes and unobserved influences. Statistical methods enable analysts to summarise data, quantify uncertainty and draw informed conclusions about populations based on finite samples. These foundations are critical not only for descriptive analysis, but also for the responsible application of regression models, classification algorithms and machine learning techniques, where the assumptions and uncertainty must be carefully considered.

The structure of this book reflects the progression from foundational concepts to applied analytical practice. It begins with descriptive statistics, focusing on the summarisation and visualisation of data through measures of central tendency, dispersion and distributional form. These tools represent the first step in any data science or economic analysis, providing essential insight into the structure, quality and limitations of a dataset prior to formal modelling. The text then introduces probability theory and sampling, thereby establishing the mathematical language required to reason rigorously about uncertainty and stochastic processes. Building on this foundation, the book advances to statistical inference, including estimation, confidence intervals, hypothesis testing and model-based reasoning. Throughout the book, Python is used as the primary computational environment to demonstrate how statistical concepts are applied in practice. Code examples are designed to mirror real analytical workflows commonly encountered in data science and economics, with an emphasis placed on interpretation, assumptions and limitations rather than just computation alone. This applied focus reflects the reality that effective data analysis requires both technical proficiency and sound statistical judgement.

This text is intended to serve as both a structured learning resource and a practical reference guide. Readers may follow the material sequentially to develop a coherent understanding of statistical foundations or consult individual sections to support revision, applied projects and professional analysis. By integrating statistical theory, economic intuition and practical implementation, the book aims to support robust, interpretable and methodologically sound data-driven analysis.

Descriptive Statistics

Formulas

Mean Formula:

$$Mean = \text{Sum of all Values/Total Number of Values}$$

Standard Deviation (Sample)

There are different formulas to calculate the standard deviation for a population and a sample. As a reminder, data professionals typically work with sample data, and they make inferences about populations based on the sample. So, let's review the formula for sample standard deviation:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

n = Total Number of Values in your Sample

x = Each individual data Value

$\bar{x}$ = The Mean of your data Values

$\sum$ = Sum (Greek Letter Sigma)

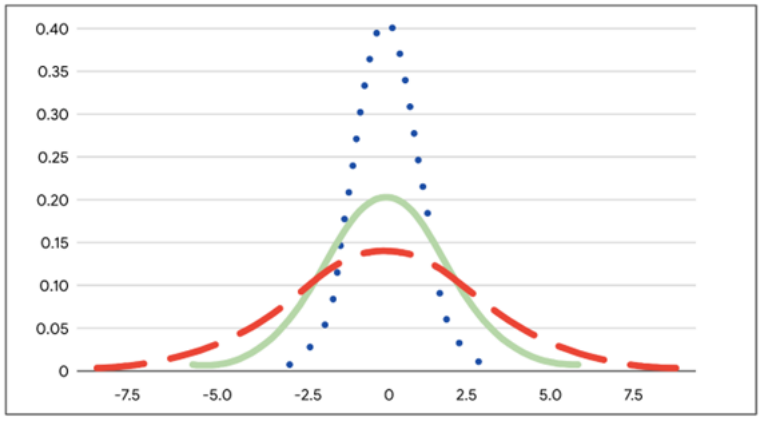


Figure 1: Measure of Dispersion: Std Deviation

Relationship between Quartiles and Percentiles

- Q1 = 25th Percentile
- Q2 = 50th Percentile
- Q3 = 75th Percentile

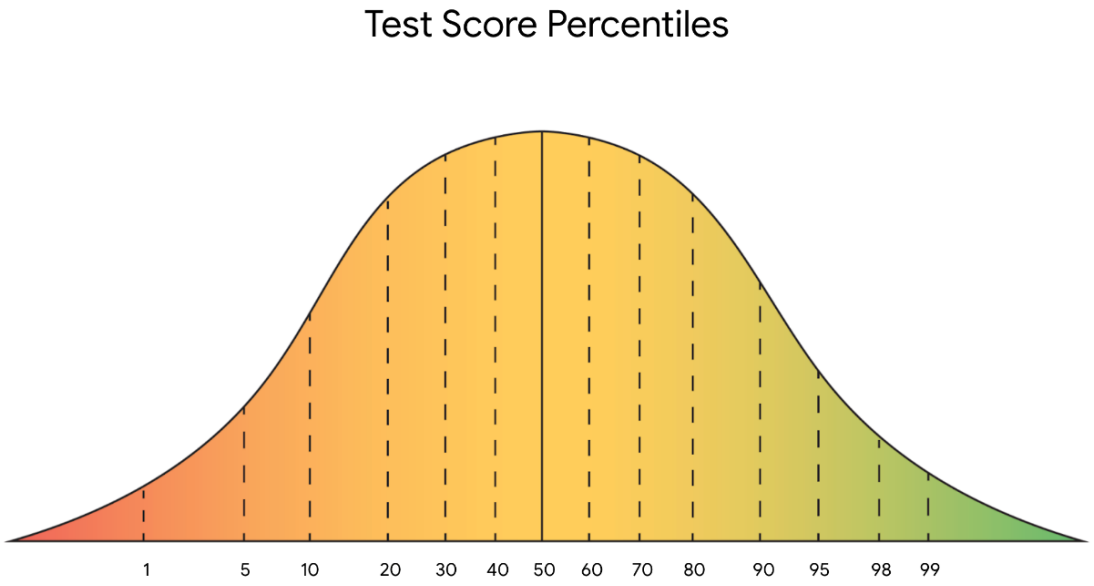


Figure 2: Measures of Position: Percentiles

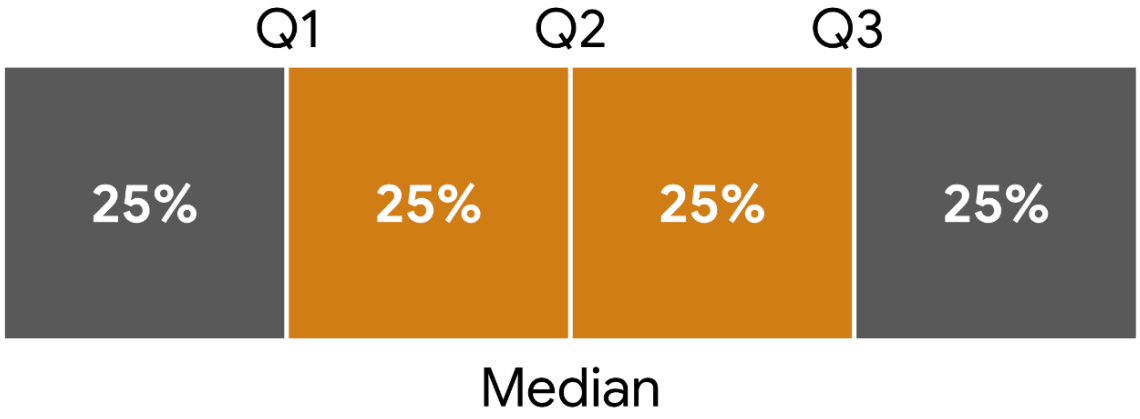


Figure 3: Measures of Position: Quartiles

IQR

$$IQR = Q3 - Q1$$

Five Number Summary

- The Minimum
- The First Quartile (Q1)
- The Median, or Second Quartile (Q2)
- The Third Quartile (Q3)
- The Maximum

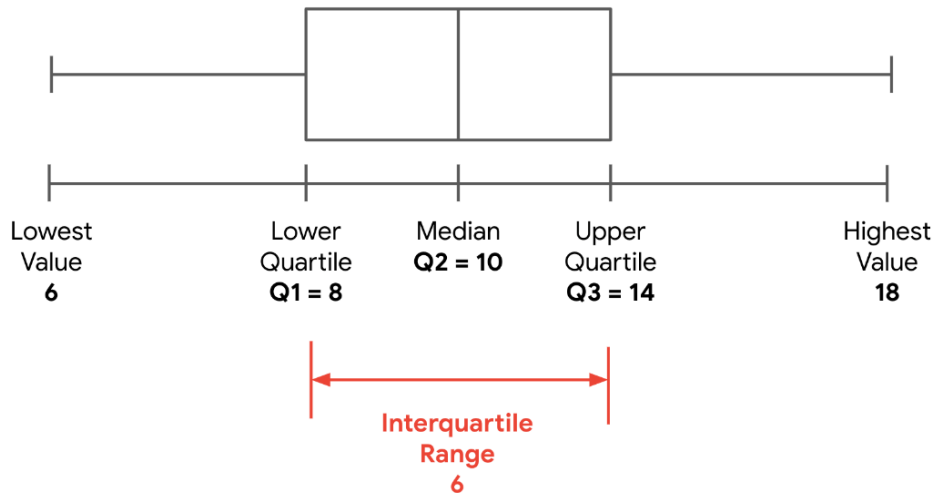


Figure 4: Measures of Position: The Interquartile Range

Descriptive Statistics in Python

Packages

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Functions

```
data = pd.read_csv('data.csv')

data.head()

data['column'].describe()
```

For the `.describe()` function above:

- The `mean` helps to clarify the centre of your dataset.
- The Categories `25%`, `50%` and `75%` refer to Q1, Q2 and Q3 respectively, Remember that Q2 is also the Median of your dataset.

```
data['categorical column'].describe()
```

For the `.describe()` function above:

- `count` This is the total number of non-null entries in the column.
- The `unique` Category, shows how many unique categories (distinct values) are present in that column.
- The `top` Category, is the most frequently occurring value (the mode) in the column.
- The `freq` Category, is the frequency (number of occurrences) of the most common value (shown in `top`)

```
range = data['column'].max() - data['column'].min()
range

#From Activity

np.mean()
np.median()
np.min()
np.max()
np.std()
```

Probability

Formulas

Classical Probability

$$\text{Classical Probability} = \frac{\text{Number of Desired Outcomes}}{\text{Total Number of Possible Outcomes}}$$

Empirical Probability

$$\text{Empirical Probability} = \frac{\text{Number of Times a Specific Event Occurs}}{\text{Total Number of Events}}$$

Three Concepts at the Foundation of Probability Theory:

- Random Experiment
- Outcome
- Event

Random Experiment: A process where outcomes cannot be predicted with certainty

All random experiments have three things in common:

- The Experiment can have more than one possible outcome
- You can represent each possible outcome in advance
- The outcome of the experiment depends on chance

The Probability of an Event

The probability that an event will occur is expressed as a number between O and 1. Probability can also be expressed as a percent.

- If the Probability of an event equals 0, there is a 0% chance that the event will occur.
- If the Probability of an event equals 1, there is a 100% chance that the event will occur.
- If the Probability of an event equals 0.5, there is a 50% chance that the event will occur - or not occur.
- If the Probability of an event is close to 0, there is a small chance that the event will occur.
- If the Probability of an event is close to 1, there is a strong chance that the event will occur.

Calculate the Probability of an Event

To calculate the probability of an event in which all possible outcomes are equally likely, you divide the number of desired outcomes by the total number of possible outcomes. You may recall that this is also the formula for classical probability:

$$\text{Probability of an Event} = \frac{\text{Number of desired outcomes}}{\text{Total Number of Possible Outcomes}}$$

Probability Notation

- The probability of event A is written as P(A).
- The probability of event B is written as P(B).
- For any event A, $0 \leq$ P(A) ≤ 1. In other words, the probability of any event A is always between 0 and 1.
- If P(A) $>$ P(B), then event A has a higher chance of occurring than event B.
- If P(A) = P(B), then event A and event B are equally likely to occur.
- P(A'), Probability of not event A.

Two events are **mutually exclusive** if they cannot occur at the same time.

Two events are **independent** if the occurrence of one event does not change the probability of the other event. This means that one event does not affect the outcome of the other event.

Three Basic Rules of Probability

- Complement Rule
- Addition Rule
- Multiplication Rule

Complement Rule *For Mutually Exclusive Events*

$$P(A') = 1 - P(A)$$

Addition Rule *For Mutually Exclusive Events*

$$P(AorB) = P(A) + P(B)$$

Multiplication Rule *For Independent Events*

$$P(AandB) = P(A) * P(B)$$

Conditional Probability: The Probability of an event occurring given that another event has already occurred.

Dependent Events: Two Events are dependent if the occurrence of one event changes the Probability of the other Event.

Conditional Probability Formula

$$P(AandB) = P(A) * P(B|A)$$

- $P(AandB)$: Probability of Event A and Event B
- $P(A)$: Probability of Event A
- $P(B|A)$: Probability of Event B given Event A

The vertical Bar (|) represents that Event B depends on Event A happening, we say this as the Probability of a B given A.

The formula can also be expressed as:

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

Bayes's Theorem

Used for determining Conditional Probability

Bayes Theorem provides a way to update the Probability of an event based on new information about the Event.

Prior Probability: The Probability of an Event before new data is collected.

Posterior Probability: The updated Probability of an Event based on new data.

Posterior means occurring after

Posterior Probability is calculated by updating the Prior Probability using Bayes Theorem

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

Bayes's theorem states that for any two events A and B, the probability of A given B equals the probability of A multiplied by the probability of B given A divided by the probability of B.

In the theorem, prior probability is the probability of event A. Posterior probability, or what you're trying to calculate, is the probability of event A given event B.

Sometimes, statisticians and data professionals use the term "likelihood" to refer to the probability of event B given event A, and the term "evidence" to refer to the probability of event B.

- $P(A)$: Prior Probability
- $P(A|B)$: Posterior Probability
- $P(B|A)$: Likelihood
- $P(B)$: Evidence

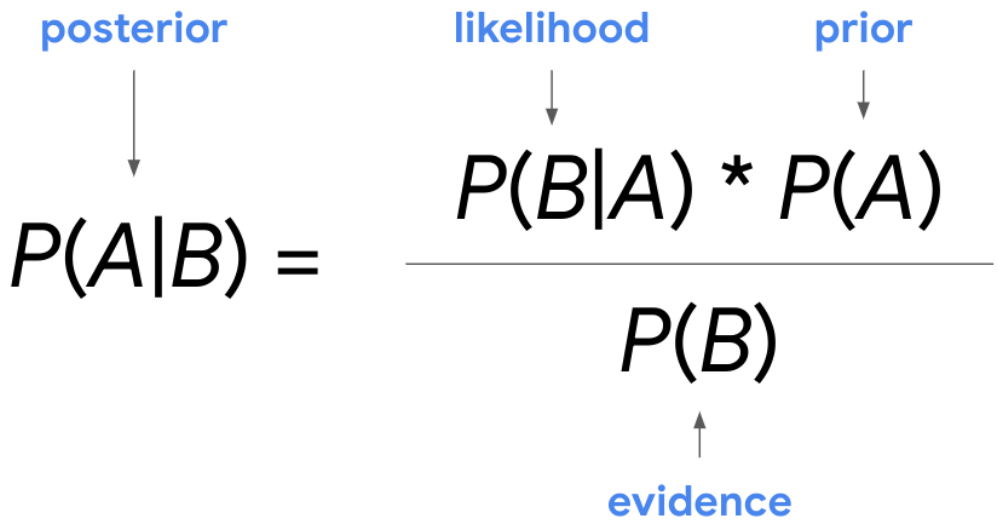


Figure 5: Bayes Theorem

Bayes's Theorem (Expanded Version)

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + P(B|notA) * P(notA)}$$

You can use the 2 versions of Bayes Theorem to deal with different types of problems.

For instance, if you don't know the Probability of Event B, in this case you can use the expanded version of Bayes Theorem, because you don't need to know the Probability of Event B.

Probability Distributions

Probability Distribution: Describes the likelihood of the possible outcomes of a Random Event.

Random Variable: Represents the values for the possible outcomes of a Random Event.

Random Variables

- Discrete
- Continuous

Discrete Random Variable: Has a countable number of possible values.

Often are whole numbers that can be counted

Continuous Random Variable: Takes all the possible values in some range of numbers

When your dealing with Continuous Variables you're dealing with decimal values rather than whole numbers.

Typically these are decimal values that can be measured such as height, weight, time or temperature.

Discrete or Continuous Variables

- Count the number of outcomes = Discrete
- Measure the outcome = Continuous

Discrete Distributions represent Discrete Random Variables

Continuous Distributions represent Continuous Random Variables

- Discrete Distributions can be represented as Tables or Histograms. With the Variable your counting on the X-Axis and the Probability on the Y-Axis.
- Continuous Distributions are represented in intervals and are represented as a Curve (or Bell Curve) with different intervals. X-Axis has the variable your measuring in Intervals (e.g. 10-17), and the Y-Axis has the Probability Density, which is not the same as Probability and is a Statistical Function.

Binomial Distribution

Binomial Distribution: A Discrete Distribution that Models the Probability of Events with only two possible outcomes, success or failure. These outcomes are Mutually Exclusive and cannot occur at the same time.

A Binomial Experiment has the following attributes: * The Experiment consists of a number of repeated trials. * Each trial has only only two possible outcomes. * The Probability of success is the same for each trial. * Each trial is independent.

Binomial Distribution Formula

$$p(x = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

k = Number of Successes

n = Number of Trials

p = The Probability of Success on a given Trial

$(n - k)$ = "n-choose-k" number of ways to obtain k successes in n trials

Can use a Histogram to visualise the Binomial Distribution

- X-Axis = Random Variable
- Y-Axis = Probability

Poisson Distribution

Poisson Distribution: Models the Probability that a certain number of Events will occur during a specific time period.

Can also be used to represent the number of events that occur in a specific space, such as a distance, area or volume.

The Poisson Distribution represents a type of Random Experiment called a Poisson Experiment. A Poisson Experiment has the following attributes:

- The number of Events in the Experiment can be counted.
- The mean number of Events that occur during a specific time period is known.
- Each Event is independent.

Poisson Distribution Formula

$$p(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

λ = The Mean number of Events that occur during a specific time period.

k = Number of Events

e = Constant equal to approximately 2.71828

! = Stands for Factorial. A function that multiplies a number by every whole number below it down to 1. e.g. 2 factorial is 2×1.

Can use a Histogram to visualise the Poisson Distribution

- X-Axis = Number of Events
- Y-Axis = Probability

Uniform Distribution

Uniform Distribution: Describes Events whose outcomes are all equally likely, or have equal Probability.

Can Visualise the Uniform Distribution with a Histogram

- X-Axis = Random Variable
- Y-Axis = Probability

Bernoulli Distribution

Bernoulli Distribution: Models Events that only have 2 possible outcomes (success or failure). Refers to only a single trial of an experiment.

Can be visualised with a Histogram

- X-Axis = Random Variable
- Y-Axis = Probability

Probability Density and Probability

A probability function is a mathematical function that provides probabilities for the possible outcomes of a random variable.

There are two types of probability functions:

- Probability Mass Functions (PMFs) represent Discrete Random Variables
- Probability Density Functions (PDFs) represent Continuous Random Variables

A probability function can be represented as an equation or a graph. The math involved in probability functions is beyond the scope of this course. For now, it's important to know that the graph of a PDF appears as a curve. You've learned about the bell curve, which refers to the graph for a normal distribution.

The Normal Distribution

Normal Distribution: A Continuous Probability Distribution that is symmetrical on both sides of the Mean and bell-shaped.

It is often called the Bell Curve

It is also known as the Gaussian Distribution

The most common Distribution in Statistics.

Normal Distribution have the following features:

- The shape is a Bell Curve
- The Mean is located at the centre of the Curve
- The Curve is symmetrical on both sides of the centre
- The total area under the Curve equals to 1

Under a Normal Distribution, the distance of a data point from the Mean is often measured in Standard Deviations.

The Values along a Curve are distributed in a regular pattern based on their distance from the Mean, this is known as the Empirical Rule

The Empirical Rule

- 68% of Values fall within 1 Standard Deviation of the Mean.
- 95% of Values fall within 2 Standard Deviations of the Mean.
- 99.7% of Values fall within 3 Standard Deviations of the Mean.

The Empirical Rule is useful for estimating data, especially for large datasets like height and weight data for an entire population.

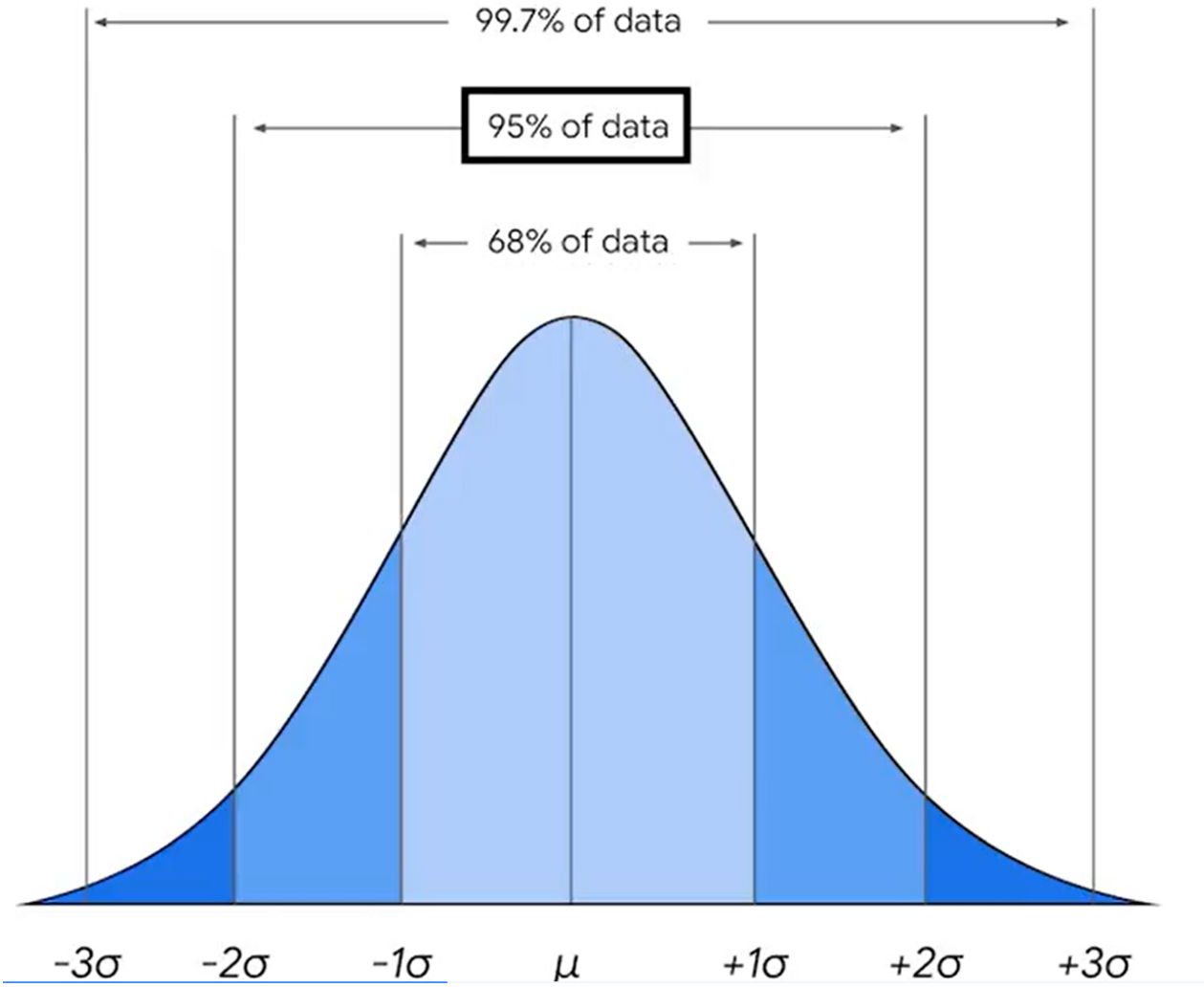


Figure 6: The Empirical Rule

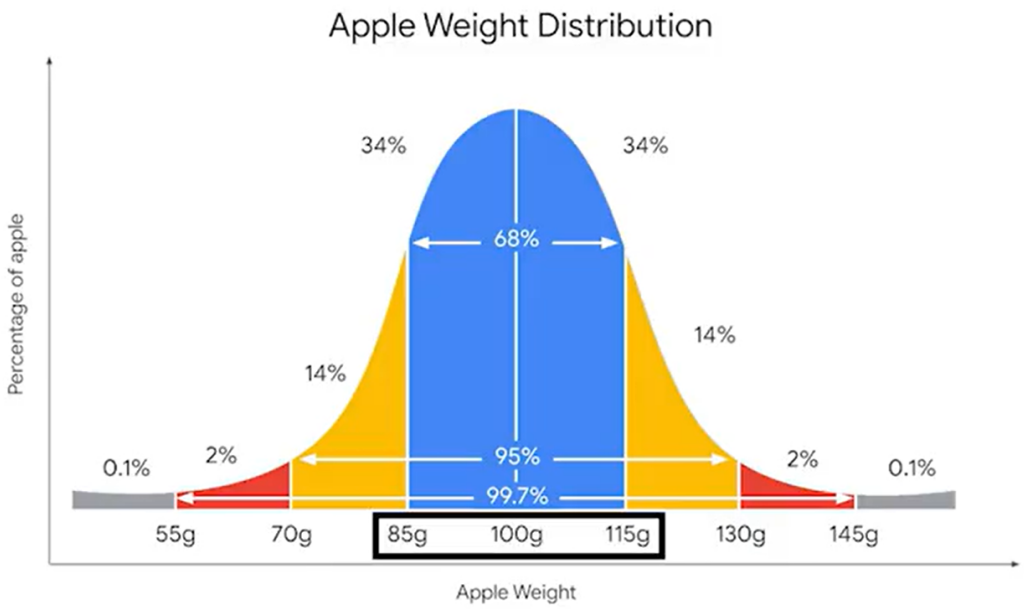


Figure 7: E.g. of the Empirical Rule

You can use the Empirical Rule to get an estimate of the Distributions of values in your dataset, such as what percentage of values will fall within 1, 2 or 3 Standard Deviations of the Mean. This saves time and helps you better understand your data.

