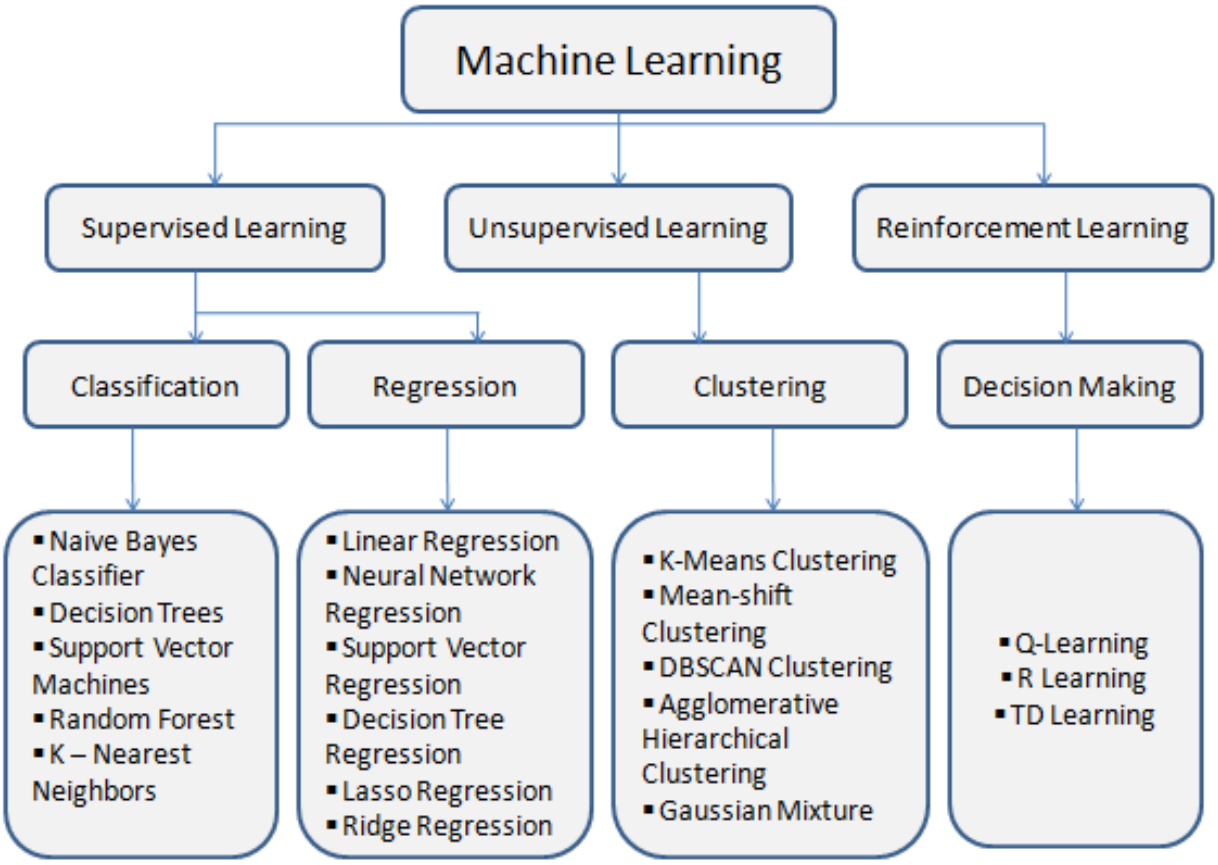
Machine learning

Machine learning (**ML**) is an umbrella term for solving problems for which development of algorithms by human programmers would be cost-prohibitive, and instead the problems are solved by helping machines 'discover' their 'own' algorithms, without needing to be explicitly told what to do by any human-developed algorithms. Recently, generative artificial neural networks have been able to surpass results of many previous approaches. Machine learning approaches have been applied to large language models, computer vision, speech recognition, email filtering, agriculture and medicine, where it is too costly to develop algorithms to perform the needed tasks.