Knowing the location of your values on a Normal Distribution is useful for detecting outliers

Typically, data professionals consider values that lie more than 3 Standard Deviations below or above the Mean to be Outliers.

Z-Scores

Z-Score: A Measure of how many Standard Deviations below or above the population Mean a data point is.

Gives you an idea of how far from the mean a data point is

- Z-Score is 0, value is equal to the Mean
- Z-Score is Positive, value is greater than the Mean
- Z-Score is Negative, value is less than the Mean

Z-Scores are also called Standard Scores because they're based on what's called the Standard Normal Distribution

A Standard Normal Distribution is just a Normal Distribution with a Mean of 0 and a Standard Deviation of 1

Z-Scores typically range from -3 to 3

Standardisation: The process of putting different variables on the same scale.

Standardisation is useful because it let's you compare scores from different datasets that may have different units, Mean values and Standard Deviations

Data professionals use Z-Scores to better understand the relationship between data values within a single dataset and between different datasets.

Z-Score Formula

$$Z = \frac{x - \mu}{\sigma}$$

x = Single data Value or Raw Score

μ = Population Mean

σ = Population Standard Deviation

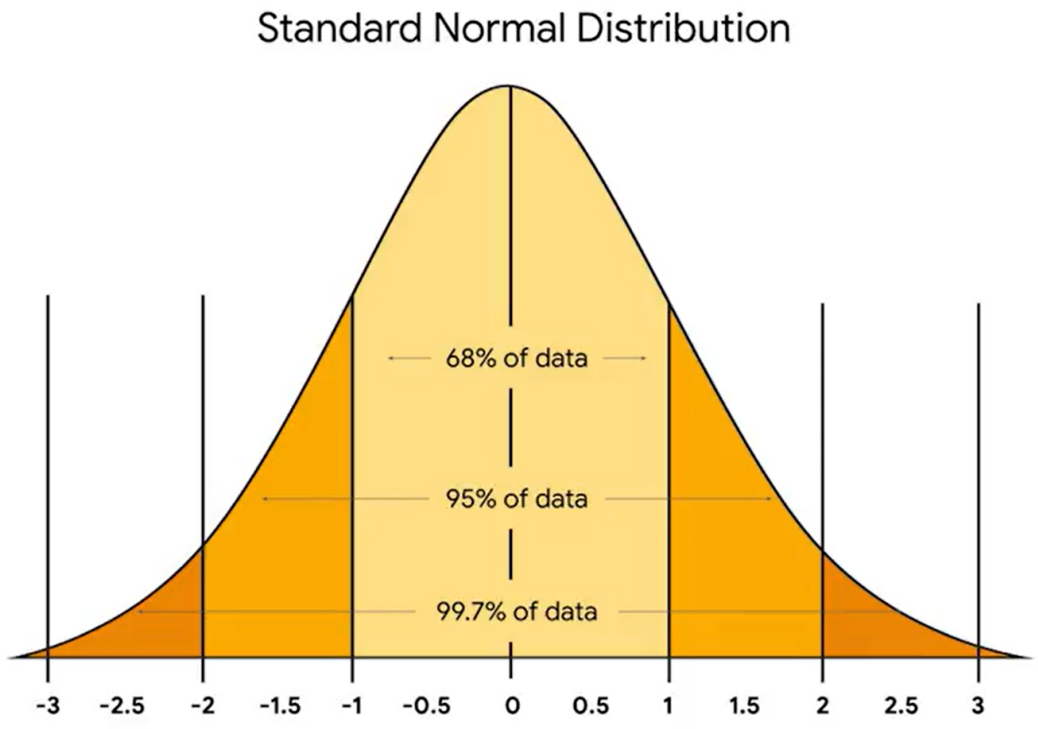


Figure 8: Standard Normal Distribution

Probability Distributions in Python

Packages

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
import statsmodels.api as sm
```

The first step is to find out what type of Probability Distribution your data is. To do this we will plot a histogram

```
data = pd.read_csv('data.csv')

data['column'].hist()
```

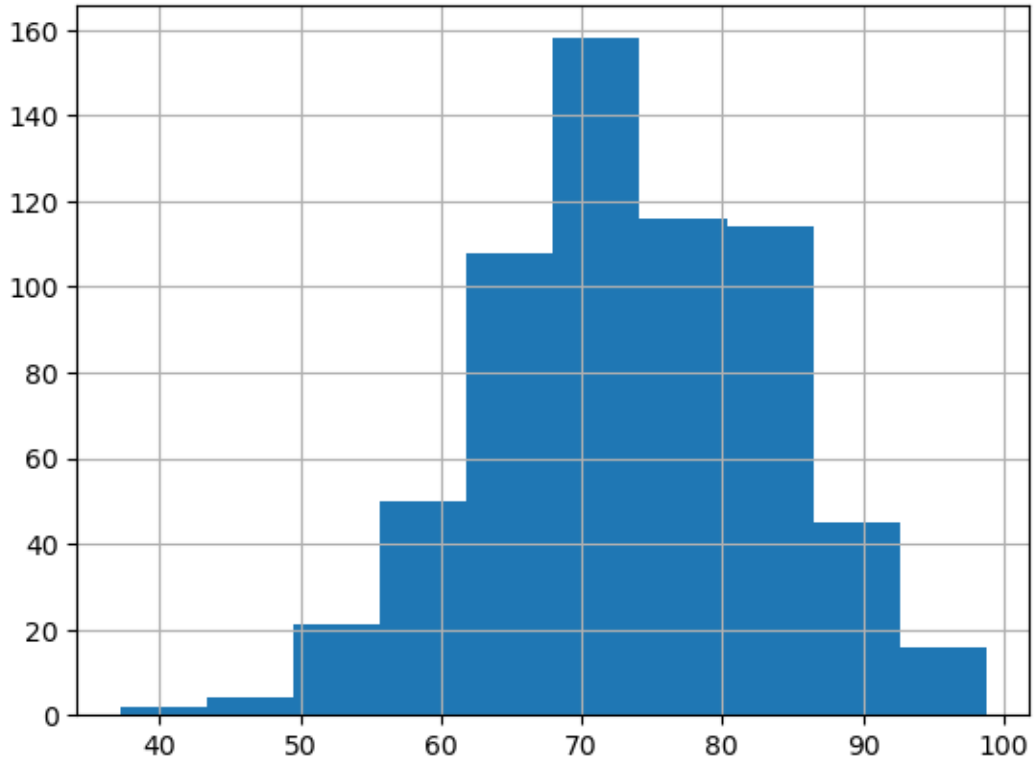


Figure 9: Python Distribution

Looks Normally Distributed

In order to verify if the data is Normally Distributed we have to see if it follows the Empirical Rule.

1 std dev of the Mean

Mean

```
mean_data = data['column'].mean()
```

Standard Deviation

```
std_data = data['column'].std()
```

Computing % of data that fall within 1 std dev of the Mean

```
upper_limit = mean_data + 1 * std_data
lower_limit = mean_data - 1 * std_data

((data['column'] >= lower_limit) & (data['column'] <= upper_limit)).mean()
```

- `lower_limit` will be 1 Standard Deviation below the Mean
- `upper_limit` will be 1 Standard Deviation above the Mean
- The Boolean Mask, tells the computer to decide if each value in the `column` is between the `lower_limit` and `upper_limit`.
- In other words to decide if each value is greater than or equal to 1 Standard Deviation below the mean and less than or equal to standard deviation above the mean.
- Use the `.mean()` function to divide the number of values that are within 1 standard deviation of the mean by the total number of values.

2 std dev of the Mean

```
upper_limit = mean_data + 2 * std_data
lower_limit = mean_data - 2 * std_data

((data['column'] >= lower_limit) & (data['column'] <= upper_limit)).mean()
```

3 std dev of the Mean

```
upper_limit = mean_data + 3 * std_data
lower_limit = mean_data - 3 * std_data

((data['column'] >= lower_limit) & (data['column'] <= upper_limit)).mean()
```

Then check if these values are close or equal to the empirical rule.

Outlier Detection with Z-Scores

First create a new column in your data called `Z-SCORE` that includes the Z-Score for each value in your original column.

Then use the `stats.zscore()` function to compute the Z-Score.

```
data['Z-SCORE'] = stats.zscore(data['column'])
data
```

Now write some code to identify the outliers with Z-Score that are greater or less than 3 Standard Deviations from the Mean.

Use the Relational Operators `>` greater than, `<` less than and the Bitwise Operator `|`.

```
data[(data['Z-SCORE'] > 3) | (data['Z-SCORE'] < -3)]
```

You'll then be able to see the outliers.

Sampling

Sampling: The process of selecting a subset of data from a population.

Representative Sample: Accurately reflects the characteristics of a population

The Sampling Process

Step 1: Identify the target Population

Step 2: Select the Sampling Frame

Step 3: Choose the Sampling Method

Step 4: Determine the Sample Size

Step 5: Collect the Sample Data

Target Population: The complete set of elements that you're interested in knowing more about.

Sampling Frame: A list of all the items in your Target population.

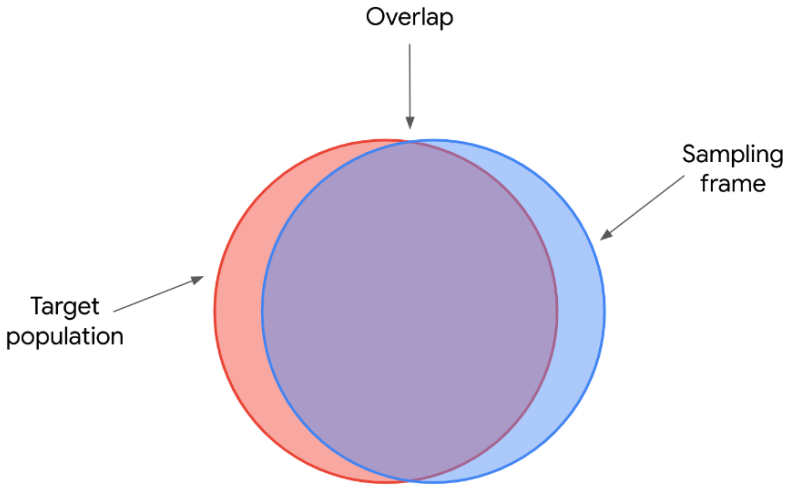


Figure 10: Sampling Frame

2 Main Sampling Methods:

- Probability Sampling
- Non-Probability Sampling

Probability Sampling: Uses Random selection to generate a Sample

Non-Probability Sampling: Based on convenience or personal preference.

Sample Size: The number of individuals or items chosen for a study or experiment

Sample Size helps determine the accuracy of the predictions you make about the population

The larger the Sample Size the more accurate your predictions

Effective Sampling ensures that your Sample data is representative of your Target population. Then when you use Sample data to make inferences about the population, you can be reasonably confident that your inferences are reliable.

Decisions you make in each step of the process can affect the quality of your sample data

Probability Sampling Methods

Simple Random Sampling: Every member of a population is selected randomly and has an equal chance of being chosen

You can Randomly select members using a Random Number Generator or by another method of Random Selection.

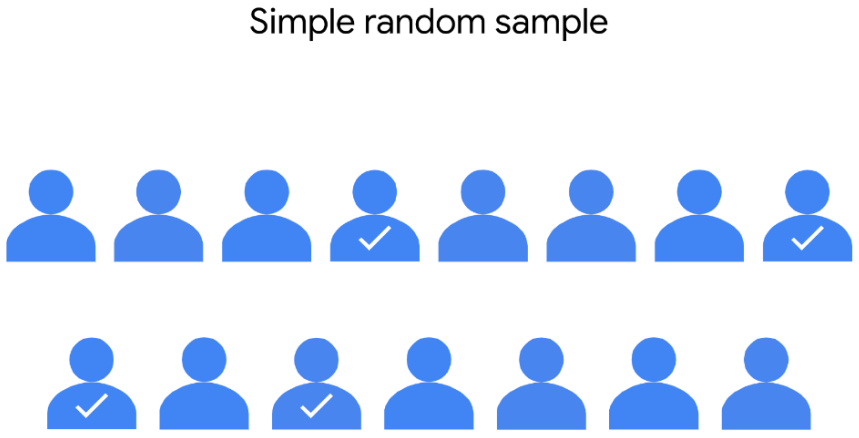


Figure 11: Simple Random Sample

Stratified Random Sampling: Divide a population into groups and randomly select some members from each group to be in the Sample

These groups are called **Strata**

Strata can be organised by age, gender, income or whatever Category your interested in studying.

Helps ensure that members from each group are included

Disadvantage: It can be difficult to identify appropriate strata for a study if you lack knowledge of a population

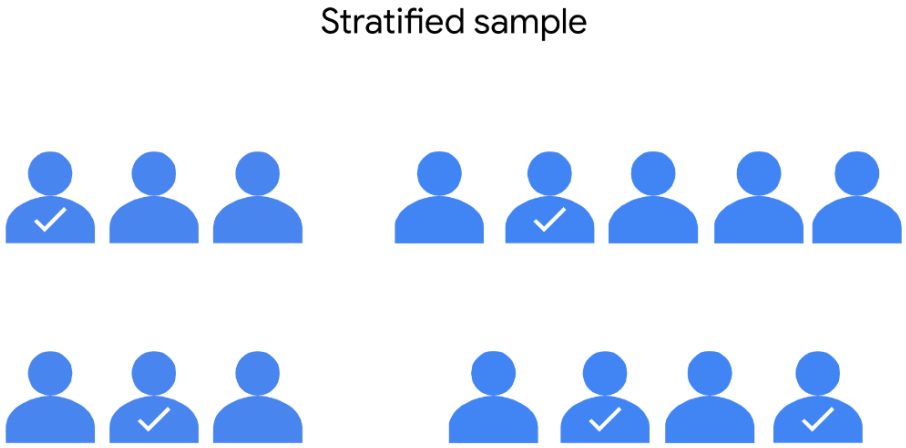


Figure 12: Stratified Random Sample

Cluster Random Sampling: Divide a population into Clusters, randomly select certain Clusters, and include all members from the chosen Clusters in the Sample.

Clusters are divided using identifying details such as age, gender, location, or whatever you want to study

Helpful when dealing with large and diverse populations that have clearly defined subgroups.

Disadvantage: May be difficult to create Clusters that accurately reflect the overall population

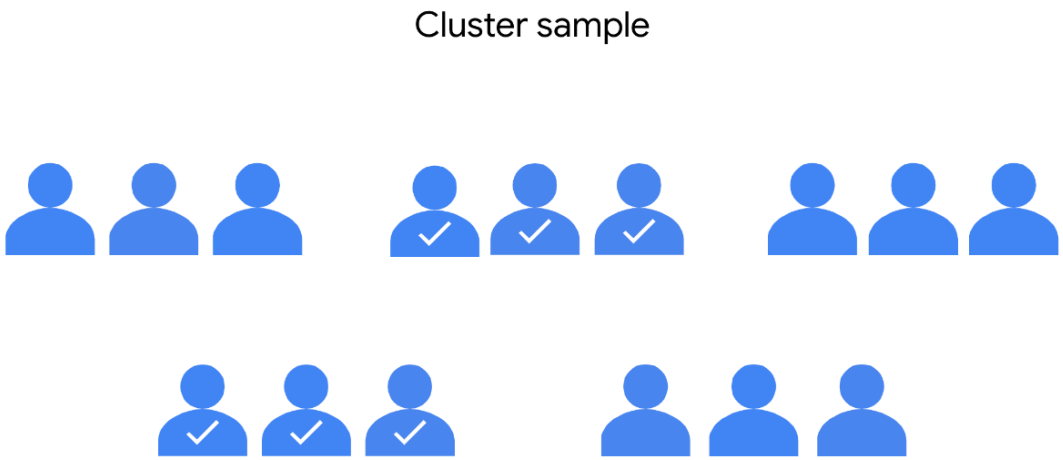


Figure 13: Cluster Random Sample

Systematic Random Sampling: Put every member of a population into an ordered sequence. Then, you choose a Random starting point in the sequence and select members for your Sample at regular intervals.

Disadvantage: You need to know the size of the population that you want to study before you begin. If you don't have this info its difficult to choose consistent intervals.

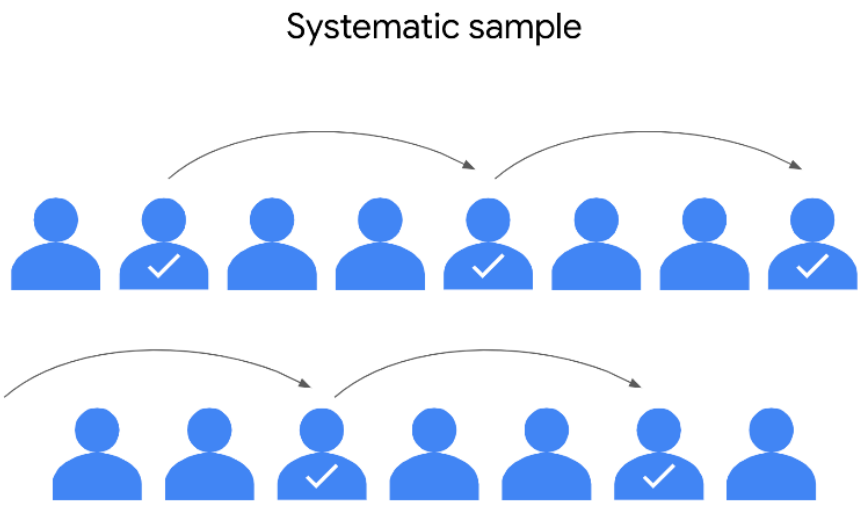


Figure 14: Systemic Random Sample

Non-Probability Sampling Methods

Sampling Bias: When a Sample is not representative of the Population as a whole.

Non-Probability Sampling Methods result in Bias.

Often less expensive and more convenient for researchers to conduct.

Sometimes due to budget, time or other reasons, its just not possible to use Probability Sampling.

Can be useful for exploratory statistics, which seek to develop initial understanding of a population, not draw conclusions or make predictions about the population as a whole.

Convenience Sampling: Choose members of a population that are easy to contact or reach.

Involves collecting a Sample from somewhere convenient to you such as your workplace, a local school or a public park.

Because these Samples are based on convenience to the researcher and not a broader Sample of the Population, Convenience Samples often show Undercoverage Bias.

Undercoverage Bias: When some members of a population are inadequately represented in a Sample.

Voluntary Response Sampling: Consists of members of a population who volunteer to participate in a Study.

Tend to suffer from Nonresponse Bias

Nonresponse Bias: When certain groups of people are less likely to provide responses.

Snowball Sampling: Researchers recruit initial participants to be in a study and then asks them to recruit other people to participate in the study.

Like a Snowball, the Sample Size gets bigger and bigger as more participants join in.

Purposive Sampling: Researchers select participants based on the purpose of their study.

Because of this, applicants who do not fit the profile are rejected.

The Researcher often intentionally excludes certain groups from the Sample to focus on a specific group they think is most relevant to their study.

Sampling Distributions

Point Estimate: Uses a Single Value to estimate a population Parameter.

Sampling Distributions: A Probability Distribution of a Sample Statistic

Sample statistics are based on randomly sampled data, and their outcomes cannot be predicted with certainty. You can use a sampling distribution to represent statistics such as the mean, median, standard deviation, range, and more.

Typically, data professionals compute sample statistics like the mean to estimate the corresponding population parameters.

Sampling Distributions represent the possible outcomes for a Sample Statistic like the mean.

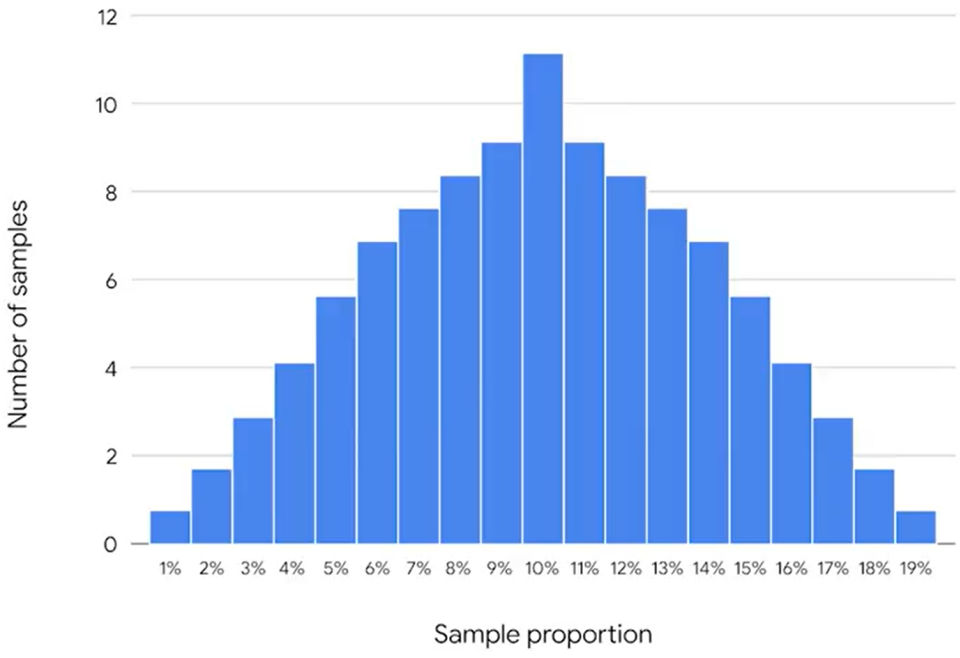


Figure 15: Sampling Distribution

Sampling Variability: How much an estimate varies between Samples.

You can use a Sampling Distribution to represent the frequency of all your different Sample Means

If your Sample Size is large enough, your Sample Mean will roughly equal the population Mean.

Standard Error of the Mean

To measure Sampling Variability, data professionals use the Standard Deviation of Sample Means to measure this Variability.

In Statistics the Standard Deviation of a Sample Statistic is called the Standard Error.

The Standard Error of the Mean measures the Variability among all your Sample Means.

- Larger Standard Error = Sample Means are more Spread out (More Variability)
- Smaller Standard Error = Sample Means are closer together (Less Variability)

The less Standard Error the more likely it is that your Sample Mean is an accurate estimate of the population Mean.

Note that the concept of Standard Error is based on the practice of repeated Sampling. In reality, researchers usually work with a single sample, its often too complicated, expensive or time consuming to take repeated Samples of a population. Instead Statisticians have derived a formula for calculating the Standard Error based on the Mathematical assumption of repeated Sampling.

Formula

$$SE = \frac{S}{\sqrt{n}}$$

S = The Sample Standard Deviation n = The Sample Size

As your Sample Size gets larger, your Standard Error gets smaller

This is because Standard Error measures the difference between your Sample Mean and the actual population Mean.

The Central Limit Theorem

Central Limit Theorem: The Sampling Distribution of the Mean approaches a Normal Distribution as the Sample Size increases.

In other words: As your Sample increases your Sampling Distribution assumes the shape of a Bell Curve, and if you take a large enough Sample of the population, the Sample Mean will be roughly equal to the population Mean.

There is no exact rule for how large a Sample Size needs to be in order for the Central Limit Theorem to apply. In general, a Sample Size of 30 or more is considered sufficient.

The Central Limit Theorem holds true for any population

You don't need to know the shape of your population Distribution in advanced in order to apply the Theorem.

If you collect a large enough Sample, the shape of your Sampling Distribution will follow a Normal Distribution.

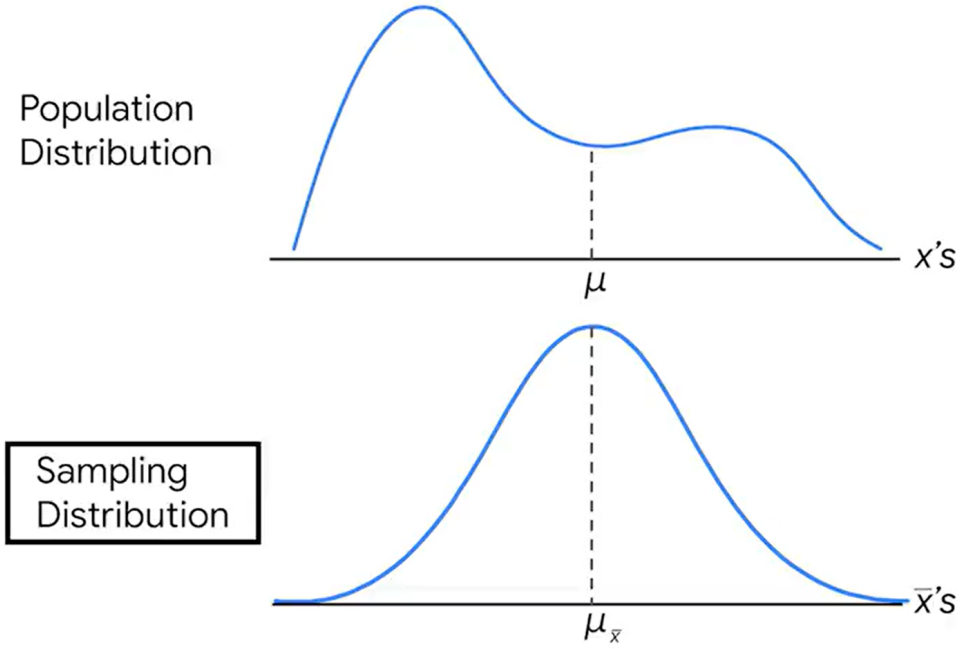


Figure 16: Central Limit Theorem Distribution

The Sampling Distribution of the Proportion

Population Proportions: The percentage of individuals or elements in a population that share a certain characteristic.

- Proportion's measure percentages or parts of a whole.
- There is also Variability for Proportions just like with the Mean.

You can use a Sampling Distribution to represent the frequency of all your different Sample Proportions.

The Central Limit Theorem also applies to Proportions.

As your Sample Size increases, the distribution of the Sample Proportion will be approximately normal. The overall average or Mean Proportion is located in the centre of the Curve.

Standard Error of the Proportion

You can Use the Standard Error of the Proportion to measure Sampling Variability.

This tells you how much a particular Sample Proportion is likely to differ from the true population Proportion.

The more Variability in your Sample Data the less likely it is that the Sample Proportion is an accurate estimate of the population Proportion

Formula

$$SE(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{n}$$

$\hat{p}$ = Population Proportion

n = Sample Size

Standard Error measures the difference between your Sample Proportion and the true population Proportion

As your Sample gets larger, your Sample Proportion gets closer to the true population Proportion.

The more accurate the estimate of the population Proportion, the smaller the Standard Error.

Typically the next step for a data Professional would be to use the Standard Error to construct a Confidence Interval. This describes the uncertainty of your estimate and gives your Stakeholders more detailed information about your results.

Sampling Distributions with Python

Packages

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
import statsmodels.api as sm
```