The mathematical foundations of ML are provided by mathematical optimization (mathematical programming) methods. Data mining is a related (parallel) field of study, focusing on exploratory data analysis through unsupervised learning.

ML is known in its application across business problems under the name predictive analytics. Although not all machine learning is statistically-based, computational statistics is an important source of the field's methods.



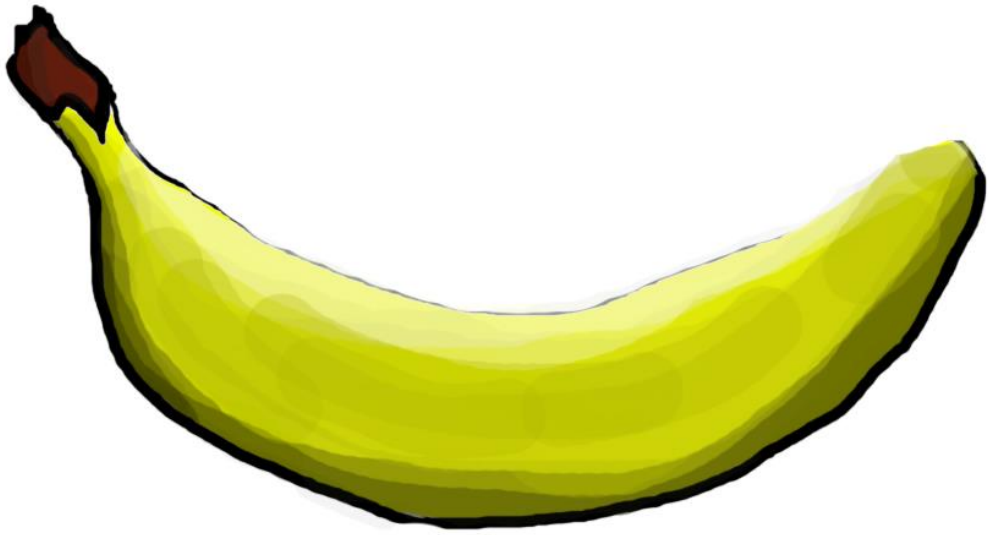
Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Supervised and Unsupervised learning:

Supervised learning: Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well-labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.
For instance, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all the different fruits one by one like this:

- If the shape of the object is rounded and has a depression at the top, is red in color, then it will be labeled as –**Apple**.
- If the shape of the object is a long curving cylinder having Green-Yellow color, then it will be labeled as –**Banana**.

Now suppose after training the data, you have given a new separate fruit, say Banana from the basket, and asked to identify it.



Since the machine has already learned the things from previous data and this time has to use it wisely. It will first classify the fruit with its shape and color and would confirm the fruit name as BANANA and put it in the Banana category. Thus the machine learns the things from training data(basket containing fruits) and then applies the knowledge to test data(new fruit).

Supervised learning is classified into two categories of algorithms:

- **Classification**: A classification problem is when the output variable is a category, such as “Red” or “blue” , “disease” or “no disease”.
- **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Supervised learning deals with or learns with “labeled” data. This implies that some data is already tagged with the correct answer.

Types:-

- Regression
- Logistic Regression
- Classification
- Naive Bayes Classifiers
- K-NN (k nearest neighbors)
- Decision Trees
- Support Vector Machine

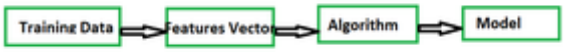
Advantages:-

- Supervised learning allows collecting data and produces data output from previous experiences.

- Helps to optimize performance criteria with the help of experience.
- Supervised machine learning helps to solve various types of real-world computation problems.
- It performs classification and regression tasks.
- It allows estimating or mapping the result to a new sample.
- We have complete control over choosing the number of classes we want in the training data.

Disadvantages:-

- Classifying big data can be challenging.
- Training for supervised learning needs a lot of computation time. So, it requires a lot of time.
- Supervised learning cannot handle all complex tasks in Machine Learning.
- Computation time is vast for supervised learning.
- It requires a labelled data set.
- It requires a training process.