.sample()

Generating a Random Sample

Syntax * `n` refers to the desired Sample Size * `replace` indicates whether you are Sampling with or without replacement * `random_state` refers to the seed of the Random Number

Sampling with Replacement: When a population element can be selected more than one time

Sampling without Replacement: When a population element can be selected only one time

For example, suppose you have a jar that contains 100 unique numbers from 1-100, you want to select a random sample of numbers from the jar. After you pick the number from the jar, you can put the number aside or you can put it back in the jar. If you put the number back in the jar, it may be selected more than once **this is sampling with replacement**. If you put the number aside it can be selected only one time **this is sampling without replacement**. For the purpose of our example **you will sample with replacement**.

Random Seed: A starting point for generating random numbers.

You can use any arbitrary number to fix the random seed and give the random number generator a starting point.

Also going forward you can use the same random seed to generate the same set of numbers, In a later video you'll work with a sample again.

```
data = pd.read_csv('data.csv')

sampled_data = data.sample(n=50, replace=True, random_state=31208)
sampled_data
```

Sample Mean

```
estimate1 = sampled_data['column'].mean()
estimate1
```

This is a Point Estimate based on the population Mean based on your Random Sample

Due to Sampling Variability, the Sample Mean is usually not exactly the same as the population Mean.

Generating a 2nd Random Sample

```
estimate2 = data['column'].sample(n=50, replace=True, random_state=56910).mean()
estimate2
```

Due to Sampling Variability, this Sample Mean is different from the Sample Mean of your previous estimate (estimate1) but there really close.

Recall that the Central Limit Theorem tells you that when the sample size is large enough, the sample mean approaches a normal distribution. And as you sample more observations from a population, the sample mean gets closer to the population mean.

The larger your sample size, the more accurate your estimate of the population mean is likely to be

Now imagine you repeated this study 10,000 times and obtained 10,000 point estimates of the mean

According to the Central Limit Theorem the mean of your sampling distribution will be roughly equal to the population mean

You can use Python to compute the mean of the sampling distribution with 10,000 samples

Computing the Mean of the Sampling Distribution of 10,000 Samples

Lets review the code step by step

- First create an empty list to store the sample mean from each sample, name this `estimate_list`
- Set up a `for` loop with the `range` function, the loop will run 10,000 times and iterate over each number of the sequence
- Specify what you want to in each iteration of the loop. The `.sample()` function takes a random sample of `50` with replacement. The `.append()` function adds a single item to the existing list, in this case it appends the value of the sample mean to each item in the list.
- Create a new dataframe for your list of 10,000 estimates
- Name a new variable `estimate_df` to store your dataframe
- Name a new variable `mean_sample_means` then create the mean for your sampling distribution of 10,000 random samples

```
estimate_list = []
for i in range(10000):
    estimate_list.append(data['column'].sample(n=50, replace=True).mean())
estimate_df = pd.DataFrame(data={'estimate': estimate_list})
```

```
mean_sample_means = estimate_df['estimate'].mean()
mean_sample_means
```

The Mean of your Sampling Distribution is essentially identical to your population Mean of your complete dataset.

Visualisation of Sampling Distribution of 10,000 Estimates

To visualise the relationship between your sampling distribution of 10,000 estimates and the normal distribution we can plot both at the same time.

```
plt.hist(estimate_df['estimate'], bins=25, density=True, alpha=0.4, label = "histogram of sampl
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100) # generate a grid of 100 values from xmin to xmax.
p = stats.norm.pdf(x, mean_sample_means, stats.tstd(estimate_df['estimate']))
plt.plot(x, p,'k', linewidth=2, label = 'normal curve from central limit theorem')
plt.axvline(x=mean_sample_means, color='g', linestyle = 'solid', label = 'population mean')
plt.axvline(x=estimate1, color='r', linestyle = '--', label = 'sample mean of the first random
plt.axvline(x=mean_sample_means, color='b', linestyle = ':', label = 'mean of sample means of 1
plt.title("Sampling distribution of sample mean")
plt.xlabel('sample mean')
plt.ylabel('density')
plt.legend(bbox_to_anchor=(1.04,1))
plt.show()
```

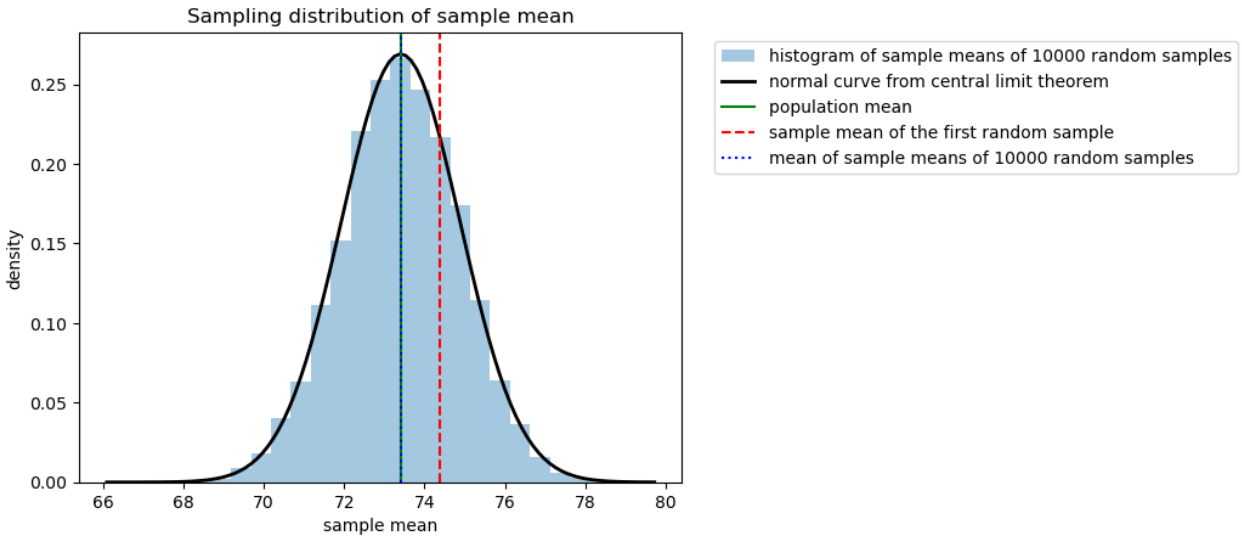


Figure 17: Sampling Distribution of Sample Mean

Activity: Calculating the Standard Error

```
standard_error = data['column'].std() / np.sqrt(len(data))
```

Confidence Intervals

Confidence Interval: A range of values that describes the uncertainty surrounding an Estimate.

Data professionals use Confidence Intervals as part of their job

You may be asked about Confidence Intervals in a job interview

Interval Estimate: Uses a range of values to estimate a population Parameter.

"A Point Estimate is useful, but a single element like 30lbs does not express the uncertainty built into any estimate. This uncertainty is due to the method of Random Sampling"

Confidence Intervals give data professionals a way to express the uncertainty caused by randomness and provide a more reliable Estimate. Confidence Level: Describes the likelihood that a particular Sampling Method will produce a Confidence Interval that includes the population Parameter.

e.g. Say you use a 95% Confidence Level to calculate a Confidence Interval between 28 and 32lbs.

This means if you took 100 Random Samples from the Penguin population and calculated a 95% Confidence Interval for each Sample, then approximately 95 of the 100 Intervals or 95% of the total would contain the actual population Mean. One such Interval will be the range of values between 28 and 32lbs.

Common Confidence Levels: * 90% * 95% * 99%

95% is a popular choice that is based on tradition in Statistical research and education, you can adjust the Confidence Level to meet the requirements of your analysis.

95% Confidence Level = Means that if you take repeated Random Samples from a population and construct a Confidence Interval for each Sample using the same Method. You can expect: - 95% of these Intervals capture the population Mean - 5% of the Intervals do not capture the population Mean

In practice, data professionals usually select one Random Sample and generate 1 Confidence Interval which may or may not contain the actual Mean. This is because repeated Random Sampling is often difficult, expensive and time consuming.

Confidence Intervals give data professionals a way to quantify the uncertainty due to Random Sampling.

In relation to an example "In other words, this method will produce an Interval that contains the population Mean with a success rate of 95%. That's a pretty good success rate." ## Correct Interpretation *e.g. Mean Weight*

Let's explore an example to get a better understanding of how to interpret a confidence interval. Imagine you want to estimate the mean weight of a population of 10,000 penguins. Instead of

weighing every single penguin, you select a sample of 100 penguins. The mean weight of your sample is 30 pounds. Based on your sample data, you construct a 95% confidence interval between 28 pounds and 32 pounds.

95% CI [28, 32]

Interpret the confidence interval

Earlier, you learned that the confidence level expresses the uncertainty of the estimation process. Let's discuss what 95% confidence means from a more technical perspective.

Technically, 95% confidence means that if you take repeated random samples from a population, and construct a confidence interval for each sample using the same method, you can expect that 95% of these intervals will capture the population mean. You can also expect that 5% of the total will not capture the population mean.

The confidence level refers to the long-term success rate of the **method**, or the estimation process based on random sampling.

For the purpose of our example, let's imagine that the mean weight of all 10,000 penguins is 31 pounds, although you wouldn't know this unless you actually weighed every penguin. So, you take a sample of the population.

Imagine you take 20 random samples of 100 penguins each from the penguin population, and calculate a 95% confidence interval for each sample. You can expect that approximately 19 of the 20 intervals, or 95% of the total, will contain the actual population mean weight of 31 pounds. One such interval will be the range of values between 28 pounds and 32 pounds.

Figure 18: Confidence Intervals

In practice, data professionals usually select one random sample and generate one confidence interval, which may or may not contain the actual population mean. This is because repeated random sampling is often difficult, expensive, and time-consuming. Confidence intervals give data professionals a way to quantify the uncertainty due to random sampling.

Common Misconceptions

1. A 95% Confidence Interval means that 95% of all the data values in your dataset fall within the Interval
2. A 95% Confidence Interval implies that 95% of all possible Sample Means fall within the range of the Interval
3. A Confidence Interval refers to the only possible source of error in your results

Note: When your interpreting a Confidence Interval remember that the uncertainty lies in an Estimation process based on Random Sampling. A 95% Confidence Level refers to the success rate of that process. In other words you can expect 95% of the Random Intervals you generate to capture the population Parameter.

The Confidence Level refers to the long-term success rate of the method or estimation process based on Random Sampling.

Pro-tip: Remember that a 95% Confidence Level refers to the success rate of the estimation process.

Steps to Constructing a Confidence Interval

1. Identify a Sample Statistic
2. Choose a Confidence Level
3. Find the Margin of Error
4. Calculate the Interval

Interval = Sample Statistic +/- Margin of Error

Margin of Error: The range of values above and below the Sample Statistic

The maximum expected difference between a population Parameter and a Sample Estimate

This is the amount that a data professional expects their estimate might vary from their actual amount.

$$\text{MoE} = \text{Z-Score} * \text{SE}$$

Confidence Level	Z-Score
90%	1.645
95%	1.96
99%	2.58

Identify a Sample Statistic

Mean or Proportion

e.g. Mean

Choose a Confidence Level

Confidence Level = 90%, 95%, 99%

e.g. 95%

Find the Margin of Error

$$\text{MoE} = \text{Z-Score} * \text{SE}$$

Standard Error for Means

$$SE = \frac{\sigma}{\sqrt{n}}$$

σ = The Proportion Standard Deviation when known, otherwise S. S = Sample Standard Deviation.
n = The Sample Size

Standard Error for Proportions

$$SE(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{n}$$

$\hat{p}$ = population Proportion

n = Sample Size

e.g. Mean

Confidence Level = 95% = Z-Score: **1.96**

$$SE = \frac{1.5}{\sqrt{100}} = 0.15$$

σ = std of population = 1.5

n = 100

$$\text{MoE} = \text{Z-Score} * \text{SE} = 1.96 * 0.15 = 0.294$$

Calculate the Interval

$$\text{Upper Limit} = \text{Sample Statistic} + \text{Margin of Error}$$

$$\text{Lower Limit} = \text{Sample Statistic} - \text{Margin of Error}$$

e.g. Mean

Sample Mean = 20.5 hrs

Sample Std = 1.7 hrs

Population std = 1.5 hrs

$$\text{Upper Limit} = \text{Sample Statistic} + \text{Margin of Error}$$

$$\text{Upper Limit} = 20.5 + 0.294 = 20.794 = 20:48(\text{hrs:min})$$

$$\text{Lower Limit} = \text{Sample Statistic} - \text{Margin of Error}$$

$$\text{Lower Limit} = 20.5 - 0.294 = 20.206 = 20:12(\text{hrs:min})$$

[20:12, 20:48]

95% CI [20:12, 20:48]

As the Confidence Level gets higher, the Confidence Interval gets wider

e.g. 99% Confidence Level

99% CI [20:07, 20:53]

This is because the wider Confidence Interval is more likely to include the actual population Parameter.

Note: For Calculating a Confidence Interval for a Proportion: As your Sample Size gets larger, your Confidence Interval gets narrower - As your Sample Size increases, your Margin of Error decreases.

Steps for Constructing a Confidence Interval of a Small Sample Size

Small Sample: T-Scores

For small sample sizes, you need to use a different distribution, called the t-distribution. Statistically speaking, this is because there is more uncertainty involved in estimating the standard error for small sample sizes. You don't need to worry about the technical details, which are beyond the scope of this course. For now, just know that if you're working with a small sample size, and your data is approximately normally distributed, you should use the t- distribution rather than the standard normal distribution. For a t-distribution, you use t-scores to make calculations about your data.

The graph of the t-distribution has a bell shape that is similar to the standard normal distribution. But, the t-distribution has bigger tails than the standard normal distribution does. The bigger tails indicate the higher frequency of outliers that come with a small dataset. As the sample size increases, the t-distribution approaches the normal distribution. When the sample size reaches 30, the distributions are practically the same, and you can use the normal distribution for your calculations.

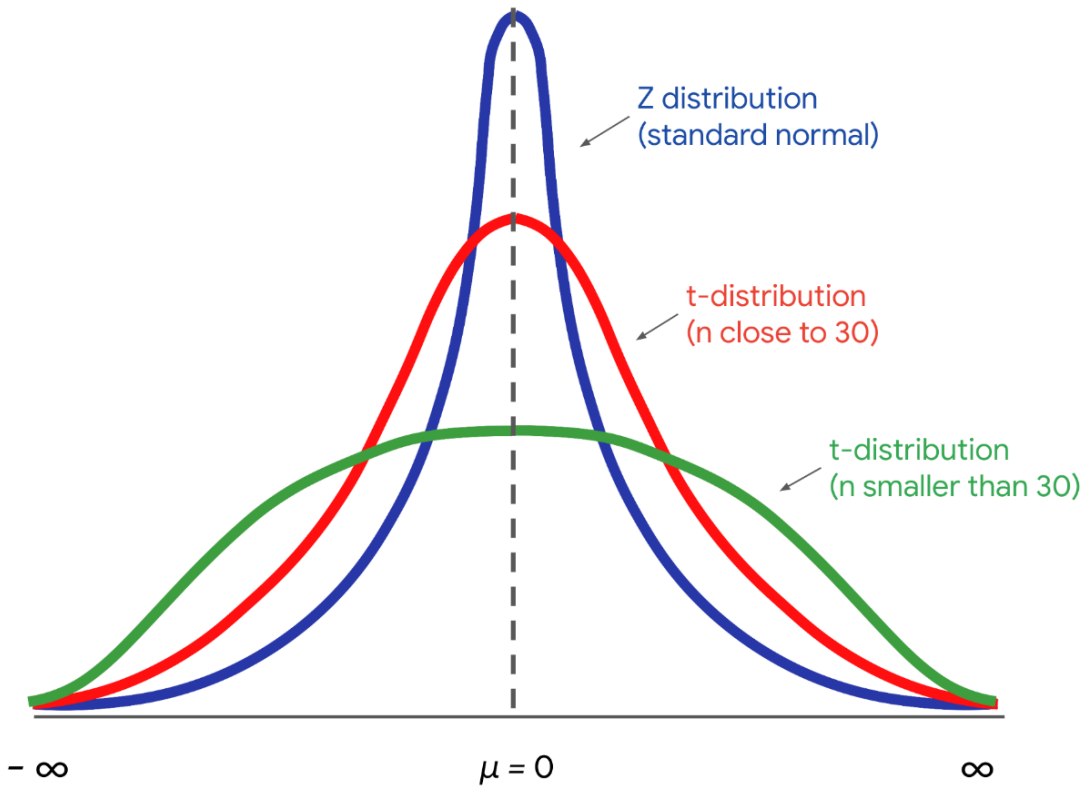


Figure 19: z and t Distribution

Construct the Confidence Interval

Will follow an example from Certificate reading, if you want full details of the example then go to the reading: Construct a Confidence Interval for a Small Sample Size

Step 1: Identify the Sample Statistic

First, identify your sample statistic. Your sample represents the average emissions rate for 15 engines. You're working with a sample mean.

Step 2: Choose a Confidence Level

Next, choose a confidence level. The engineering team requests that you choose a 95% confidence level.

Step 3: Find the Margin of Error

Your third step is to find the margin of error. For a small sample size, you calculate the margin of error by multiplying the t-score by the standard error.

The t-distribution is defined by a parameter called the degree of freedom. In our context, the degree of freedom is the sample size - 1, or 15-1 = 14. Given your degree of freedom and your confidence level, you can use a programming language like Python or other statistical software to calculate your t-score.

Based on a degree of freedom of 14, and a confidence level of 95%, your **t-score is 2.145.**

Now you can calculate the standard error, which measures the variability of your sample statistic.

Here's the formula for the standard error of the mean that you've used before:

Standard Error (Means)

$$SE(x) = \frac{S}{\sqrt{(n)}}$$

In the formula, the letter s refers to sample standard deviation, and the letter n refers to sample size.

Your sample standard deviation is 35, and your sample size is 15. The calculation gives you a **standard error of about 9.04.**

The margin of error is your t-score multiplied by your standard error. This is **2.145 * 9.04 = 19.39**.

Step 4: Calculate the Interval

Finally, calculate your confidence interval. The upper limit of your interval is the sample mean plus the margin of error. This is 430 + 19.39 = 449.39 grams of CO2 per mile.

The lower limit is the sample mean minus the margin of error. This is 430 - 19.39 = 410.61 grams of CO2 per mile.

You have a 95% confidence interval that stretches from 410.61 grams of CO2 per mile to 449.39 grams of CO2 per mile.

95% CI [410.61, 449.39]

The confidence interval gives the engineering team important information. The upper limit of your interval is below the target of 460 grams of CO2 per mile. This result provides solid statistical evidence that the emissions rate for the new engine will meet emissions standards.

Note: Confidence intervals for small sample sizes only deal with population means, and not population proportions. The statistical reason for this distinction is rather technical, so you don't need to worry about it for now.

Confidence Intervals in Python

Packages

```
import numpy as np
import pandas as pd
from scipy import stats
```

Making the Sampled Data

```
data = pd.read_csv('data.csv')
sampled_data = data.sample(n=50, replace=True, random_state=31208)
sampled_data
```

stats.norm.interval()

Syntax * `alpha` refers to the Confidence level * `loc` refers to the Sample Mean * `scale` refers to the Sample Standard Error

For `loc`

```
sample_mean = sampled_data['column'].mean()
sample_mean
```

Calculating the Standard Error

for `scale`

`.shape` function returns the number of rows and columns in a dataframe `.shape[0]` returns only the number of rows, which is the same number as your Sample Size.

```
estimated_standard_error = sampled_data['column'].std() / np.sqrt(sampled_data.shape[0])
```

Constructing the Interval

Code from Video

```
stats.norm.interval(alpha=0.95, loc=sample_mean, scale=estimated_standard_error)
```

Corrected Code

```
stats.norm.interval(0.95, loc=sample_mean, scale=estimated_standard_error)
```

e.g. Video Example

(np.float64(71.42241096968617), np.float64(77.02478903031381))

Confidence Interval: 95% CI [71.4, 77.0]

Activity

Choose your Sample Statistic

```
sample_mean = data['column'].mean()
sample mean
```

Choose your Confidence Level

```
confidence_level = 0.95
confidence_level
```

Calculate your Margin of Error

```
z_value = 1.96
standard_error = data['column'].std() / np.sqrt(data.shape[0])

margin_of_error = standard_error * z_value
```

Calculate your Interval

```
upper_ci_limit = sample_mean + margin_of_error
lower_ci_limit = sample_mean - margin_of_error
(lower_ci_limit, upper_ci_limit)
```

Hypothesis Tests

Hypothesis Testing: A Statistical Procedure that uses Sample data to evaluate an assumption about a population Parameter.

Data professionals conduct a Hypothesis Test to decide whether the evidence from their Sample data supports either the Null Hypothesis or the Alternative Hypothesis.

Statistical Significance: The claim that the results of a test or experiment are not explainable by chance alone.

Steps for Performing a Hypothesis Test

1. State the Null Hypothesis and the Alternative Hypothesis
2. Choose a Significance Level
3. Find the P-Value
4. Reject or Fail to Reject the Null Hypothesis

Null Hypothesis: A statement that is assumed to be true unless there is convincing evidence to the contrary.

The Null Hypothesis typically assumes that there is no effect in the population, and that your observed data occurs by chance.

Alternative Hypothesis: A statement that contradicts the Null Hypothesis and is accepted as true only if there is convincing evidence for it.

The Alternative Hypothesis typically assumes that there is an effect in the population, and that your observed data does not occur by chance

Significance Level: The Probability of Rejecting the Null Hypothesis when it is true.

P-Value: The Probability of observing results as or more extreme than the observed when the Null Hypothesis is true.

A lower P-Value means there is stronger evidence for the Alternative Hypothesis

Drawing a Conclusion

If P-Value $<$ (less than) Significance Level: Reject the Null Hypothesis

If P-Value $>$ (greater than) Significance Level: Fail to Reject the Null Hypothesis

	Null hypothesis (H_0)	Alternative hypothesis (H_a)
Claims	There is no effect in the population.	There is an effect in the population.
Language	• No effect • No difference • No relationship • No change	• An effect • A difference • A relationship • A change
Symbols	Equality (=, ≤, ≥)	Inequality (≠, <, >)

Figure 20: Null vs. Alternative Hypothesis

Types of Errors

Type I Error (False Positive): The Rejection of a Null Hypothesis that is actually true.

A Significance Level of 5% means that you're willing to accept a 5% chance that you are wrong when you Reject the Null Hypothesis.

To reduce your chance of making a Type I Error, choose a lower Significance Level

However choosing a lower Significance Level means your more likely to make a Type II Error or False Negative.

Type II Error (False Negative): The failure to Reject a Null Hypothesis which is actually false.

The Probability of making a Type I Error is called alpha (α). Your Significance Level or alpha (α) represents the probability of making a Type I Error.

The Probability of making a Type II Error is called beta (β), and beta is related to the power of a Hypothesis Test (power = 1 - β). Power refers to the likelihood that a test can correctly detect a real effect when there is one.

You can reduce your risk of making a Type II Error by ensuring your test has enough power. In data work, power is usually set at 0.80 or 80%. The higher the statistical power, the lower the Probability of making a Type II Error. To increase power, you can increase your Sample Size or your Significance Level

	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct Outcome! (True positive)
Fail to reject null hypothesis	Correct Outcome! (True negative)	Type II Error (False negative)

Figure 21: Type I and II Errors

One-Sample Tests

One-Sample Test: Determines whether or not a population Parameter like a Mean or Proportion is equal to a specific value.

One-Sample z-test

One-Sample z-test Assumptions

- **The data is a Random Sample of a Normally Distributed population**
- **The population Standard Deviation is known**

Test Statistic: A value that shows how closely your observed data matches the Distribution expected under the Null Hypothesis.

For a Z-Test the Test Statistic is a Z-Score

One-Sample z-test Formula for Means

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

$\bar{x}$ = Sample Mean

μ = Population Mean

σ = Population Standard Deviation

n = Sample Size

e.g.

$$Z = \frac{\bar{38} - 40}{\frac{5}{\sqrt{50}}} = -2.82$$

In this example the Z-Score is far to the left, almost 3 std dev below the mean. For a Normal Distribution, the Probability of getting a value less than your Z-Score (-2.82) is calculated by taking the area under the Curve to the left of the Z-Score. This is called a ***left-tailed test*** *because your P-Value is located on the left tail of the Distribution. The area under this part of the Curve is the same as your P-Value*

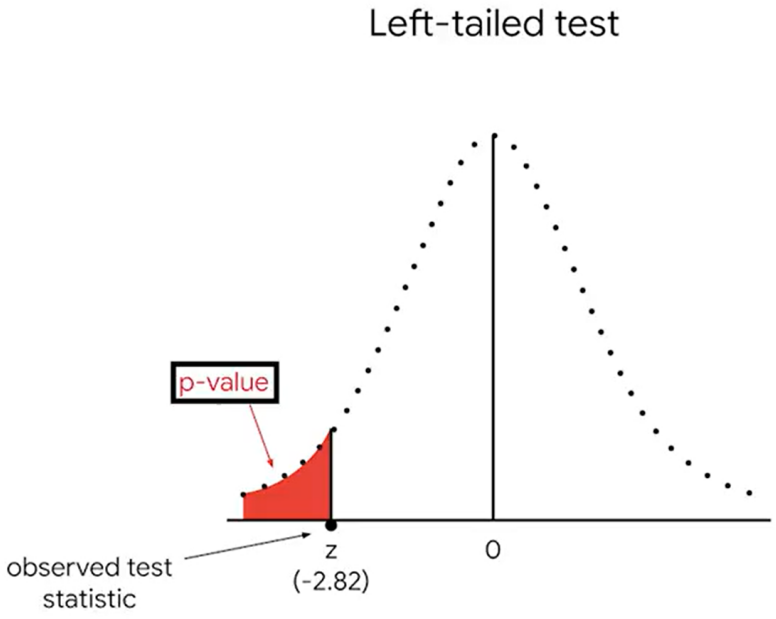


Figure 22: Left-tailed test

In a different testing scenario, your Test Statistic might be 2.45, and you might be interested in values as high or higher than the Z-Score 2.45. In that case your P-Value would be located on the right-tail of the Distribution and you would be conducting a ***right-tailed test***

One-Tailed and Two-Tailed Tests

- Left-Tailed Test: When the H_a states that the actual value of the Parameter is less than the value in the H_o.
- Right-Tailed Test: When the H_a states that the actual value of the Parameter is greater than the value in the H_o.
- Two-Tailed Test: When the H_a states that the actual value of the Parameter does not equal the value in the H_o.

Note: P-Value for a Two-Tailed Test is always two times the P-Value for a One-Tailed Test.

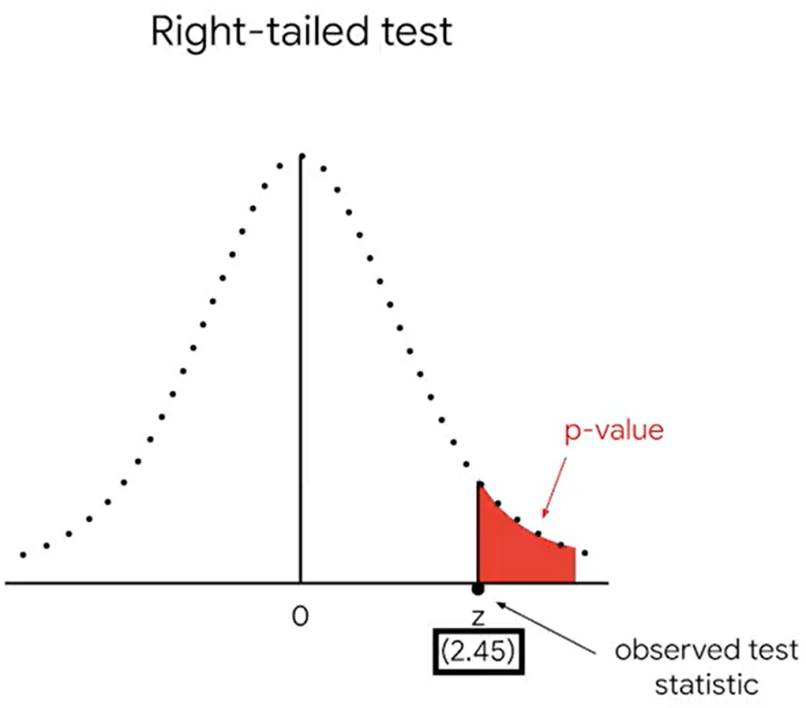


Figure 23: Right-tailed test

A two-tailed test results when the alternative hypothesis states that the actual value of the parameter does not equal the value in the null hypothesis.

Two-Sample Tests

Two-Sample Test: Determines whether or not two population Parameters such as two Means or two Proportions are equal to each other.

Two-Sample t-test

Two-Sample t-test for Means Assumptions

- **The two Samples are independent of each other**
- **For each Sample, the data is drawn Randomly from a Normally Distributed population**
- **The population Standard Deviation is unknown**

In practice the population Standard Deviation is usually unknown because its difficult to get complete data on large populations. So data professionals use a t-test for practical applications.

t-scores are based on the t-distribution

t-distribution has bigger tails than the standard Normal Distribution does.

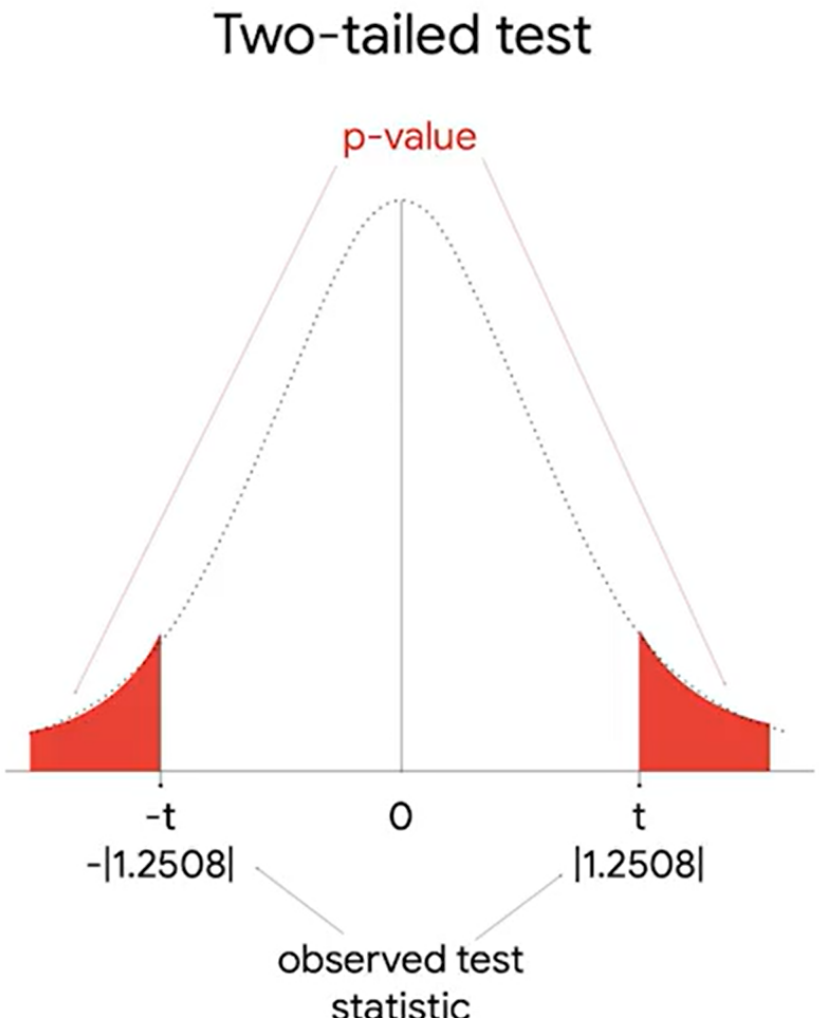


Figure 24: Two-tailed test

The bigger tails indicate the higher frequency of outliers that come with small datasets.

As the Sample Size increases, the t-distribution approaches the Normal Distribution

Two-Sample t-test Formula for Means

$$t = \frac{(\bar{x_1} - \bar{x_2})}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}}$$

$\bar{x_1}$, $\bar{x_2}$ = Sample Means

n_1, n_2 = Sample Sizes

s_1^2, s_2^2 = Sample Variances

Two-Sample z-test

For Technical reasons t-tests do not apply for Proportions

Two-Sample z-test Formula for Proportions

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p_0}(1 - \hat{p_0})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$\hat{p}_{1/2}$ = Sample Proportions for both groups

$n_{1/2}$ = Sample Sizes for both groups

$\hat{p_0}$ = Pooled Proportion

Pooled Proportion ($\hat{p_0}$): Weighted average of the Proportions from your 2 Samples.

Experimental Design: Refers to planning an experiment in order to collect data to answer your research question.

Pooled Proportions

Pooled: Combining multiple Sample estimates into one single, more stable estimate, assuming they came from the same Population (under H_0).

Used when H_0 assumes equality (e.g. $p_1 = p_2$)

Used when you want to leverage more data to improve the accuracy of a shared Parameter estimate

Shows up in: * Pooled Variance in t-tests (when assuming population Variances) * Pooled Proportions in z-tests (when assuming population Proportions)

For performing Two-Tailed z-tests for population Proportions

We use the Pooled Proportion when the H_0 assumes the two population Proportions are equal (e.g. $p_1 = p_2$).

Logic

We should combine (pool) the information from both Samples to get the best estimate of the common Proportion.

Formula

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$x_{1/2}$ = Number of successes in Samples 1 and 2

$n_{1/2}$ = Sample Sizes

$\hat{p}$ = Pooled Proportion (i.e. Combined Success Rate)

Hypothesis Testing in Python

Packages

```
import pandas as pd
from scipy import stats
```

Making the 2 Samples

```
data = pd.read_csv('data.csv')

state21 = data[data['statename'] == 'state21']
state28 = data[data['statename'] == 'state28']

sampled_state21 = state21.sample(n=20, random_state=13490, replace=True)
sampled_state28 = state28.sample(n=20, random_state=39103, replace=True)
```

Means of the 2 Sample

```
sampled_state21['column'].mean()
sampled_state28['column'].mean()
```

Observed Difference in Means

```
sampled_state21['column'].mean() - sampled_state28['column'].mean()
```

stats.ttest_ind()

Syntax

- `a` refers to observations from your first Sample
- `b` refers to observations from your second Sample
- `equal_var` is a Boolean or True/False Statement which indicates whether the population Variance of the two Samples is assumed to be equal.
- `alternative` specifies the direction of the Hypothesis Test, whether your performing a Two-Tailed or One-Tailed Test. Has 3 Arguments: `'two-sided'` = Two-Tailed, `'less'` = left, `'greater'` = right. If you omit the `alternative` argument it defaults to `'two-sided'`.

If you don't have access to the data for the entire population, you don't want to assume anything about the Variance

To avoid making a wrong assumption, set `equal_var` to `False`.

```
stats.ttest_ind(a=sampled_state21['column'], b=sampled_state28['column'], equal_var=False)
```

e.g. Output

TtestResult(statistic=np.float64(2.8980444277268735), pvalue=np.float64(0.006421719142765237), df=np.float64(35.20796133045557))

Activity

Two-Sample t-test with alternative = 'less'

```
stats.ttest_ind(a=data['column'], b=data['column'], equal_var=False, alternative='less')
```

alternative way of writing the code out

```
tstat, pvalue = stats.ttest_ind(a=data['column'], b=data['column'], equal_var=False,
                                alternative='less')

print(tstat)
print(pvalue)
```

stats.ttest_1samp()

One-Sample t-test with popmean and alternative = 'greater'

```
stats.ttest_1samp(a=data['column'], popmean=10, alternative='greater')
```

- `popmean` is the population Mean you are comparing your Sample against.

alternative way of writing the code out

```
tstat, pvalue = stats.ttest_1samp(a=data['column'], popmean=10, alternative='greater')

print(tstat)
print(pvalue)
```

Certificate Readings

- Measures of Central Tendency: The Mean, The Median, and The Mode
- Measures of Dispersion: Range, Variance, and Standard Deviation
- Measures of Position: Percentiles and Quartiles
- Fundamental Concepts of Probability
- The Probability of Multiple Events
- Calculate Conditional Probability for Dependent Events
- Calculate Conditional Probability with Bayes's Theorem
- Discrete Probability Distributions
- Model Data with the Normal Distribution
- The Stages of the Sampling Process
- The Relationship between Sample and Population
- Probability Sampling Methods
- Non-Probability Sampling Methods
- The Sampling Distribution of the Mean
- Infer Population Parameters with the Central Limit Theorem
- Confidence Interval: Correct and Incorrect Interpretations
- Construct a Confidence Interval for a Small Sample Size
- Differences between the Null and Alternative Hypotheses
- Type I and Type II Errors
- Determine if Data has Statistical Significance
- One-Tailed and Two-Tailed Tests

Python Notebooks

- Compute Descriptive Statistics with Python
- Work with Probability Distributions in Python
- Sampling Distributions with Python
- Confidence Intervals with Python
- Exemplar_Explore Confidence Intervals
- Activity_Explore Confidence Intervals
- Use Python to Conduct a Hypothesis Test
- Exemplar_Explore Hypothesis Testing
- Activity_Explore Hypothesis Testing
- Course 4 End of Course Portfolio Project

Applied Statistics for Data Science with Python is a practical guide to the statistical foundations that drive modern data analysis, machine learning and data-driven decision making. Designed for students, analysts and professionals, this book bridges classical statistical theory with hands-on implementation in Python.

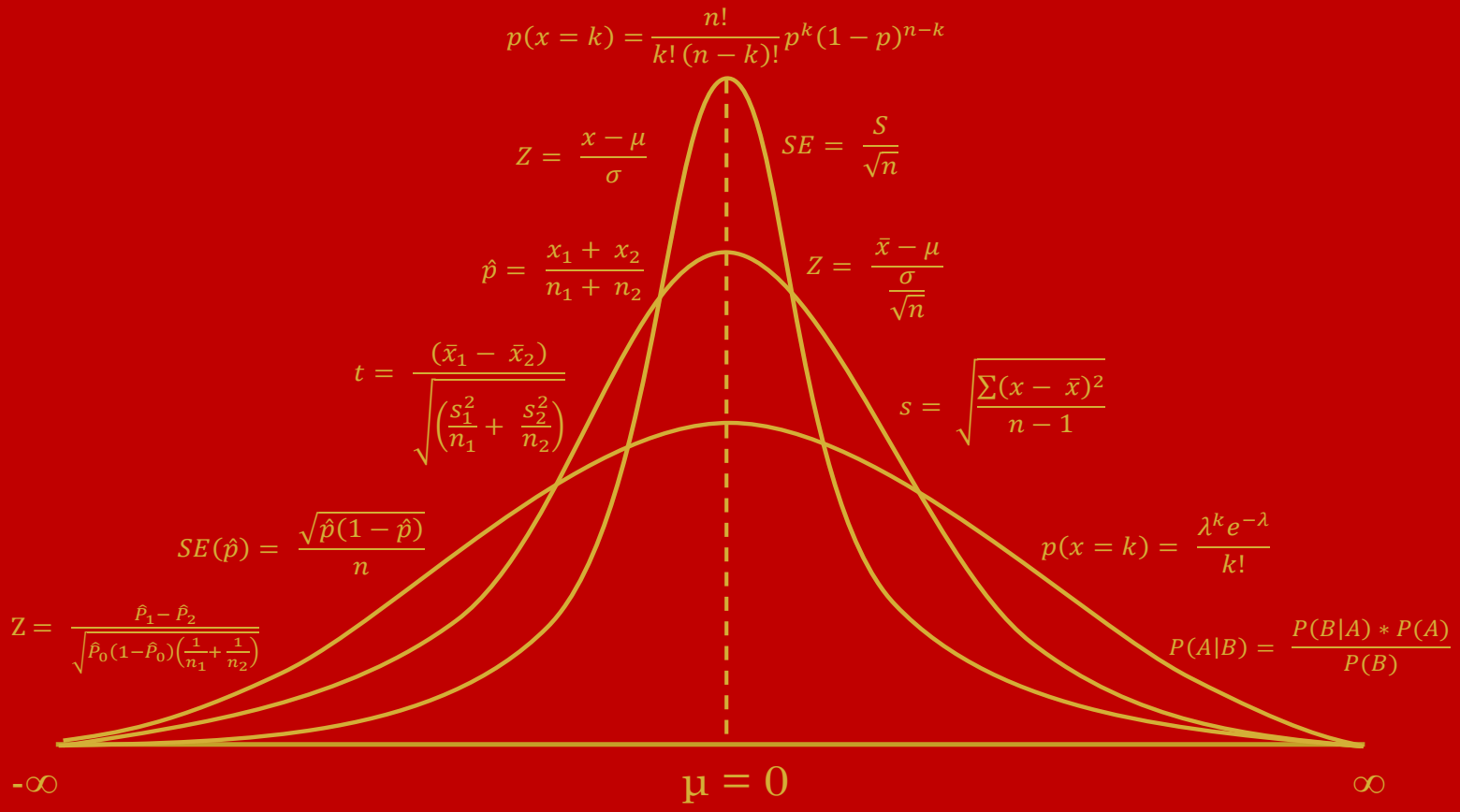
This book covers everything from descriptive statistics to probability and sampling methodology with their distributions to interpreting and constructing confidence intervals and hypothesis testing. With each of the techniques being built up from its first principles and then connected to real-world data analysis. Rather than treating statistics as abstract mathematics, the focus is on interpretation, assumptions and how results should be used to make informed conclusions from data.