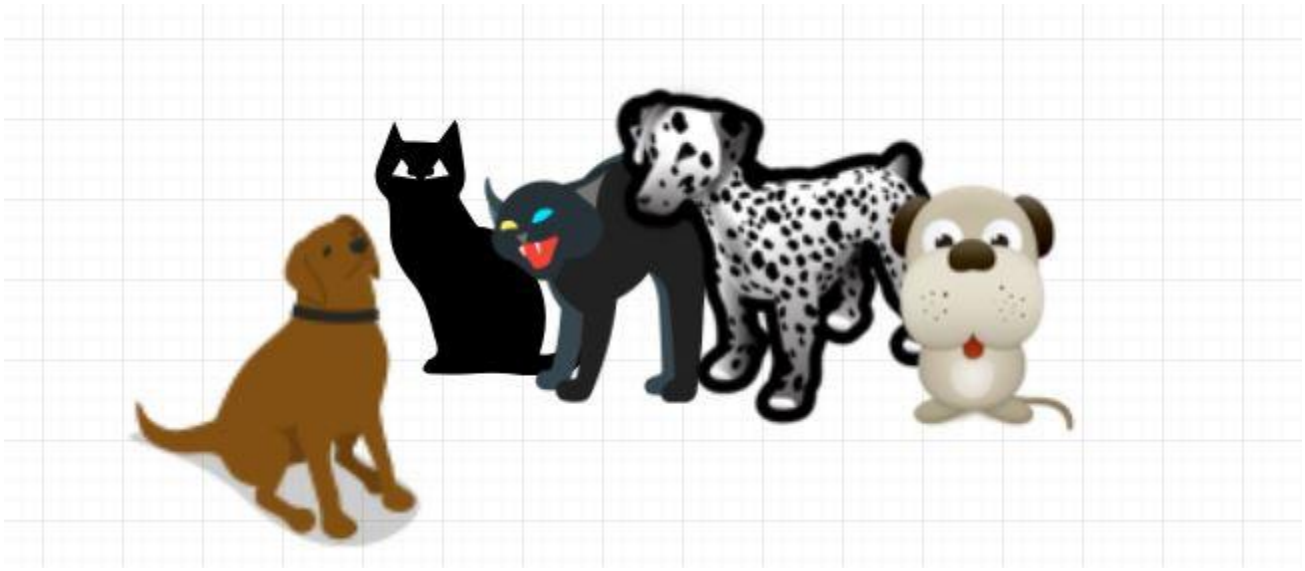
Steps

Unsupervised learning

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.

For instance, suppose it is given an image having both dogs and cats which it has never seen.



Thus the machine has no idea about the features of dogs and cats so we can't categorize it as 'dogs and cats '. But it can categorize them according to their similarities, patterns, and differences, i.e., we can easily categorize the above picture into two parts. The first may contain all pics having **dogs** in them and the second part may contain all pics having **cats** in them. Here you didn't learn anything before, which means no training data or examples.
It allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with unlabelled data.

Unsupervised learning is classified into two categories of algorithms:

- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Types of Unsupervised Learning:-

Clustering

1. Exclusive (partitioning)
2. Agglomerative
3. Overlapping
4. Probabilistic

Clustering Types:-

1. Hierarchical clustering
2. K-means clustering
3. Principal Component Analysis
4. Singular Value Decomposition
5. Independent Component Analysis