Python is used throughout the entirety of this book to demonstrate how statistical concepts translate into working code. These examples illustrate data preparation, statistical modelling and the interpretation of analytical outputs using industry-standard tools. The objective is to move beyond numerical outputs and move towards meaningful interpretation and evidence-based decision-making.

Whether used as a learning resource or a reference guide, this book offers a comprehensive foundation for statistical analysis in contemporary data science.

Scan for my portfolio

Applied Statistics for Data Science with Python



Alexander Thompson BSc (Hons)

Applied Statistics for Data Science with Python

Alexander Thompson BSc (Hons)

28th January 2026

Contents

Preface

This is the Statistics checklist for computing statistics in Python, and is sourced from the statistical information from the Google Advanced Data Analytics Professional Certificate. This book covers in depth statistical methodology and how to apply them to the data science field, covering the topics of probability and its distributions to sampling and the central limit theorem to confidence intervals and hypothesis tests. These concepts enable oneself to be able to extend their analytical outreach with the end result being an improved analysis of the data.

In regards to the data science field, these concepts play the role of a bridge, connecting exploratory data analysis to regression and machine learning modelling. This is due to the statistical concepts explaining everything from the basics, such as descriptive statistics and probabilities to the more advanced concepts such as hypothesis testing, with regression modelling building directly on top of this. With respect to machine learning, these statistical concepts are heavily applied, such as with the Naive Bayes machine learning technique using posterior probability and the Random Forest Technique using sampling with replacement. Therefore, these concepts bridge the gap between basic data analytics and high-level data science and is why this book is essential for covering these concepts.

These concepts are carried out using the programming language Python using the appropriate functions and mainly using the statsmodels library. Due to the updates made over time to the Python language the code in this book will have changed over time as the different versions of Python get released. Python is a great tool for statistical analysis and works excellently within the data analytics framework. Using Python allows the analyst to go straight from the data preparation to machine learning model development all in the same script or notebook.

This book is the full and complete guide to descriptive and inferential statistics and their applications via Python, and is intended for use as a handbook for revision and applied work. Whether it is revising statistical concepts for a project or teaching another analyst about statistics and how to apply them, this book is the definitive guide and will help any data scientist reach their goals. A majority of the concepts covered here in this book I am very familiar with due to my Bachelors of Science degree in Economics, where hypothesis testing and descriptive statistics were prevalent. Therefore, my added input into these concepts will aid in the understanding and application of these concepts and how they are applied into the real analytical world.

"Fill your mind, not to impress others but to endure yourself"

Introduction

Statistics underpins nearly every stage of modern data science, from exploratory data analysis to predictive modelling and evidence-based decision making under uncertainty. In applied contexts such as economics, finance and machine learning, statistical reasoning is essential not only for producing analytical outputs but also for interpreting the results correctly, assessing their ability and communicating insights with clarity and precision.

Within economics, statistical methods have long been used to analyse relationships between variables, test theoretical models and draw inferences about populations, markets and behaviours from observed data. These same principles now form the backbone of contemporary data science, where analysts work with increasingly large complex datasets to model outcomes, forecast trends and inform strategic or policy decisions. In both domains, the ability to reason statistically provides the crucial link between raw data and meaningful interpretation. At its core, statistics offers a structured framework for understanding variation. Real-world data are inherently imperfect and are shaped by randomness, measurement error, sampling processes and unobserved influences. Statistical methods enable analysts to summarise data, quantify uncertainty and draw informed conclusions about populations based on finite samples. These foundations are critical not only for descriptive analysis, but also for the responsible application of regression models, classification algorithms and machine learning techniques, where the assumptions and uncertainty must be carefully considered.

The structure of this book reflects the progression from foundational concepts to applied analytical practice. It begins with descriptive statistics, focusing on the summarisation and visualisation of data through measures of central tendency, dispersion and distributional form. These tools represent the first step in any data science or economic analysis, providing essential insight into the structure, quality and limitations of a dataset prior to formal modelling. The text then introduces probability theory and sampling, thereby establishing the mathematical language required to reason rigorously about uncertainty and stochastic processes. Building on this foundation, the book advances to statistical inference, including estimation, confidence intervals, hypothesis testing and model-based reasoning. Throughout the book, Python is used as the primary computational environment to demonstrate how statistical concepts are applied in practice. Code examples are designed to mirror real analytical workflows commonly encountered in data science and economics, with an emphasis placed on interpretation, assumptions and limitations rather than just computation alone. This applied focus reflects the reality that effective data analysis requires both technical proficiency and sound statistical judgement.

This text is intended to serve as both a structured learning resource and a practical reference guide. Readers may follow the material sequentially to develop a coherent understanding of statistical foundations or consult individual sections to support revision, applied projects and professional analysis. By integrating statistical theory, economic intuition and practical implementation, the book aims to support robust, interpretable and methodologically sound data-driven analysis.

Descriptive Statistics

Formulas

Mean Formula:

$$Mean = \text{Sum of all Values/Total Number of Values}$$

Standard Deviation (Sample)

There are different formulas to calculate the standard deviation for a population and a sample. As a reminder, data professionals typically work with sample data, and they make inferences about populations based on the sample. So, let's review the formula for sample standard deviation:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

n = Total Number of Values in your Sample

x = Each individual data Value

$\bar{x}$ = The Mean of your data Values

$\sum$ = Sum (Greek Letter Sigma)

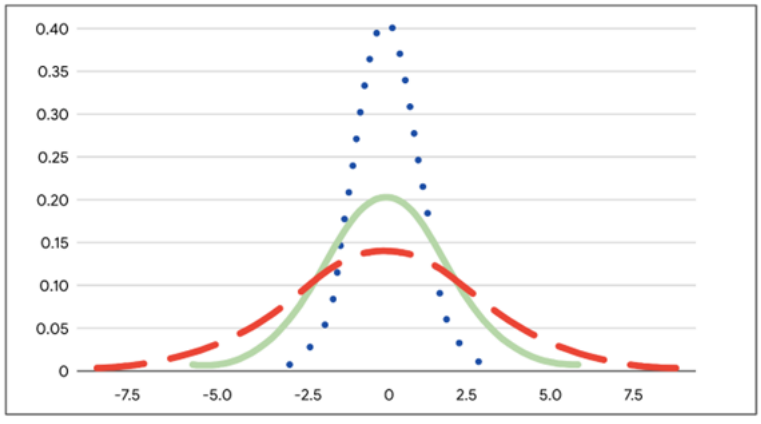


Figure 1: Measure of Dispersion: Std Deviation

Relationship between Quartiles and Percentiles

- Q1 = 25th Percentile
- Q2 = 50th Percentile
- Q3 = 75th Percentile

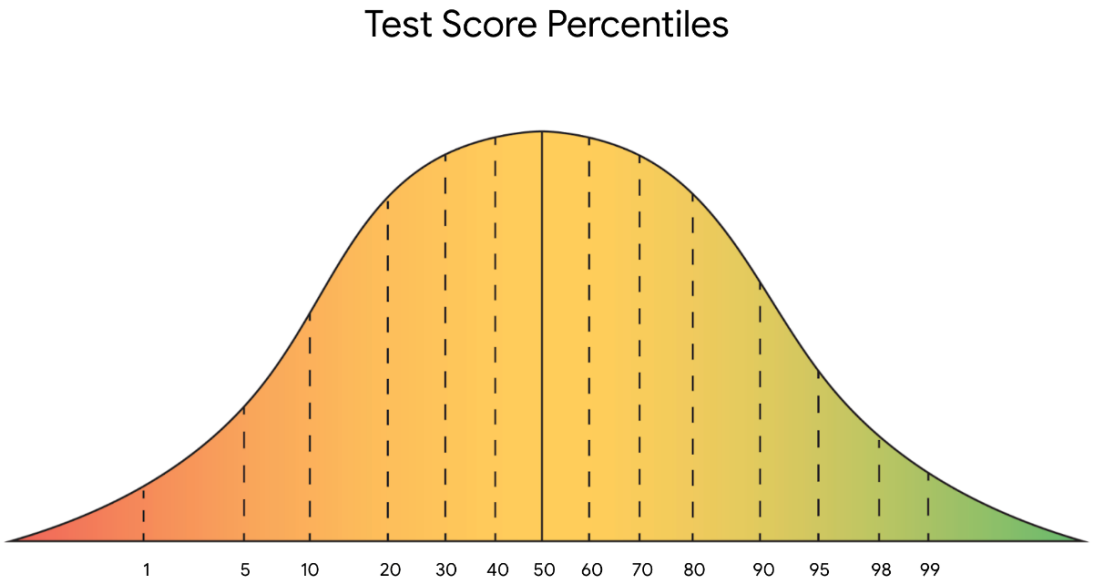


Figure 2: Measures of Position: Percentiles

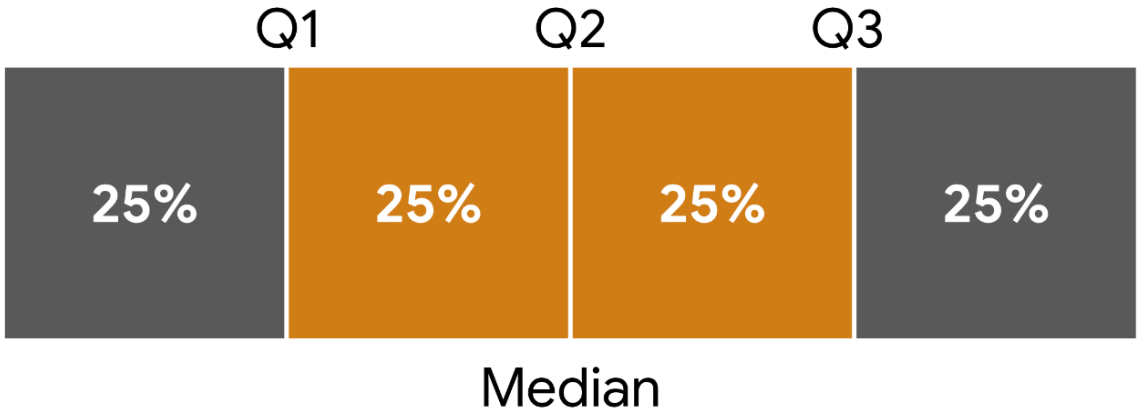


Figure 3: Measures of Position: Quartiles

IQR

$$IQR = Q3 - Q1$$

Five Number Summary

- The Minimum
- The First Quartile (Q1)
- The Median, or Second Quartile (Q2)
- The Third Quartile (Q3)
- The Maximum

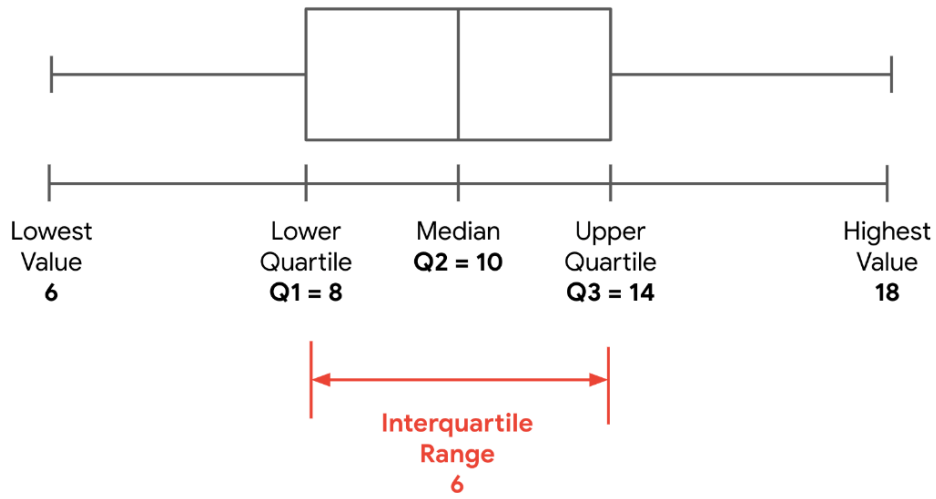


Figure 4: Measures of Position: The Interquartile Range

Descriptive Statistics in Python

Packages

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Functions

```
data = pd.read_csv('data.csv')

data.head()

data['column'].describe()
```

For the `.describe()` function above:

- The `mean` helps to clarify the centre of your dataset.
- The Categories `25%`, `50%` and `75%` refer to Q1, Q2 and Q3 respectively, Remember that Q2 is also the Median of your dataset.

```
data['categorical column'].describe()
```

For the `.describe()` function above:

- `count` This is the total number of non-null entries in the column.
- The `unique` Category, shows how many unique categories (distinct values) are present in that column.
- The `top` Category, is the most frequently occurring value (the mode) in the column.
- The `freq` Category, is the frequency (number of occurrences) of the most common value (shown in `top`)

```
range = data['column'].max() - data['column'].min()
range

#From Activity

np.mean()
np.median()
np.min()
np.max()
np.std()
```

Probability

Formulas

Classical Probability

$$\text{Classical Probability} = \frac{\text{Number of Desired Outcomes}}{\text{Total Number of Possible Outcomes}}$$

Empirical Probability

$$\text{Empirical Probability} = \frac{\text{Number of Times a Specific Event Occurs}}{\text{Total Number of Events}}$$

Three Concepts at the Foundation of Probability Theory:

- Random Experiment
- Outcome
- Event

Random Experiment: A process where outcomes cannot be predicted with certainty

All random experiments have three things in common:

- The Experiment can have more than one possible outcome
- You can represent each possible outcome in advance
- The outcome of the experiment depends on chance

The Probability of an Event

The probability that an event will occur is expressed as a number between O and 1. Probability can also be expressed as a percent.

- If the Probability of an event equals 0, there is a 0% chance that the event will occur.
- If the Probability of an event equals 1, there is a 100% chance that the event will occur.
- If the Probability of an event equals 0.5, there is a 50% chance that the event will occur - or not occur.
- If the Probability of an event is close to 0, there is a small chance that the event will occur.
- If the Probability of an event is close to 1, there is a strong chance that the event will occur.

Calculate the Probability of an Event

To calculate the probability of an event in which all possible outcomes are equally likely, you divide the number of desired outcomes by the total number of possible outcomes. You may recall that this is also the formula for classical probability:

$$\text{Probability of an Event} = \frac{\text{Number of desired outcomes}}{\text{Total Number of Possible Outcomes}}$$

Probability Notation

- The probability of event A is written as P(A).
- The probability of event B is written as P(B).
- For any event A, $0 \leq$ P(A) ≤ 1. In other words, the probability of any event A is always between 0 and 1.
- If P(A) $>$ P(B), then event A has a higher chance of occurring than event B.
- If P(A) = P(B), then event A and event B are equally likely to occur.
- P(A'), Probability of not event A.

Two events are **mutually exclusive** if they cannot occur at the same time.

Two events are **independent** if the occurrence of one event does not change the probability of the other event. This means that one event does not affect the outcome of the other event.

Three Basic Rules of Probability

- Complement Rule
- Addition Rule
- Multiplication Rule

Complement Rule *For Mutually Exclusive Events*

$$P(A') = 1 - P(A)$$

Addition Rule *For Mutually Exclusive Events*

$$P(AorB) = P(A) + P(B)$$

Multiplication Rule *For Independent Events*

$$P(AandB) = P(A) * P(B)$$

Conditional Probability: The Probability of an event occurring given that another event has already occurred.

Dependent Events: Two Events are dependent if the occurrence of one event changes the Probability of the other Event.

Conditional Probability Formula

$$P(AandB) = P(A) * P(B|A)$$

- $P(AandB)$: Probability of Event A and Event B
- $P(A)$: Probability of Event A
- $P(B|A)$: Probability of Event B given Event A

The vertical Bar (|) represents that Event B depends on Event A happening, we say this as the Probability of a B given A.

The formula can also be expressed as:

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

Bayes's Theorem

Used for determining Conditional Probability

Bayes Theorem provides a way to update the Probability of an event based on new information about the Event.

Prior Probability: The Probability of an Event before new data is collected.

Posterior Probability: The updated Probability of an Event based on new data.

Posterior means occurring after

Posterior Probability is calculated by updating the Prior Probability using Bayes Theorem

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

Bayes's theorem states that for any two events A and B, the probability of A given B equals the probability of A multiplied by the probability of B given A divided by the probability of B.

In the theorem, prior probability is the probability of event A. Posterior probability, or what you're trying to calculate, is the probability of event A given event B.

Sometimes, statisticians and data professionals use the term "likelihood" to refer to the probability of event B given event A, and the term "evidence" to refer to the probability of event B.

- $P(A)$: Prior Probability
- $P(A|B)$: Posterior Probability
- $P(B|A)$: Likelihood
- $P(B)$: Evidence

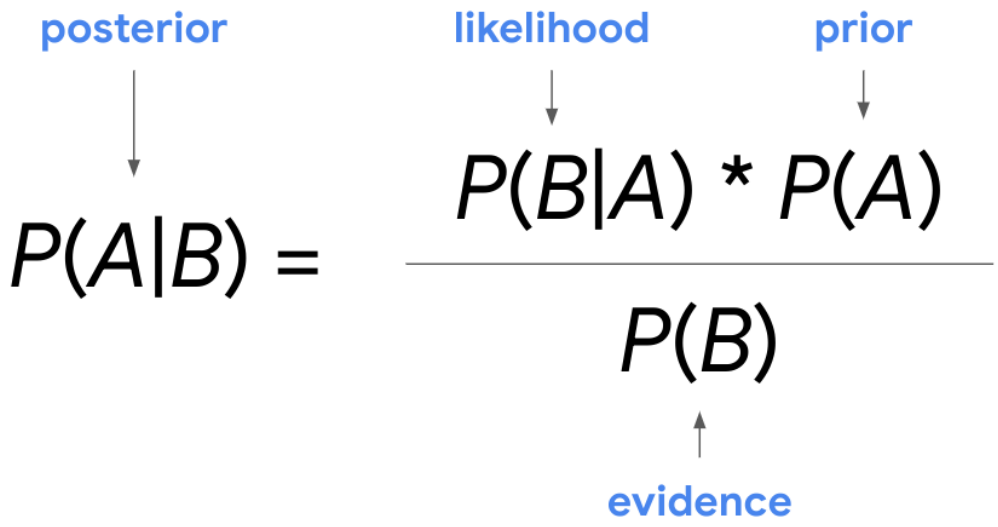


Figure 5: Bayes Theorem

Bayes's Theorem (Expanded Version)

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + P(B|notA) * P(notA)}$$

You can use the 2 versions of Bayes Theorem to deal with different types of problems.

For instance, if you don't know the Probability of Event B, in this case you can use the expanded version of Bayes Theorem, because you don't need to know the Probability of Event B.

Probability Distributions

Probability Distribution: Describes the likelihood of the possible outcomes of a Random Event.

Random Variable: Represents the values for the possible outcomes of a Random Event.

Random Variables

- Discrete
- Continuous

Discrete Random Variable: Has a countable number of possible values.

Often are whole numbers that can be counted

Continuous Random Variable: Takes all the possible values in some range of numbers

When your dealing with Continuous Variables you're dealing with decimal values rather than whole numbers.

Typically these are decimal values that can be measured such as height, weight, time or temperature.

Discrete or Continuous Variables

- Count the number of outcomes = Discrete
- Measure the outcome = Continuous

Discrete Distributions represent Discrete Random Variables

Continuous Distributions represent Continuous Random Variables

- Discrete Distributions can be represented as Tables or Histograms. With the Variable your counting on the X-Axis and the Probability on the Y-Axis.
- Continuous Distributions are represented in intervals and are represented as a Curve (or Bell Curve) with different intervals. X-Axis has the variable your measuring in Intervals (e.g. 10-17), and the Y-Axis has the Probability Density, which is not the same as Probability and is a Statistical Function.

Binomial Distribution

Binomial Distribution: A Discrete Distribution that Models the Probability of Events with only two possible outcomes, success or failure. These outcomes are Mutually Exclusive and cannot occur at the same time.

A Binomial Experiment has the following attributes: * The Experiment consists of a number of repeated trials. * Each trial has only only two possible outcomes. * The Probability of success is the same for each trial. * Each trial is independent.

Binomial Distribution Formula

$$p(x = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

k = Number of Successes

n = Number of Trials

p = The Probability of Success on a given Trial

$(n - k)$ = "n-choose-k" number of ways to obtain k successes in n trials

Can use a Histogram to visualise the Binomial Distribution

- X-Axis = Random Variable
- Y-Axis = Probability

Poisson Distribution

Poisson Distribution: Models the Probability that a certain number of Events will occur during a specific time period.

Can also be used to represent the number of events that occur in a specific space, such as a distance, area or volume.

The Poisson Distribution represents a type of Random Experiment called a Poisson Experiment. A Poisson Experiment has the following attributes:

- The number of Events in the Experiment can be counted.
- The mean number of Events that occur during a specific time period is known.
- Each Event is independent.

Poisson Distribution Formula

$$p(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

λ = The Mean number of Events that occur during a specific time period.

k = Number of Events

e = Constant equal to approximately 2.71828

! = Stands for Factorial. A function that multiplies a number by every whole number below it down to 1. e.g. 2 factorial is 2×1.

Can use a Histogram to visualise the Poisson Distribution

- X-Axis = Number of Events
- Y-Axis = Probability

Uniform Distribution

Uniform Distribution: Describes Events whose outcomes are all equally likely, or have equal Probability.

Can Visualise the Uniform Distribution with a Histogram

- X-Axis = Random Variable
- Y-Axis = Probability

Bernoulli Distribution

Bernoulli Distribution: Models Events that only have 2 possible outcomes (success or failure). Refers to only a single trial of an experiment.

Can be visualised with a Histogram

- X-Axis = Random Variable
- Y-Axis = Probability

Probability Density and Probability

A probability function is a mathematical function that provides probabilities for the possible outcomes of a random variable.

There are two types of probability functions:

- Probability Mass Functions (PMFs) represent Discrete Random Variables
- Probability Density Functions (PDFs) represent Continuous Random Variables

A probability function can be represented as an equation or a graph. The math involved in probability functions is beyond the scope of this course. For now, it's important to know that the graph of a PDF appears as a curve. You've learned about the bell curve, which refers to the graph for a normal distribution.

The Normal Distribution

Normal Distribution: A Continuous Probability Distribution that is symmetrical on both sides of the Mean and bell-shaped.

It is often called the Bell Curve

It is also known as the Gaussian Distribution

The most common Distribution in Statistics.

Normal Distribution have the following features:

- The shape is a Bell Curve
- The Mean is located at the centre of the Curve
- The Curve is symmetrical on both sides of the centre
- The total area under the Curve equals to 1

Under a Normal Distribution, the distance of a data point from the Mean is often measured in Standard Deviations.

The Values along a Curve are distributed in a regular pattern based on their distance from the Mean, this is known as the Empirical Rule

The Empirical Rule

- 68% of Values fall within 1 Standard Deviation of the Mean.
- 95% of Values fall within 2 Standard Deviations of the Mean.
- 99.7% of Values fall within 3 Standard Deviations of the Mean.

The Empirical Rule is useful for estimating data, especially for large datasets like height and weight data for an entire population.