Statistics

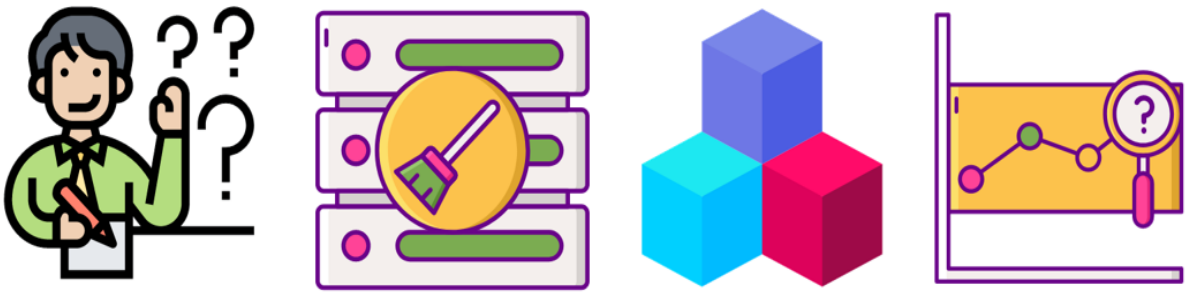
Basic statistics

Statistics is a core component of data analytics and machine learning. It helps you analyze and visualize data to find unseen patterns. If you are interested in machine learning and want to grow your career in it, then learning statistics along with programming should be the first step. In this article, you will learn all the concepts in statistics for machine learning.

What Is Statistics?

Statistics is a branch of mathematics that deals with collecting, analyzing, interpreting, and visualizing empirical data. Descriptive statistics and inferential statistics are the two major areas of statistics. Descriptive statistics are for describing the properties of sample and population data (what has happened). Inferential statistics use those properties to test hypotheses, reach conclusions, and make predictions (what can you expect).

Use of Statistics in Machine Learning



- Asking questions about the data
- Cleaning and preprocessing the data
- Selecting the right features

- Model evaluation
- Model prediction

With this basic understanding, it’s time to dive deep into learning all the crucial concepts related to statistics for machine learning.

Variables:

The values that are altering according to circumstances are referred to as variables. A variable can occurs in any form, such as trait, factor or a statement that will constantly be changing according to the changes in the applied environment. Such variables in statistics are broadly divided into four categories such as independent variables, dependent variables, categorical and continuous variables. Apart from these, quantitative and qualitative variables hold data as nominal, ordinal, interval and ratio. Each type of data has unique attributes.

Different Types of Variables in Statistics

In statistics, the variable is an algebraic term that denotes the unknown value that is not a fixed value which is in numerical format. Such types of variables are implemented for many types of research for easy computations. So there are many different types of variables available that can be applied in varied domains. Many other variables are discussed in minimally are listed are active variable which the researcher evaluates. A variable that occurs before the independent variable is called an antecedent variable.

1. Independent Variables

The independent variable is the one that is computed in research to view the impact of dependent variables. It is also called as resultant variables, predictor or experimental variables. For example, A manager asks 100 employees to complete a project. He should know the capacity of the individual employee. He wants to know the reason behind smart guys and failure guys. The first reason is that some will be working hard for day and night to complete the project within the estimated time, and the other one is

that some guys are born intelligent and smarter than others. The variable which is similar to an independent variable is called a covariate variable but is impacted by the dependent variable but not as common as a variable of interest.

2. Dependent Variables

The dependent variable is also called a criterion variable which is applied in non-experimental circumstances. The dependent variable has relied on the independent variable. From the above-mentioned example, the project's productivity or completion is the main criteria that are dependent on estimated time and IQ. Here, the independent variables are IQ and estimated time, which may or may not reflect in an employee's productivity. So the extension of estimated time or enhancing the IQ of a person doesn't make any sense in employee's productivity as it is not predictable.

Hence, the managers' focus is to work on the independent variables such as allotted time and IQ that leads to certain changes in employee's productivity that are the dependent variables. So both the variables are connected in some measures. The variables which get affected by other variables in econometrics is termed as endogenous variables. A hidden variable impacts the relationship between the dependent and independent variable called lurking variables. When an independent variable is not impacted by any other variables and is restricted to a certain extent are called an explanatory variable.

3. Categorical Variables

It is a wide category of variable which is infinite and has no numerical data. These variables are called as qualitative variables or attribute variable in terms of statistics

software. Such variables are further divided into nominal variables, ordinal and dichotomous variables. Nominal variables don't have any intrinsic order. For instance, a developer classifies his environment into different types of networks based on their structure, such as P2P, cloud computing, pervasive computing, IoT. So here, the type of network is a nominal variable comprised of four categories. The varied categories present in the nominal variable can be known as the nominal variable levels or groups. Dichotomous variables are also called binary values, which have only two categories.