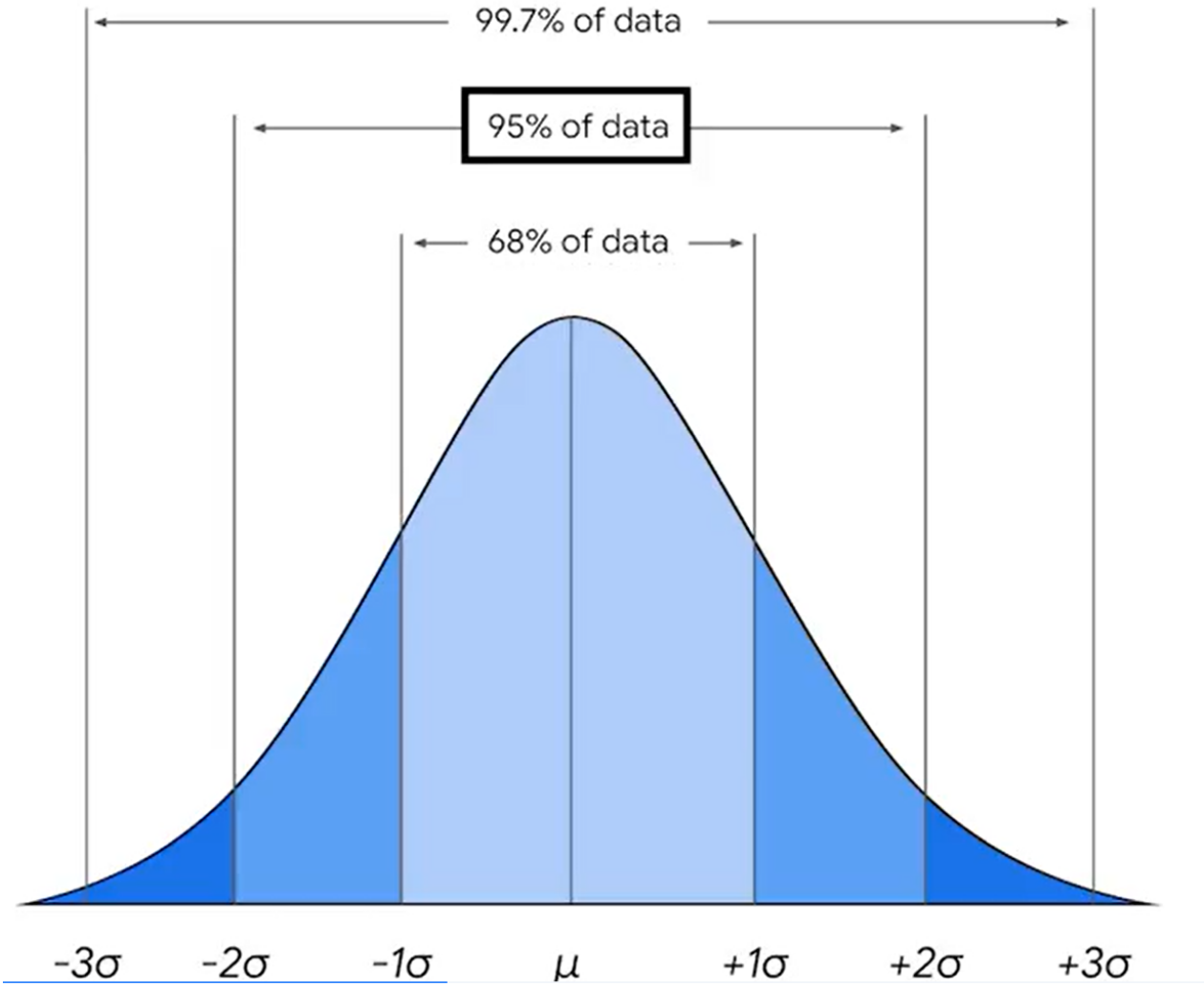


Figure 6: The Empirical Rule

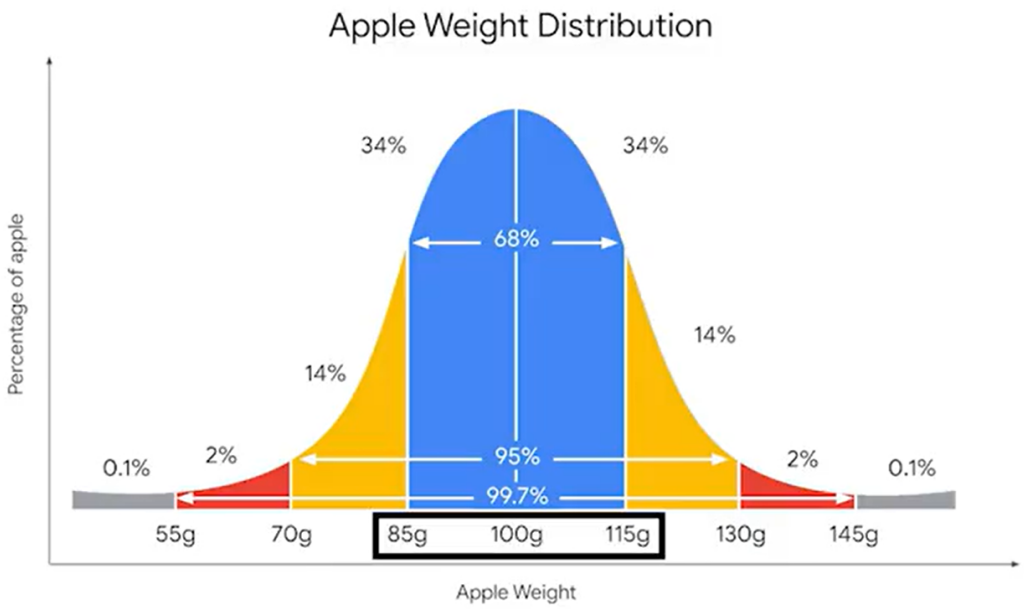


Figure 7: E.g. of the Empirical Rule

You can use the Empirical Rule to get an estimate of the Distributions of values in your dataset, such as what percentage of values will fall within 1, 2 or 3 Standard Deviations of the Mean. This saves time and helps you better understand your data.

Knowing the location of your values on a Normal Distribution is useful for detecting outliers

Typically, data professionals consider values that lie more than 3 Standard Deviations below or above the Mean to be Outliers.

Z-Scores

Z-Score: A Measure of how many Standard Deviations below or above the population Mean a data point is.

Gives you an idea of how far from the mean a data point is

- Z-Score is 0, value is equal to the Mean
- Z-Score is Positive, value is greater than the Mean
- Z-Score is Negative, value is less than the Mean

Z-Scores are also called Standard Scores because they're based on what's called the Standard Normal Distribution

A Standard Normal Distribution is just a Normal Distribution with a Mean of 0 and a Standard Deviation of 1

Z-Scores typically range from -3 to 3

Standardisation: The process of putting different variables on the same scale.

Standardisation is useful because it let's you compare scores from different datasets that may have different units, Mean values and Standard Deviations

Data professionals use Z-Scores to better understand the relationship between data values within a single dataset and between different datasets.

Z-Score Formula

$$Z = \frac{x - \mu}{\sigma}$$

x = Single data Value or Raw Score

μ = Population Mean

σ = Population Standard Deviation

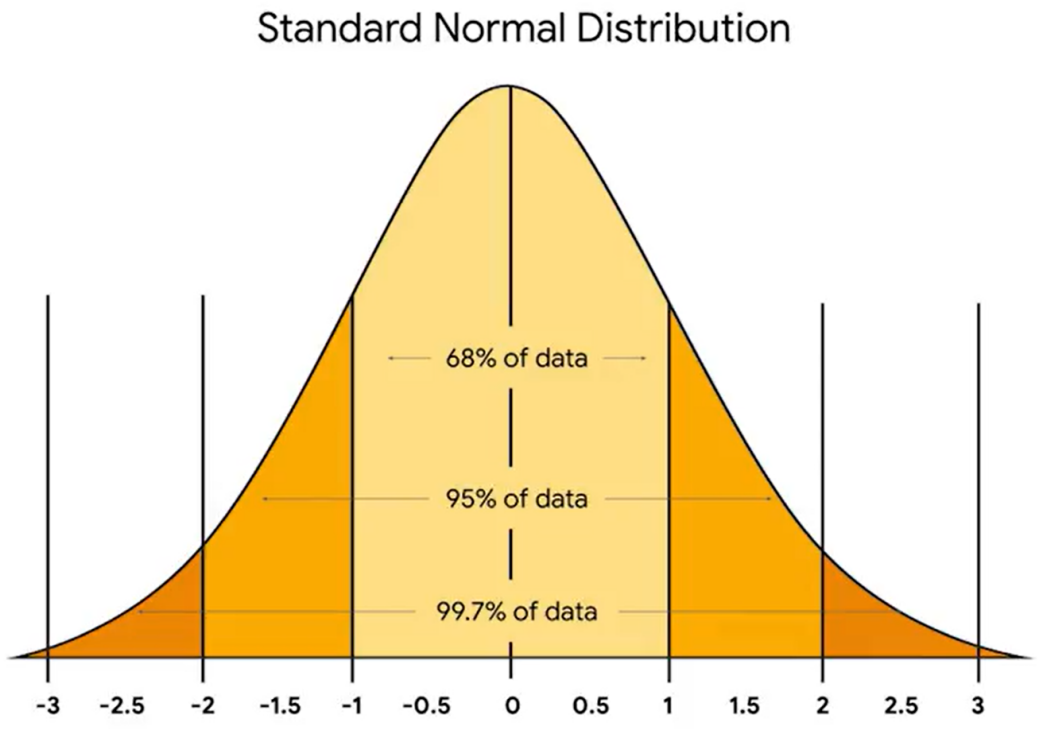


Figure 8: Standard Normal Distribution

Probability Distributions in Python

Packages

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
import statsmodels.api as sm
```

The first step is to find out what type of Probability Distribution your data is. To do this we will plot a histogram

```
data = pd.read_csv('data.csv')

data['column'].hist()
```

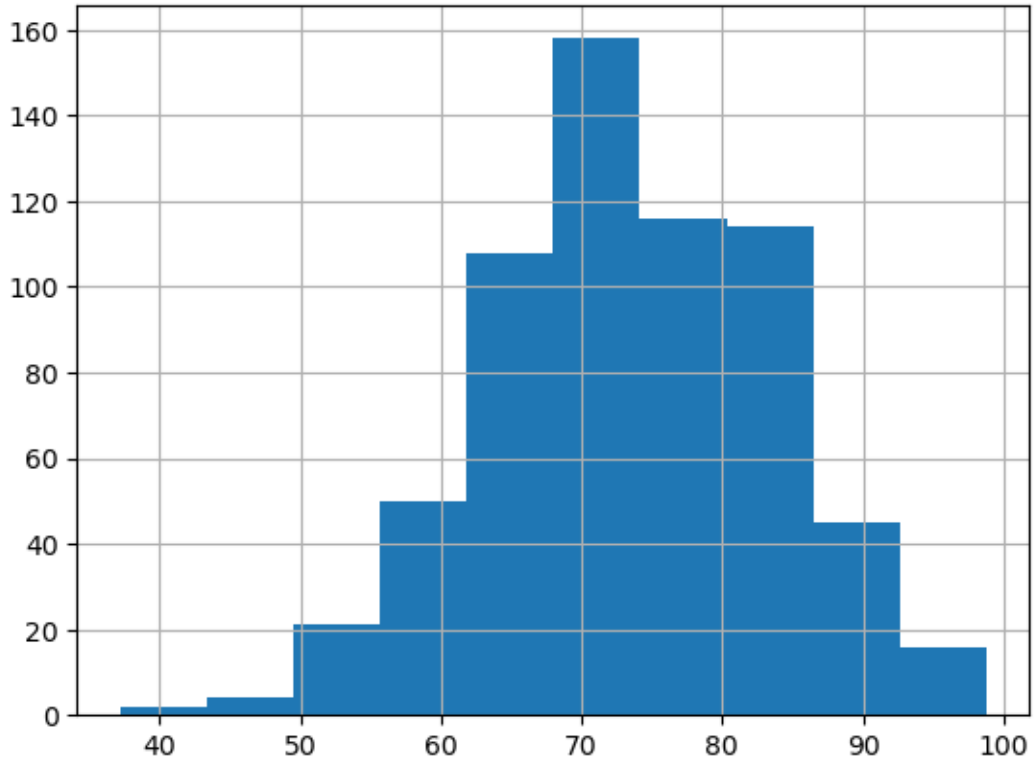


Figure 9: Python Distribution

Looks Normally Distributed

In order to verify if the data is Normally Distributed we have to see if it follows the Empirical Rule.

1 std dev of the Mean

Mean

```
mean_data = data['column'].mean()
```

Standard Deviation

```
std_data = data['column'].std()
```

Computing % of data that fall within 1 std dev of the Mean

```
upper_limit = mean_data + 1 * std_data
lower_limit = mean_data - 1 * std_data

((data['column'] >= lower_limit) & (data['column'] <= upper_limit)).mean()
```

- `lower_limit` will be 1 Standard Deviation below the Mean
- `upper_limit` will be 1 Standard Deviation above the Mean
- The Boolean Mask, tells the computer to decide if each value in the `column` is between the `lower_limit` and `upper_limit`.
- In other words to decide if each value is greater than or equal to 1 Standard Deviation below the mean and less than or equal to standard deviation above the mean.
- Use the `.mean()` function to divide the number of values that are within 1 standard deviation of the mean by the total number of values.

2 std dev of the Mean

```
upper_limit = mean_data + 2 * std_data
lower_limit = mean_data - 2 * std_data

((data['column'] >= lower_limit) & (data['column'] <= upper_limit)).mean()
```

3 std dev of the Mean

```
upper_limit = mean_data + 3 * std_data
lower_limit = mean_data - 3 * std_data

((data['column'] >= lower_limit) & (data['column'] <= upper_limit)).mean()
```

Then check if these values are close or equal to the empirical rule.

Outlier Detection with Z-Scores

First create a new column in your data called `Z-SCORE` that includes the Z-Score for each value in your original column.

Then use the `stats.zscore()` function to compute the Z-Score.

```
data['Z-SCORE'] = stats.zscore(data['column'])
data
```

Now write some code to identify the outliers with Z-Score that are greater or less than 3 Standard Deviations from the Mean.

Use the Relational Operators `>` greater than, `<` less than and the Bitwise Operator `|`.

```
data[(data['Z-SCORE'] > 3) | (data['Z-SCORE'] < -3)]
```

You'll then be able to see the outliers.

Sampling

Sampling: The process of selecting a subset of data from a population.

Representative Sample: Accurately reflects the characteristics of a population

The Sampling Process

Step 1: Identify the target Population

Step 2: Select the Sampling Frame

Step 3: Choose the Sampling Method

Step 4: Determine the Sample Size

Step 5: Collect the Sample Data

Target Population: The complete set of elements that you're interested in knowing more about.

Sampling Frame: A list of all the items in your Target population.

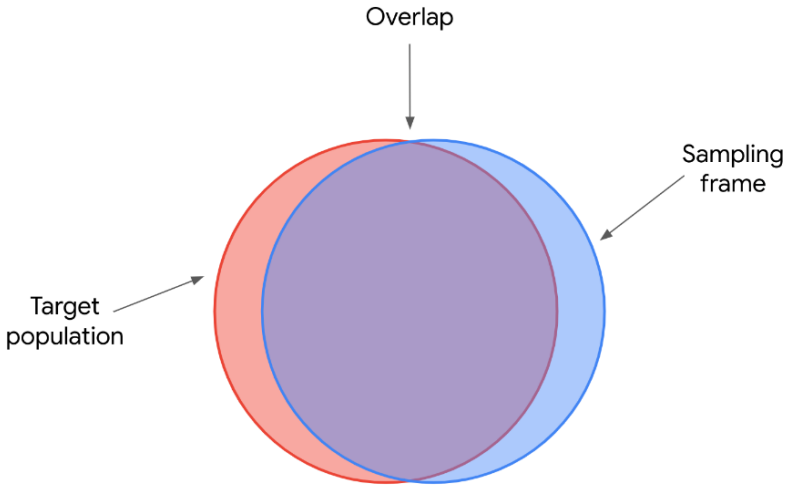


Figure 10: Sampling Frame

2 Main Sampling Methods:

- Probability Sampling
- Non-Probability Sampling

Probability Sampling: Uses Random selection to generate a Sample

Non-Probability Sampling: Based on convenience or personal preference.

Sample Size: The number of individuals or items chosen for a study or experiment

Sample Size helps determine the accuracy of the predictions you make about the population

The larger the Sample Size the more accurate your predictions

Effective Sampling ensures that your Sample data is representative of your Target population. Then when you use Sample data to make inferences about the population, you can be reasonably confident that your inferences are reliable.

Decisions you make in each step of the process can affect the quality of your sample data

Probability Sampling Methods

Simple Random Sampling: Every member of a population is selected randomly and has an equal chance of being chosen

You can Randomly select members using a Random Number Generator or by another method of Random Selection.

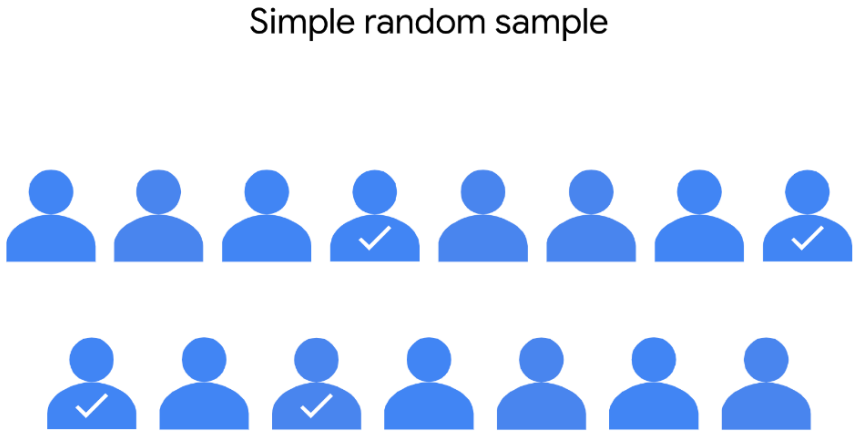


Figure 11: Simple Random Sample

Stratified Random Sampling: Divide a population into groups and randomly select some members from each group to be in the Sample

These groups are called **Strata**

Strata can be organised by age, gender, income or whatever Category your interested in studying.

Helps ensure that members from each group are included

Disadvantage: It can be difficult to identify appropriate strata for a study if you lack knowledge of a population

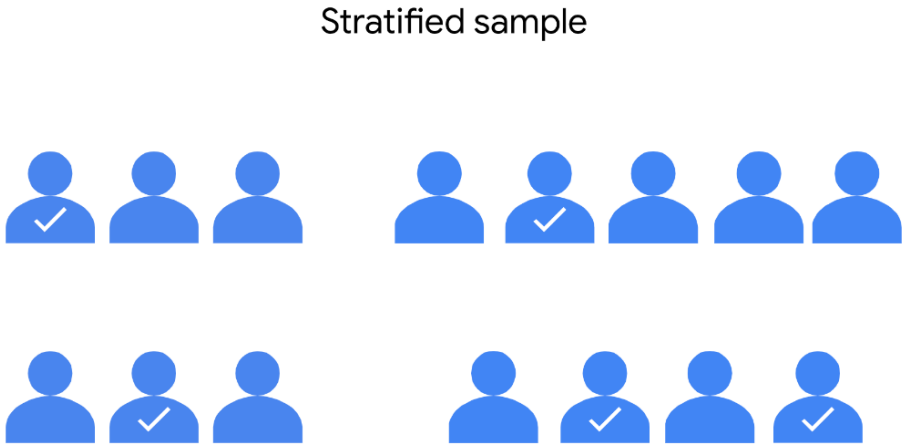


Figure 12: Stratified Random Sample

Cluster Random Sampling: Divide a population into Clusters, randomly select certain Clusters, and include all members from the chosen Clusters in the Sample.

Clusters are divided using identifying details such as age, gender, location, or whatever you want to study

Helpful when dealing with large and diverse populations that have clearly defined subgroups.

Disadvantage: May be difficult to create Clusters that accurately reflect the overall population

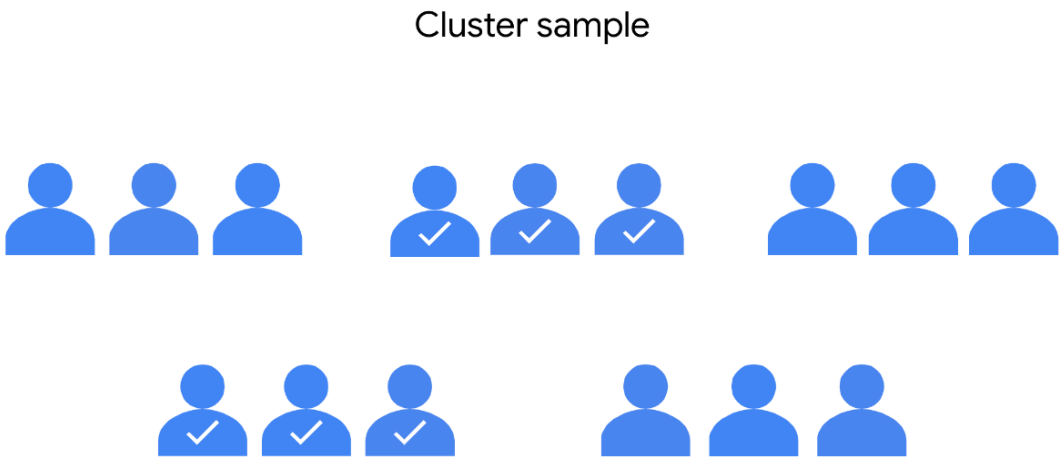


Figure 13: Cluster Random Sample

Systematic Random Sampling: Put every member of a population into an ordered sequence. Then, you choose a Random starting point in the sequence and select members for your Sample at regular intervals.

Disadvantage: You need to know the size of the population that you want to study before you begin. If you don't have this info its difficult to choose consistent intervals.

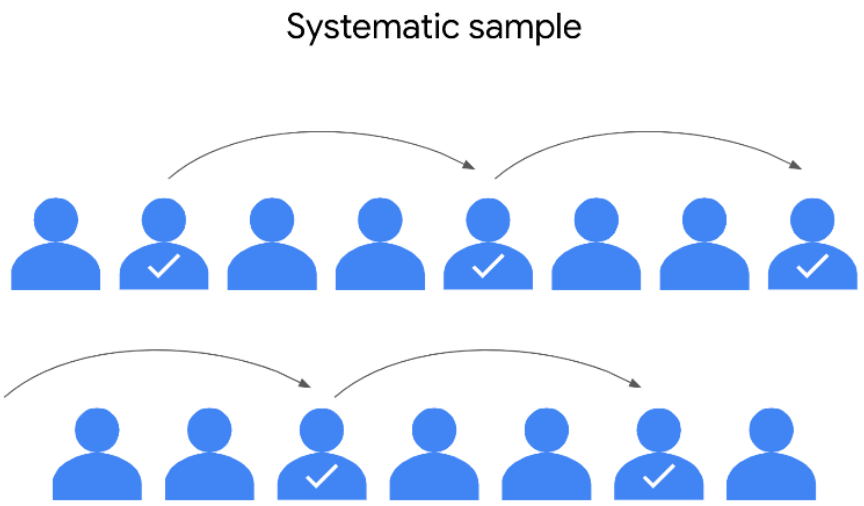


Figure 14: Systemic Random Sample

Non-Probability Sampling Methods

Sampling Bias: When a Sample is not representative of the Population as a whole.

Non-Probability Sampling Methods result in Bias.

Often less expensive and more convenient for researchers to conduct.

Sometimes due to budget, time or other reasons, its just not possible to use Probability Sampling.

Can be useful for exploratory statistics, which seek to develop initial understanding of a population, not draw conclusions or make predictions about the population as a whole.

Convenience Sampling: Choose members of a population that are easy to contact or reach.

Involves collecting a Sample from somewhere convenient to you such as your workplace, a local school or a public park.

Because these Samples are based on convenience to the researcher and not a broader Sample of the Population, Convenience Samples often show Undercoverage Bias.

Undercoverage Bias: When some members of a population are inadequately represented in a Sample.

Voluntary Response Sampling: Consists of members of a population who volunteer to participate in a Study.

Tend to suffer from Nonresponse Bias

Nonresponse Bias: When certain groups of people are less likely to provide responses.

Snowball Sampling: Researchers recruit initial participants to be in a study and then asks them to recruit other people to participate in the study.

Like a Snowball, the Sample Size gets bigger and bigger as more participants join in.

Purposive Sampling: Researchers select participants based on the purpose of their study.

Because of this, applicants who do not fit the profile are rejected.

The Researcher often intentionally excludes certain groups from the Sample to focus on a specific group they think is most relevant to their study.

Sampling Distributions

Point Estimate: Uses a Single Value to estimate a population Parameter.

Sampling Distributions: A Probability Distribution of a Sample Statistic

Sample statistics are based on randomly sampled data, and their outcomes cannot be predicted with certainty. You can use a sampling distribution to represent statistics such as the mean, median, standard deviation, range, and more.

Typically, data professionals compute sample statistics like the mean to estimate the corresponding population parameters.

Sampling Distributions represent the possible outcomes for a Sample Statistic like the mean.

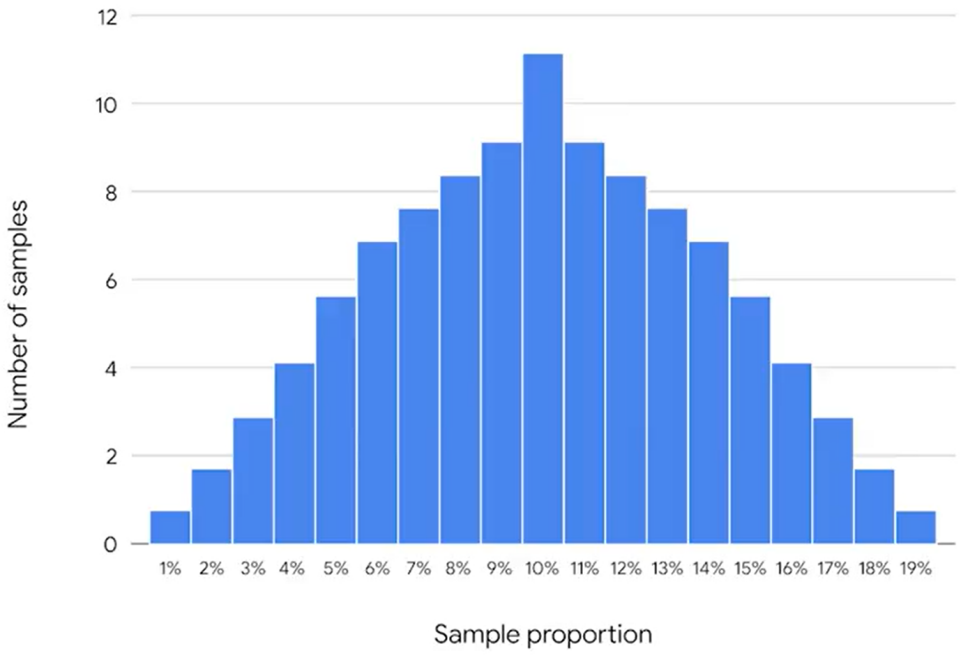


Figure 15: Sampling Distribution

Sampling Variability: How much an estimate varies between Samples.

You can use a Sampling Distribution to represent the frequency of all your different Sample Means

If your Sample Size is large enough, your Sample Mean will roughly equal the population Mean.

Standard Error of the Mean

To measure Sampling Variability, data professionals use the Standard Deviation of Sample Means to measure this Variability.

In Statistics the Standard Deviation of a Sample Statistic is called the Standard Error.

The Standard Error of the Mean measures the Variability among all your Sample Means.

- Larger Standard Error = Sample Means are more Spread out (More Variability)
- Smaller Standard Error = Sample Means are closer together (Less Variability)

The less Standard Error the more likely it is that your Sample Mean is an accurate estimate of the population Mean.

Note that the concept of Standard Error is based on the practice of repeated Sampling. In reality, researchers usually work with a single sample, its often too complicated, expensive or time consuming to take repeated Samples of a population. Instead Statisticians have derived a formula for calculating the Standard Error based on the Mathematical assumption of repeated Sampling.

Formula

$$SE = \frac{S}{\sqrt{n}}$$

S = The Sample Standard Deviation n = The Sample Size

As your Sample Size gets larger, your Standard Error gets smaller

This is because Standard Error measures the difference between your Sample Mean and the actual population Mean.

The Central Limit Theorem

Central Limit Theorem: The Sampling Distribution of the Mean approaches a Normal Distribution as the Sample Size increases.

In other words: As your Sample increases your Sampling Distribution assumes the shape of a Bell Curve, and if you take a large enough Sample of the population, the Sample Mean will be roughly equal to the population Mean.

There is no exact rule for how large a Sample Size needs to be in order for the Central Limit Theorem to apply. In general, a Sample Size of 30 or more is considered sufficient.

The Central Limit Theorem holds true for any population

You don't need to know the shape of your population Distribution in advanced in order to apply the Theorem.

If you collect a large enough Sample, the shape of your Sampling Distribution will follow a Normal Distribution.

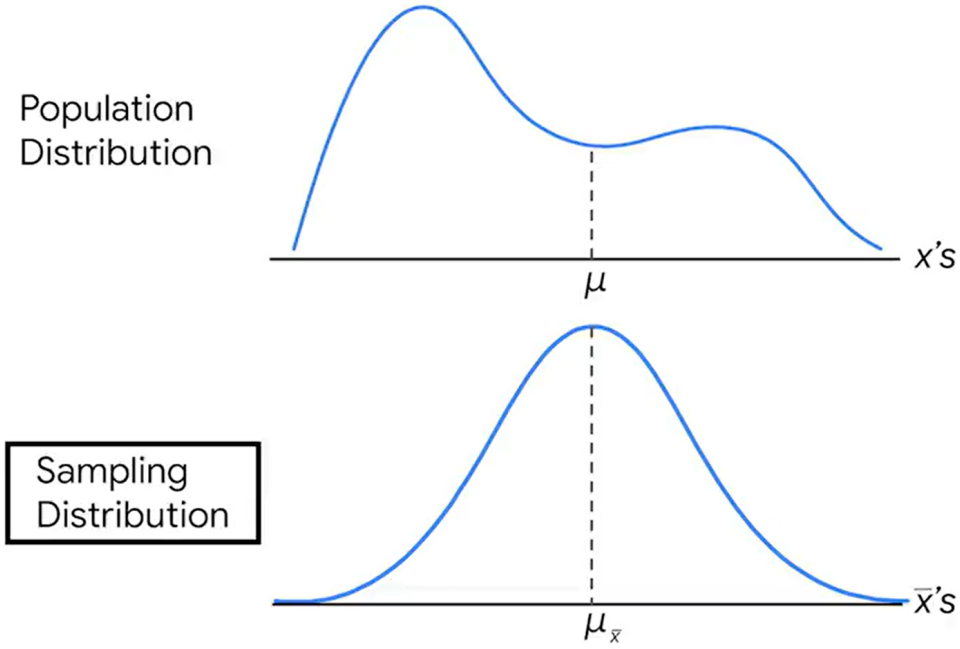


Figure 16: Central Limit Theorem Distribution

The Sampling Distribution of the Proportion

Population Proportions: The percentage of individuals or elements in a population that share a certain characteristic.

- Proportion's measure percentages or parts of a whole.
- There is also Variability for Proportions just like with the Mean.

You can use a Sampling Distribution to represent the frequency of all your different Sample Proportions.

The Central Limit Theorem also applies to Proportions.

As your Sample Size increases, the distribution of the Sample Proportion will be approximately normal. The overall average or Mean Proportion is located in the centre of the Curve.

Standard Error of the Proportion

You can Use the Standard Error of the Proportion to measure Sampling Variability.

This tells you how much a particular Sample Proportion is likely to differ from the true population Proportion.

The more Variability in your Sample Data the less likely it is that the Sample Proportion is an accurate estimate of the population Proportion

Formula

$$SE(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{n}$$

$\hat{p}$ = Population Proportion

n = Sample Size

Standard Error measures the difference between your Sample Proportion and the true population Proportion

As your Sample gets larger, your Sample Proportion gets closer to the true population Proportion.

The more accurate the estimate of the population Proportion, the smaller the Standard Error.

Typically the next step for a data Professional would be to use the Standard Error to construct a Confidence Interval. This describes the uncertainty of your estimate and gives your Stakeholders more detailed information about your results.

Sampling Distributions with Python

Packages

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
import statsmodels.api as sm
```

.sample()

Generating a Random Sample

Syntax * `n` refers to the desired Sample Size * `replace` indicates whether you are Sampling with or without replacement * `random_state` refers to the seed of the Random Number

Sampling with Replacement: When a population element can be selected more than one time

Sampling without Replacement: When a population element can be selected only one time

For example, suppose you have a jar that contains 100 unique numbers from 1-100, you want to select a random sample of numbers from the jar. After you pick the number from the jar, you can put the number aside or you can put it back in the jar. If you put the number back in the jar, it may be selected more than once **this is sampling with replacement**. If you put the number aside it can be selected only one time **this is sampling without replacement**. For the purpose of our example **you will sample with replacement**.

Random Seed: A starting point for generating random numbers.

You can use any arbitrary number to fix the random seed and give the random number generator a starting point.

Also going forward you can use the same random seed to generate the same set of numbers, In a later video you'll work with a sample again.

```
data = pd.read_csv('data.csv')

sampled_data = data.sample(n=50, replace=True, random_state=31208)
sampled_data
```

Sample Mean

```
estimate1 = sampled_data['column'].mean()
estimate1
```

This is a Point Estimate based on the population Mean based on your Random Sample

Due to Sampling Variability, the Sample Mean is usually not exactly the same as the population Mean.

Generating a 2nd Random Sample

```
estimate2 = data['column'].sample(n=50, replace=True, random_state=56910).mean()
estimate2
```

Due to Sampling Variability, this Sample Mean is different from the Sample Mean of your previous estimate (estimate1) but there really close.

Recall that the Central Limit Theorem tells you that when the sample size is large enough, the sample mean approaches a normal distribution. And as you sample more observations from a population, the sample mean gets closer to the population mean.

The larger your sample size, the more accurate your estimate of the population mean is likely to be

Now imagine you repeated this study 10,000 times and obtained 10,000 point estimates of the mean

According to the Central Limit Theorem the mean of your sampling distribution will be roughly equal to the population mean

You can use Python to compute the mean of the sampling distribution with 10,000 samples

Computing the Mean of the Sampling Distribution of 10,000 Samples

Lets review the code step by step

- First create an empty list to store the sample mean from each sample, name this `estimate_list`
- Set up a `for` loop with the `range` function, the loop will run 10,000 times and iterate over each number of the sequence
- Specify what you want to in each iteration of the loop. The `.sample()` function takes a random sample of `50` with replacement. The `.append()` function adds a single item to the existing list, in this case it appends the value of the sample mean to each item in the list.
- Create a new dataframe for your list of 10,000 estimates
- Name a new variable `estimate_df` to store your dataframe
- Name a new variable `mean_sample_means` then create the mean for your sampling distribution of 10,000 random samples

```
estimate_list = []
for i in range(10000):
    estimate_list.append(data['column'].sample(n=50, replace=True).mean())
estimate_df = pd.DataFrame(data={'estimate': estimate_list})
```

```
mean_sample_means = estimate_df['estimate'].mean()
mean_sample_means
```

The Mean of your Sampling Distribution is essentially identical to your population Mean of your complete dataset.

Visualisation of Sampling Distribution of 10,000 Estimates

To visualise the relationship between your sampling distribution of 10,000 estimates and the normal distribution we can plot both at the same time.

```
plt.hist(estimate_df['estimate'], bins=25, density=True, alpha=0.4, label = "histogram of sampl
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100) # generate a grid of 100 values from xmin to xmax.
p = stats.norm.pdf(x, mean_sample_means, stats.tstd(estimate_df['estimate']))
plt.plot(x, p,'k', linewidth=2, label = 'normal curve from central limit theorem')
plt.axvline(x=mean_sample_means, color='g', linestyle = 'solid', label = 'population mean')
plt.axvline(x=estimate1, color='r', linestyle = '--', label = 'sample mean of the first random
plt.axvline(x=mean_sample_means, color='b', linestyle = ':', label = 'mean of sample means of 1
plt.title("Sampling distribution of sample mean")
plt.xlabel('sample mean')
plt.ylabel('density')
plt.legend(bbox_to_anchor=(1.04,1))
plt.show()
```

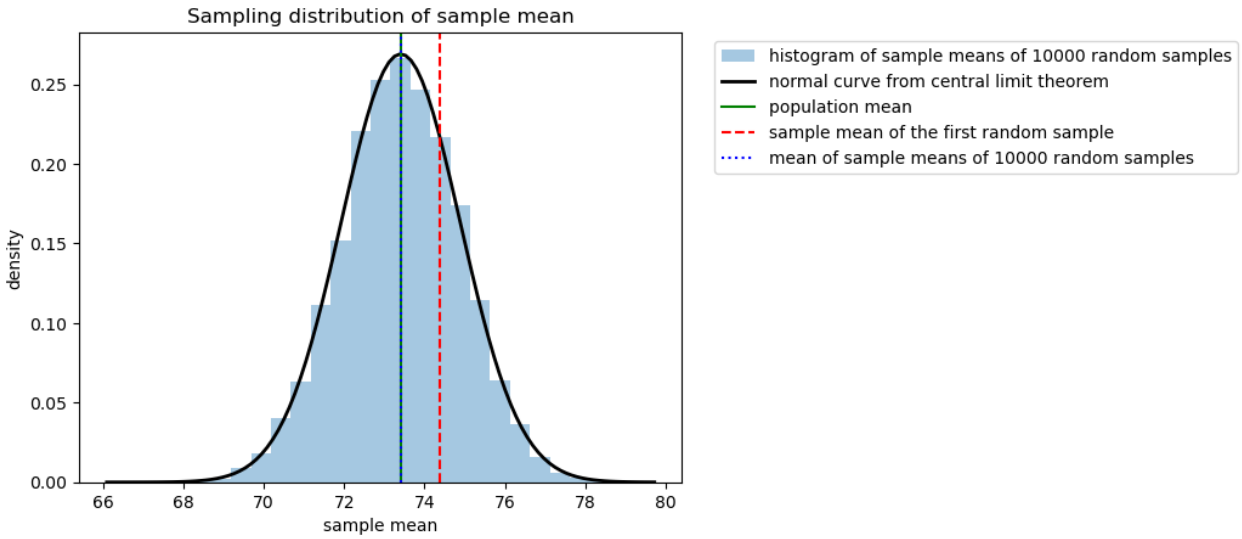


Figure 17: Sampling Distribution of Sample Mean

Activity: Calculating the Standard Error

```
standard_error = data['column'].std() / np.sqrt(len(data))
```

Confidence Intervals

Confidence Interval: A range of values that describes the uncertainty surrounding an Estimate.

Data professionals use Confidence Intervals as part of their job

You may be asked about Confidence Intervals in a job interview

Interval Estimate: Uses a range of values to estimate a population Parameter.

"A Point Estimate is useful, but a single element like 30lbs does not express the uncertainty built into any estimate. This uncertainty is due to the method of Random Sampling"

Confidence Intervals give data professionals a way to express the uncertainty caused by randomness and provide a more reliable Estimate. Confidence Level: Describes the likelihood that a particular Sampling Method will produce a Confidence Interval that includes the population Parameter.

e.g. Say you use a 95% Confidence Level to calculate a Confidence Interval between 28 and 32lbs.

This means if you took 100 Random Samples from the Penguin population and calculated a 95% Confidence Interval for each Sample, then approximately 95 of the 100 Intervals or 95% of the total would contain the actual population Mean. One such Interval will be the range of values between 28 and 32lbs.

Common Confidence Levels: * 90% * 95% * 99%

95% is a popular choice that is based on tradition in Statistical research and education, you can adjust the Confidence Level to meet the requirements of your analysis.

95% Confidence Level = Means that if you take repeated Random Samples from a population and construct a Confidence Interval for each Sample using the same Method. You can expect: - 95% of these Intervals capture the population Mean - 5% of the Intervals do not capture the population Mean

In practice, data professionals usually select one Random Sample and generate 1 Confidence Interval which may or may not contain the actual Mean. This is because repeated Random Sampling is often difficult, expensive and time consuming.

Confidence Intervals give data professionals a way to quantify the uncertainty due to Random Sampling.

In relation to an example "In other words, this method will produce an Interval that contains the population Mean with a success rate of 95%. That's a pretty good success rate." ## Correct Interpretation *e.g. Mean Weight*

Let's explore an example to get a better understanding of how to interpret a confidence interval. Imagine you want to estimate the mean weight of a population of 10,000 penguins. Instead of

weighing every single penguin, you select a sample of 100 penguins. The mean weight of your sample is 30 pounds. Based on your sample data, you construct a 95% confidence interval between 28 pounds and 32 pounds.

95% CI [28, 32]

Interpret the confidence interval

Earlier, you learned that the confidence level expresses the uncertainty of the estimation process. Let's discuss what 95% confidence means from a more technical perspective.

Technically, 95% confidence means that if you take repeated random samples from a population, and construct a confidence interval for each sample using the same method, you can expect that 95% of these intervals will capture the population mean. You can also expect that 5% of the total will not capture the population mean.

The confidence level refers to the long-term success rate of the **method**, or the estimation process based on random sampling.

For the purpose of our example, let's imagine that the mean weight of all 10,000 penguins is 31 pounds, although you wouldn't know this unless you actually weighed every penguin. So, you take a sample of the population.

Imagine you take 20 random samples of 100 penguins each from the penguin population, and calculate a 95% confidence interval for each sample. You can expect that approximately 19 of the 20 intervals, or 95% of the total, will contain the actual population mean weight of 31 pounds. One such interval will be the range of values between 28 pounds and 32 pounds.

Figure 18: Confidence Intervals

In practice, data professionals usually select one random sample and generate one confidence interval, which may or may not contain the actual population mean. This is because repeated random sampling is often difficult, expensive, and time-consuming. Confidence intervals give data professionals a way to quantify the uncertainty due to random sampling.

Common Misconceptions

1. A 95% Confidence Interval means that 95% of all the data values in your dataset fall within the Interval
2. A 95% Confidence Interval implies that 95% of all possible Sample Means fall within the range of the Interval
3. A Confidence Interval refers to the only possible source of error in your results

Note: When your interpreting a Confidence Interval remember that the uncertainty lies in an Estimation process based on Random Sampling. A 95% Confidence Level refers to the success rate of that process. In other words you can expect 95% of the Random Intervals you generate to capture the population Parameter.

The Confidence Level refers to the long-term success rate of the method or estimation process based on Random Sampling.

Pro-tip: Remember that a 95% Confidence Level refers to the success rate of the estimation process.

Steps to Constructing a Confidence Interval

1. Identify a Sample Statistic
2. Choose a Confidence Level
3. Find the Margin of Error
4. Calculate the Interval

Interval = Sample Statistic +/- Margin of Error

Margin of Error: The range of values above and below the Sample Statistic

The maximum expected difference between a population Parameter and a Sample Estimate

This is the amount that a data professional expects their estimate might vary from their actual amount.

$$\text{MoE} = \text{Z-Score} * \text{SE}$$

Confidence Level	Z-Score
90%	1.645
95%	1.96
99%	2.58

Identify a Sample Statistic

Mean or Proportion

e.g. Mean

Choose a Confidence Level

Confidence Level = 90%, 95%, 99%

e.g. 95%

Find the Margin of Error

$$\text{MoE} = \text{Z-Score} * \text{SE}$$

Standard Error for Means

$$SE = \frac{\sigma}{\sqrt{n}}$$

σ = The Proportion Standard Deviation when known, otherwise S. S = Sample Standard Deviation.
n = The Sample Size

Standard Error for Proportions

$$SE(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{n}$$

$\hat{p}$ = population Proportion

n = Sample Size

e.g. Mean

Confidence Level = 95% = Z-Score: **1.96**

$$SE = \frac{1.5}{\sqrt{100}} = 0.15$$

σ = std of population = 1.5

n = 100

$$\text{MoE} = \text{Z-Score} * \text{SE} = 1.96 * 0.15 = 0.294$$

Calculate the Interval

$$\text{Upper Limit} = \text{Sample Statistic} + \text{Margin of Error}$$

$$\text{Lower Limit} = \text{Sample Statistic} - \text{Margin of Error}$$

e.g. Mean

Sample Mean = 20.5 hrs

Sample Std = 1.7 hrs

Population std = 1.5 hrs

$$\text{Upper Limit} = \text{Sample Statistic} + \text{Margin of Error}$$

$$\text{Upper Limit} = 20.5 + 0.294 = 20.794 = 20:48(\text{hrs:min})$$

$$\text{Lower Limit} = \text{Sample Statistic} - \text{Margin of Error}$$

$$\text{Lower Limit} = 20.5 - 0.294 = 20.206 = 20:12(\text{hrs:min})$$

[20:12, 20:48]

95% CI [20:12, 20:48]

As the Confidence Level gets higher, the Confidence Interval gets wider

e.g. 99% Confidence Level

99% CI [20:07, 20:53]

This is because the wider Confidence Interval is more likely to include the actual population Parameter.

Note: For Calculating a Confidence Interval for a Proportion: As your Sample Size gets larger, your Confidence Interval gets narrower - As your Sample Size increases, your Margin of Error decreases.

Steps for Constructing a Confidence Interval of a Small Sample Size

Small Sample: T-Scores

For small sample sizes, you need to use a different distribution, called the t-distribution. Statistically speaking, this is because there is more uncertainty involved in estimating the standard error for small sample sizes. You don't need to worry about the technical details, which are beyond the scope of this course. For now, just know that if you're working with a small sample size, and your data is approximately normally distributed, you should use the t- distribution rather than the standard normal distribution. For a t-distribution, you use t-scores to make calculations about your data.

The graph of the t-distribution has a bell shape that is similar to the standard normal distribution. But, the t-distribution has bigger tails than the standard normal distribution does. The bigger tails indicate the higher frequency of outliers that come with a small dataset. As the sample size increases, the t-distribution approaches the normal distribution. When the sample size reaches 30, the distributions are practically the same, and you can use the normal distribution for your calculations.