For example, if we question a person that he owns a car, he would reply only with yes or no. such types of two distinct variables that are nominal are called as dichotomous. It just accounts for only two values, such as 0 or 1. It could be yes or no, short or long, etc. Ordinal variables are nominal variables that include two or multiple categories. If you see any hotel feedback form, it has five ratings such as excellent, good, better, poor and very poor. So we can rank the level with the help of ordinal variables that hold meaning to the research. It is unambiguous, and values can be considered for decision making.

4. Continuous Variables

The variables which measure some count or quantity and don't have any boundaries are termed as continuous variables. It can be segregated into ratio or interval, or discrete variables. Interval variables have their centralized attribute, which is calibrated along with a range with some numerical values. The example can be temperature calibrated in Celsius or Fahrenheit doesn't give any two different meaning; they display the optimum temperature, and it's strictly not a ratio variable.

It can account for only a certain set of values, such as several bikes in a parking area are discrete as the floor holds only a limited portion to park bikes. Ratio variables occur with intervals; it has an extra condition that zero on any measurement denotes that there is no value of that variable. In simple, the distance of four meters is twice the distance of two meters. It operates on the ratio of measurements. Apart from these mentioned variables, a dummy variable can be applied in regression analysis to establish a relationship to unlinked categorical variables. For instance, if the user had categories ”has pet” and ”owns a home” can assign as 1 to ”’has pet” and 0 to ”’owns a home”.

A factor that remains constant in an experiment is termed as a control variable. In an experiment, if the scientist wants to test the plant’s light for its growth, he should control the value of water and soil quality. The additional variable which has a hidden impact on the obtained experimental values are called confounding variables.

Random Variable:

A random variable is a rule that assigns a numerical value to each outcome in a sample space. Random variables may be either discrete or continuous. A random variable is said to be discrete if it assumes only specified values in an interval. Otherwise, it is continuous. We generally denote the random variables with capital letters such as X and Y. When X takes values 1, 2, 3, …, it is said to have a discrete random variable.

Population and Sample

Population:

In statistics, the population comprises all observations (data points) about the subject under study.

An example of a population is studying the voters in an election. In the 2019 Lok Sabha elections, nearly 900 million voters were eligible to vote in 543 constituencies.

Sample:

In statistics, a sample is a subset of the population. It is a small portion of the total observed population.

An example of a sample is analyzing the first-time voters for an opinion poll.

Population Distribution, Sample Distribution and Sampling Distribution

Population Distribution

Population distribution refers to the distribution of a particular characteristic or variable among all individuals or units in a specific population. For example, the population distribution of heights in a country would refer to the distribution of heights among all individuals living in that country.

The population is the whole set of values, or individuals, you are interested in. For example, if you want to know the average height of the residents of India, that is your population, i.e., the population of India.

Population characteristic are mean (μ), Standard deviation (σ) , proportion (P) , median, percentiles etc. The value of a population characteristic is fixed. This characteristics are called population distribution. They are symbolized by Greek characters as they are population parameters.

Sample Distribution

Sample distribution refers to the distribution of a particular characteristic or variable among the individuals or units selected from a population. For example, if we take a random sample of 100 individuals from a country's population and measure their heights, the distribution of heights in the sample is called the sample distribution.

The sample is a subset of the population, and is the set of values you actually use in your estimation. Let's think 1000 individual you have selected for your study to know about average height of the residents of India. This sample has some quantity computed from values e.g. mean (x), Standard deviation (s), sample proportion etc. This is called sample distribution. The mean and standard deviation are symbolized by Roman characters as they are sample statistics.