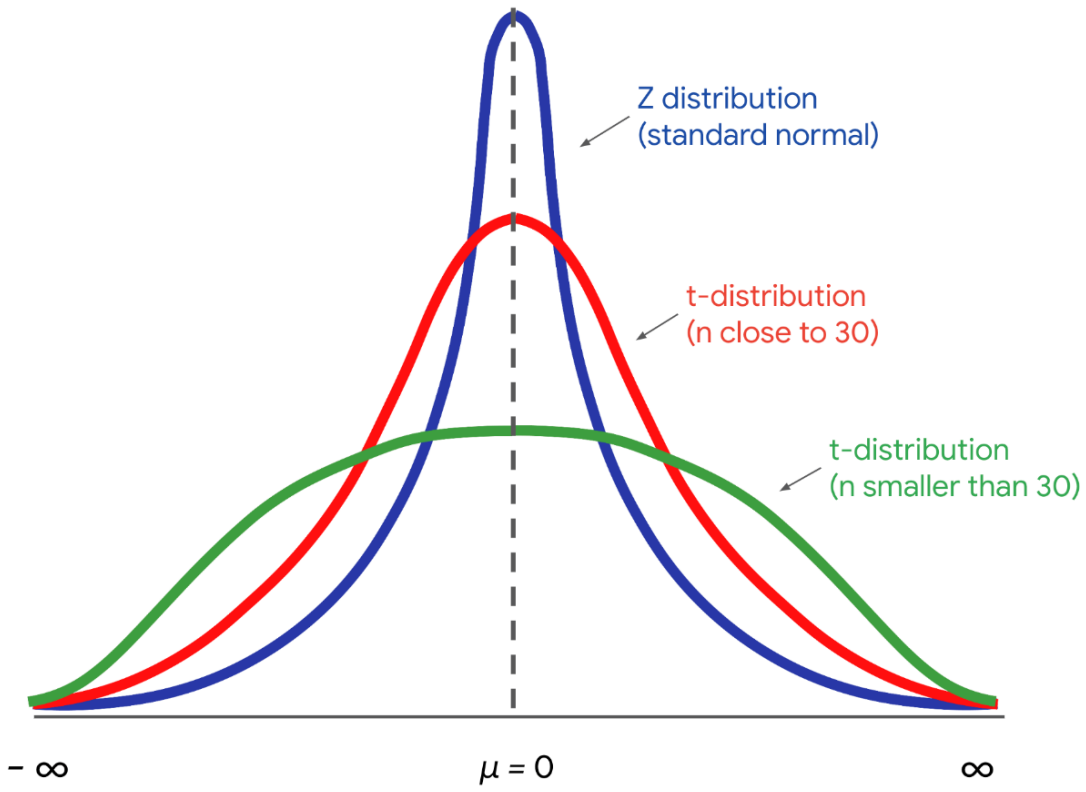


Figure 19: z and t Distribution

Construct the Confidence Interval

Will follow an example from Certificate reading, if you want full details of the example then go to the reading: Construct a Confidence Interval for a Small Sample Size

Step 1: Identify the Sample Statistic

First, identify your sample statistic. Your sample represents the average emissions rate for 15 engines. You're working with a sample mean.

Step 2: Choose a Confidence Level

Next, choose a confidence level. The engineering team requests that you choose a 95% confidence level.

Step 3: Find the Margin of Error

Your third step is to find the margin of error. For a small sample size, you calculate the margin of error by multiplying the t-score by the standard error.

The t-distribution is defined by a parameter called the degree of freedom. In our context, the degree of freedom is the sample size - 1, or 15-1 = 14. Given your degree of freedom and your confidence level, you can use a programming language like Python or other statistical software to calculate your t-score.

Based on a degree of freedom of 14, and a confidence level of 95%, your **t-score is 2.145.**

Now you can calculate the standard error, which measures the variability of your sample statistic.

Here's the formula for the standard error of the mean that you've used before:

Standard Error (Means)

$$SE(x) = \frac{S}{\sqrt{(n)}}$$

In the formula, the letter s refers to sample standard deviation, and the letter n refers to sample size.

Your sample standard deviation is 35, and your sample size is 15. The calculation gives you a **standard error of about 9.04.**

The margin of error is your t-score multiplied by your standard error. This is **2.145 * 9.04 = 19.39**.

Step 4: Calculate the Interval

Finally, calculate your confidence interval. The upper limit of your interval is the sample mean plus the margin of error. This is 430 + 19.39 = 449.39 grams of CO2 per mile.

The lower limit is the sample mean minus the margin of error. This is 430 - 19.39 = 410.61 grams of CO2 per mile.

You have a 95% confidence interval that stretches from 410.61 grams of CO2 per mile to 449.39 grams of CO2 per mile.

95% CI [410.61, 449.39]

The confidence interval gives the engineering team important information. The upper limit of your interval is below the target of 460 grams of CO2 per mile. This result provides solid statistical evidence that the emissions rate for the new engine will meet emissions standards.

Note: Confidence intervals for small sample sizes only deal with population means, and not population proportions. The statistical reason for this distinction is rather technical, so you don't need to worry about it for now.

Confidence Intervals in Python

Packages

```
import numpy as np
import pandas as pd
from scipy import stats
```

Making the Sampled Data

```
data = pd.read_csv('data.csv')
sampled_data = data.sample(n=50, replace=True, random_state=31208)
sampled_data
```

stats.norm.interval()

Syntax * `alpha` refers to the Confidence level * `loc` refers to the Sample Mean * `scale` refers to the Sample Standard Error

For `loc`

```
sample_mean = sampled_data['column'].mean()
sample_mean
```

Calculating the Standard Error

for `scale`

`.shape` function returns the number of rows and columns in a dataframe `.shape[0]` returns only the number of rows, which is the same number as your Sample Size.

```
estimated_standard_error = sampled_data['column'].std() / np.sqrt(sampled_data.shape[0])
```

Constructing the Interval

Code from Video

```
stats.norm.interval(alpha=0.95, loc=sample_mean, scale=estimated_standard_error)
```

Corrected Code

```
stats.norm.interval(0.95, loc=sample_mean, scale=estimated_standard_error)
```

e.g. Video Example

(np.float64(71.42241096968617), np.float64(77.02478903031381))

Confidence Interval: 95% CI [71.4, 77.0]

Activity

Choose your Sample Statistic

```
sample_mean = data['column'].mean()
sample mean
```

Choose your Confidence Level

```
confidence_level = 0.95
confidence_level
```

Calculate your Margin of Error

```
z_value = 1.96
standard_error = data['column'].std() / np.sqrt(data.shape[0])

margin_of_error = standard_error * z_value
```

Calculate your Interval

```
upper_ci_limit = sample_mean + margin_of_error
lower_ci_limit = sample_mean - margin_of_error
(lower_ci_limit, upper_ci_limit)
```

Hypothesis Tests

Hypothesis Testing: A Statistical Procedure that uses Sample data to evaluate an assumption about a population Parameter.

Data professionals conduct a Hypothesis Test to decide whether the evidence from their Sample data supports either the Null Hypothesis or the Alternative Hypothesis.

Statistical Significance: The claim that the results of a test or experiment are not explainable by chance alone.

Steps for Performing a Hypothesis Test

1. State the Null Hypothesis and the Alternative Hypothesis
2. Choose a Significance Level
3. Find the P-Value
4. Reject or Fail to Reject the Null Hypothesis

Null Hypothesis: A statement that is assumed to be true unless there is convincing evidence to the contrary.

The Null Hypothesis typically assumes that there is no effect in the population, and that your observed data occurs by chance.

Alternative Hypothesis: A statement that contradicts the Null Hypothesis and is accepted as true only if there is convincing evidence for it.

The Alternative Hypothesis typically assumes that there is an effect in the population, and that your observed data does not occur by chance

Significance Level: The Probability of Rejecting the Null Hypothesis when it is true.

P-Value: The Probability of observing results as or more extreme than the observed when the Null Hypothesis is true.

A lower P-Value means there is stronger evidence for the Alternative Hypothesis

Drawing a Conclusion

If P-Value $<$ (less than) Significance Level: Reject the Null Hypothesis

If P-Value $>$ (greater than) Significance Level: Fail to Reject the Null Hypothesis

	Null hypothesis (H_0)	Alternative hypothesis (H_a)
Claims	There is no effect in the population.	There is an effect in the population.
Language	• No effect • No difference • No relationship • No change	• An effect • A difference • A relationship • A change
Symbols	Equality (=, ≤, ≥)	Inequality (≠, <, >)

Figure 20: Null vs. Alternative Hypothesis

Types of Errors

Type I Error (False Positive): The Rejection of a Null Hypothesis that is actually true.

A Significance Level of 5% means that you're willing to accept a 5% chance that you are wrong when you Reject the Null Hypothesis.

To reduce your chance of making a Type I Error, choose a lower Significance Level

However choosing a lower Significance Level means your more likely to make a Type II Error or False Negative.

Type II Error (False Negative): The failure to Reject a Null Hypothesis which is actually false.

The Probability of making a Type I Error is called alpha (α). Your Significance Level or alpha (α) represents the probability of making a Type I Error.

The Probability of making a Type II Error is called beta (β), and beta is related to the power of a Hypothesis Test (power = 1 - β). Power refers to the likelihood that a test can correctly detect a real effect when there is one.

You can reduce your risk of making a Type II Error by ensuring your test has enough power. In data work, power is usually set at 0.80 or 80%. The higher the statistical power, the lower the Probability of making a Type II Error. To increase power, you can increase your Sample Size or your Significance Level

	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct Outcome! (True positive)
Fail to reject null hypothesis	Correct Outcome! (True negative)	Type II Error (False negative)

Figure 21: Type I and II Errors

One-Sample Tests

One-Sample Test: Determines whether or not a population Parameter like a Mean or Proportion is equal to a specific value.

One-Sample z-test

One-Sample z-test Assumptions

- **The data is a Random Sample of a Normally Distributed population**
- **The population Standard Deviation is known**

Test Statistic: A value that shows how closely your observed data matches the Distribution expected under the Null Hypothesis.

For a Z-Test the Test Statistic is a Z-Score

One-Sample z-test Formula for Means

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

$\bar{x}$ = Sample Mean

μ = Population Mean

σ = Population Standard Deviation

n = Sample Size

e.g.

$$Z = \frac{\bar{38} - 40}{\frac{5}{\sqrt{50}}} = -2.82$$

In this example the Z-Score is far to the left, almost 3 std dev below the mean. For a Normal Distribution, the Probability of getting a value less than your Z-Score (-2.82) is calculated by taking the area under the Curve to the left of the Z-Score. This is called a ***left-tailed test*** *because your P-Value is located on the left tail of the Distribution. The area under this part of the Curve is the same as your P-Value*

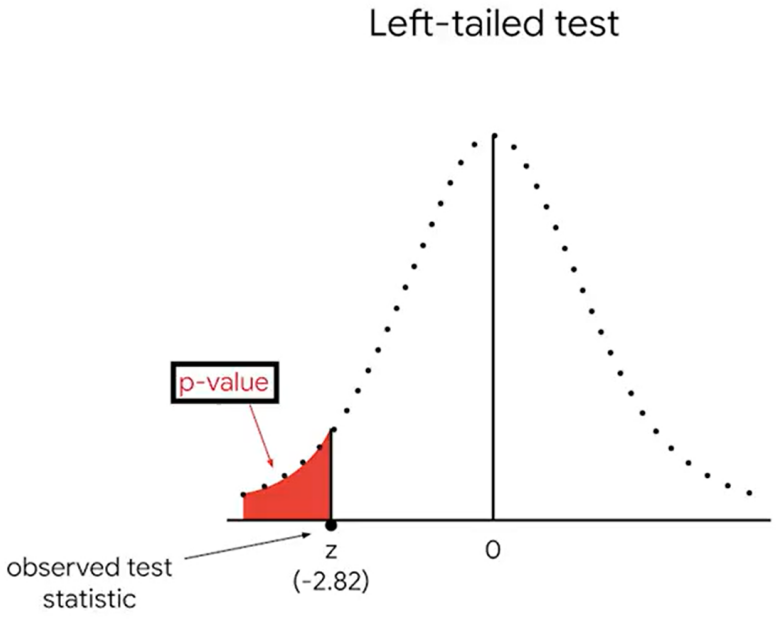


Figure 22: Left-tailed test

In a different testing scenario, your Test Statistic might be 2.45, and you might be interested in values as high or higher than the Z-Score 2.45. In that case your P-Value would be located on the right-tail of the Distribution and you would be conducting a ***right-tailed test***

One-Tailed and Two-Tailed Tests

- Left-Tailed Test: When the H_a states that the actual value of the Parameter is less than the value in the H_o.
- Right-Tailed Test: When the H_a states that the actual value of the Parameter is greater than the value in the H_o.
- Two-Tailed Test: When the H_a states that the actual value of the Parameter does not equal the value in the H_o.

Note: P-Value for a Two-Tailed Test is always two times the P-Value for a One-Tailed Test.

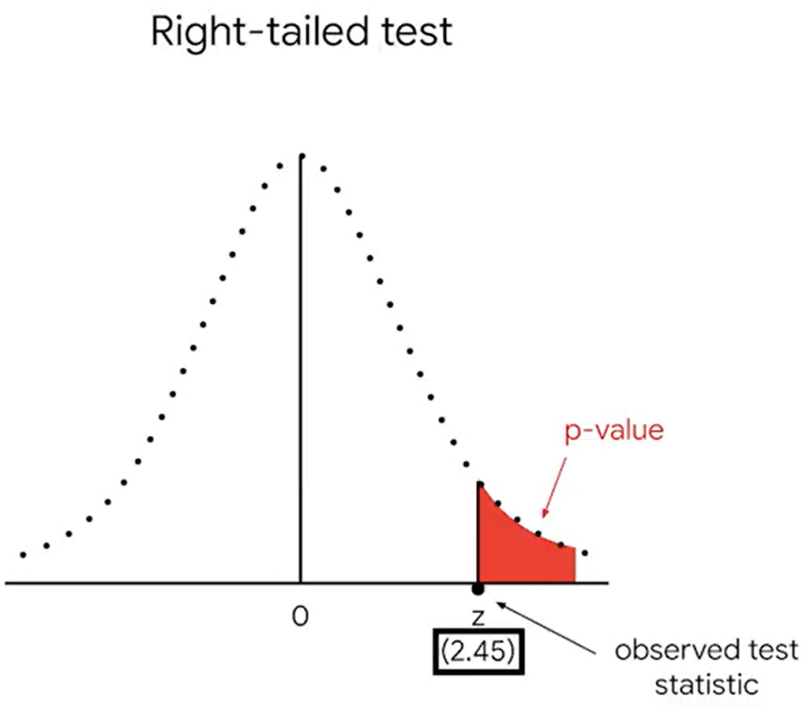


Figure 23: Right-tailed test

A two-tailed test results when the alternative hypothesis states that the actual value of the parameter does not equal the value in the null hypothesis.

Two-Sample Tests

Two-Sample Test: Determines whether or not two population Parameters such as two Means or two Proportions are equal to each other.

Two-Sample t-test

Two-Sample t-test for Means Assumptions

- **The two Samples are independent of each other**
- **For each Sample, the data is drawn Randomly from a Normally Distributed population**
- **The population Standard Deviation is unknown**

In practice the population Standard Deviation is usually unknown because its difficult to get complete data on large populations. So data professionals use a t-test for practical applications.

t-scores are based on the t-distribution

t-distribution has bigger tails than the standard Normal Distribution does.

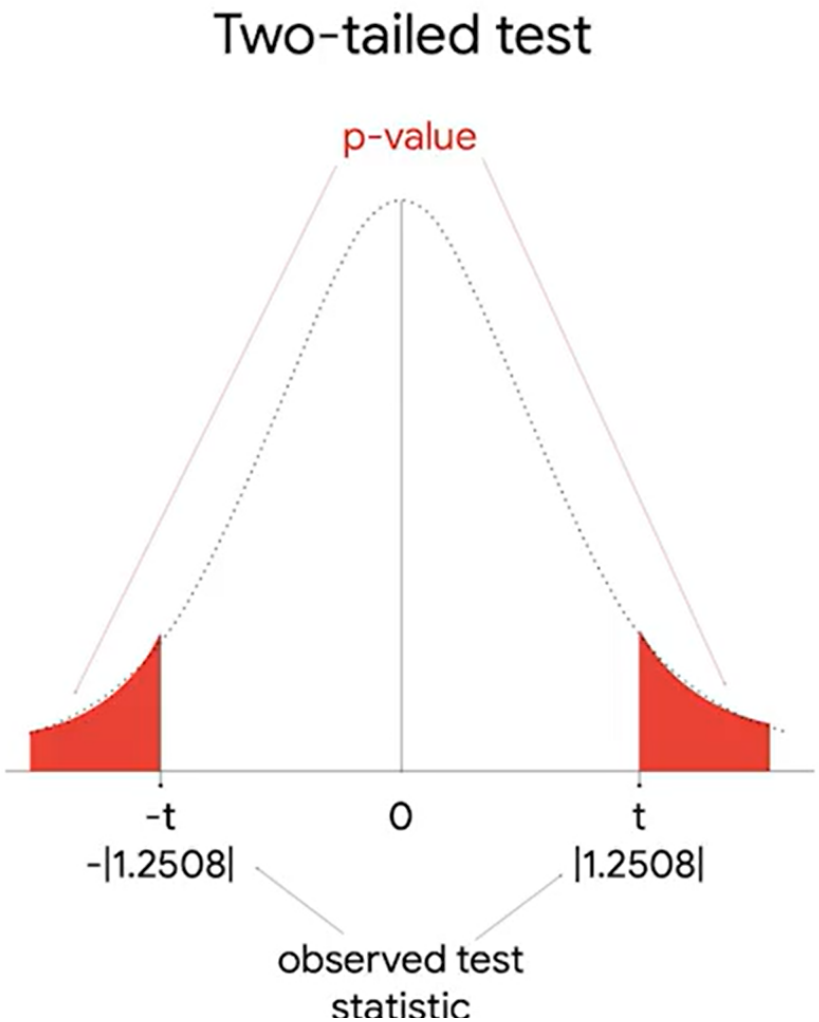


Figure 24: Two-tailed test

The bigger tails indicate the higher frequency of outliers that come with small datasets.

As the Sample Size increases, the t-distribution approaches the Normal Distribution

Two-Sample t-test Formula for Means

$$t = \frac{(\bar{x_1} - \bar{x_2})}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}}$$

$\bar{x_1}$, $\bar{x_2}$ = Sample Means

n_1, n_2 = Sample Sizes

s_1^2, s_2^2 = Sample Variances

Two-Sample z-test

For Technical reasons t-tests do not apply for Proportions

Two-Sample z-test Formula for Proportions

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p_0}(1 - \hat{p_0})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$\hat{p}_{1/2}$ = Sample Proportions for both groups

$n_{1/2}$ = Sample Sizes for both groups

$\hat{p_0}$ = Pooled Proportion

Pooled Proportion ($\hat{p_0}$): Weighted average of the Proportions from your 2 Samples.

Experimental Design: Refers to planning an experiment in order to collect data to answer your research question.

Pooled Proportions

Pooled: Combining multiple Sample estimates into one single, more stable estimate, assuming they came from the same Population (under H_0).

Used when H_0 assumes equality (e.g. $p_1 = p_2$)

Used when you want to leverage more data to improve the accuracy of a shared Parameter estimate

Shows up in: * Pooled Variance in t-tests (when assuming population Variances) * Pooled Proportions in z-tests (when assuming population Proportions)

For performing Two-Tailed z-tests for population Proportions

We use the Pooled Proportion when the H_0 assumes the two population Proportions are equal (e.g. $p_1 = p_2$).

Logic

We should combine (pool) the information from both Samples to get the best estimate of the common Proportion.

Formula

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$x_{1/2}$ = Number of successes in Samples 1 and 2

$n_{1/2}$ = Sample Sizes

$\hat{p}$ = Pooled Proportion (i.e. Combined Success Rate)

Hypothesis Testing in Python

Packages

```
import pandas as pd
from scipy import stats
```

Making the 2 Samples

```
data = pd.read_csv('data.csv')

state21 = data[data['statename'] == 'state21']
state28 = data[data['statename'] == 'state28']

sampled_state21 = state21.sample(n=20, random_state=13490, replace=True)
sampled_state28 = state28.sample(n=20, random_state=39103, replace=True)
```

Means of the 2 Sample

```
sampled_state21['column'].mean()
sampled_state28['column'].mean()
```

Observed Difference in Means

```
sampled_state21['column'].mean() - sampled_state28['column'].mean()
```

stats.ttest_ind()

Syntax

- `a` refers to observations from your first Sample
- `b` refers to observations from your second Sample
- `equal_var` is a Boolean or True/False Statement which indicates whether the population Variance of the two Samples is assumed to be equal.
- `alternative` specifies the direction of the Hypothesis Test, whether your performing a Two-Tailed or One-Tailed Test. Has 3 Arguments: `'two-sided'` = Two-Tailed, `'less'` = left, `'greater'` = right. If you omit the `alternative` argument it defaults to `'two-sided'`.

If you don't have access to the data for the entire population, you don't want to assume anything about the Variance

To avoid making a wrong assumption, set `equal_var` to `False`.

```
stats.ttest_ind(a=sampled_state21['column'], b=sampled_state28['column'], equal_var=False)
```

e.g. Output

TtestResult(statistic=np.float64(2.8980444277268735), pvalue=np.float64(0.006421719142765237), df=np.float64(35.20796133045557))

Activity

Two-Sample t-test with alternative = 'less'

```
stats.ttest_ind(a=data['column'], b=data['column'], equal_var=False, alternative='less')
```

alternative way of writing the code out

```
tstat, pvalue = stats.ttest_ind(a=data['column'], b=data['column'], equal_var=False,
                                alternative='less')

print(tstat)
print(pvalue)
```

stats.ttest_1samp()

One-Sample t-test with popmean and alternative = 'greater'

```
stats.ttest_1samp(a=data['column'], popmean=10, alternative='greater')
```

- `popmean` is the population Mean you are comparing your Sample against.

alternative way of writing the code out

```
tstat, pvalue = stats.ttest_1samp(a=data['column'], popmean=10, alternative='greater')

print(tstat)
print(pvalue)
```

Certificate Readings

- Measures of Central Tendency: The Mean, The Median, and The Mode
- Measures of Dispersion: Range, Variance, and Standard Deviation
- Measures of Position: Percentiles and Quartiles
- Fundamental Concepts of Probability
- The Probability of Multiple Events
- Calculate Conditional Probability for Dependent Events
- Calculate Conditional Probability with Bayes's Theorem
- Discrete Probability Distributions
- Model Data with the Normal Distribution
- The Stages of the Sampling Process
- The Relationship between Sample and Population
- Probability Sampling Methods
- Non-Probability Sampling Methods
- The Sampling Distribution of the Mean
- Infer Population Parameters with the Central Limit Theorem
- Confidence Interval: Correct and Incorrect Interpretations
- Construct a Confidence Interval for a Small Sample Size
- Differences between the Null and Alternative Hypotheses
- Type I and Type II Errors
- Determine if Data has Statistical Significance
- One-Tailed and Two-Tailed Tests

Python Notebooks

- Compute Descriptive Statistics with Python
- Work with Probability Distributions in Python
- Sampling Distributions with Python
- Confidence Intervals with Python
- Exemplar_Explore Confidence Intervals
- Activity_Explore Confidence Intervals
- Use Python to Conduct a Hypothesis Test
- Exemplar_Explore Hypothesis Testing
- Activity_Explore Hypothesis Testing
- Course 4 End of Course Portfolio Project

Applied Statistics for Data Science with Python is a practical guide to the statistical foundations that drive modern data analysis, machine learning and data-driven decision making. Designed for students, analysts and professionals, this book bridges classical statistical theory with hands-on implementation in Python.

This book covers everything from descriptive statistics to probability and sampling methodology with their distributions to interpreting and constructing confidence intervals and hypothesis testing. With each of the techniques being built up from its first principles and then connected to real-world data analysis. Rather than treating statistics as abstract mathematics, the focus is on interpretation, assumptions and how results should be used to make informed conclusions from data.

Python is used throughout the entirety of this book to demonstrate how statistical concepts translate into working code. These examples illustrate data preparation, statistical modelling and the interpretation of analytical outputs using industry-standard tools. The objective is to move beyond numerical outputs and move towards meaningful interpretation and evidence-based decision-making.

Whether used as a learning resource or a reference guide, this book offers a comprehensive foundation for statistical analysis in contemporary data science.

Scan for my portfolio