Sampling Distribution

Sampling distribution refers to the distribution of a statistic (such as the mean, standard deviation, etc.) calculated from multiple random samples of the same size drawn from a population. For example, if we take multiple random samples of 100 individuals from a country's population and calculate the mean height of each sample, the distribution of these means is called the sampling distribution. The central limit theorem states that under certain conditions, the sampling distribution of the sample mean will be approximately normal, regardless of the shape of the population distribution.

Measures of Central Tendency:

Measures of central tendency are the measures that are used to describe the distribution of data using a single value. Mean, Median and Mode are the three measures of central tendency.

Mean:

The arithmetic mean is the average of all the data points.

If there are n numbers of observations and xi is the ith observation, then mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Consider the data frame below that has the names of seven employees and their salaries.

	Name	Salary
0	Jane	50000
1	Michael	54000
2	Willian	50000
3	Rosy	189000
4	Hana	55000
5	Ferdie	40000
6	Graeme	59000

To find the mean or the average salary of the employees, you can use the mean() functions in Python.

```
print(df['Salary'].mean())
```

```
71000.0
```

Median:

Median is the middle value that divides the data into two equal parts once it sorts the data in ascending order.

If the total number of data points (n) is odd, the median is the value at position (n+1)/2.

When the total number of observations (n) is even, the median is the average value of observations at n/2 and (n+2)/2 positions.

The median() function in Python can help you find the median value of a column. From the above data frame, you can find the median salary as:

```
print(df['Salary'].median())
```

```
54000.0
```

Mode:

The mode is the observation (value) that occurs most frequently in the data set. There can be over one mode in a dataset.

Given below are the heights of students (in cm) in a class:

155, 157, 160, 159, 162, 160, 161, 165, 160, 158

Mode = 160 cm.

The mode salary from the data frame can be calculated as:

```
print(df['Salary'].mode())
```

```
0    50000
dtype: int64
```

Measures of dispersion

The measures of central tendency are not adequate to describe data. Two data sets can have the same mean but they can be entirely different. Thus to describe data, one needs to know the extent of variability. This is given by the measures of dispersion. Range, interquartile range, and standard deviation are the three commonly used measures of dispersion.

Variance and Standard Deviation

Variance is used to measure the variability in the data from the mean.

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

Consider the below dataset.

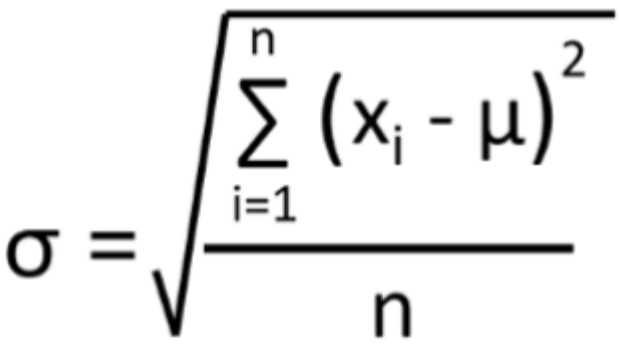
	Name	Salary	Hours	Grade
0	Dan	50000	41	50
1	Joann	54000	40	50
2	Pedro	50000	36	46
3	Rosie	189000	17	95
4	Ethan	55000	35	50
5	Vicky	40000	39	5
6	Frederic	59000	40	57

To calculate the variance of the Grade, use the following:

```
print(df['Grade'].var())
```

```
685.6190476190476
```

Standard deviation in statistics is the square root of the variance. Variance and standard deviation represent the measures of fit, meaning how well the mean represents the data.



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

You can find the standard deviation using the std() function in Python.

```
print(df['Grade'].std())
```

```
26.184328282754315
```

Range and Interquartile Range

Range:

The Range in statistics is the difference between the maximum and the minimum value of the dataset.

```
import numpy as np

data = [10, 12, 13, 16, 16, 20, 25, 31, 34, 41, 43, 45, 50, 54, 60, 61, 68,
        75, 87, 91, 95]

range = max(data) - min(data)

print(range)
```

```
85
```

Interquartile Range (IQR):

The IQR is a measure of the distance between the 1st quartile (Q1) and 3rd quartile (Q3).

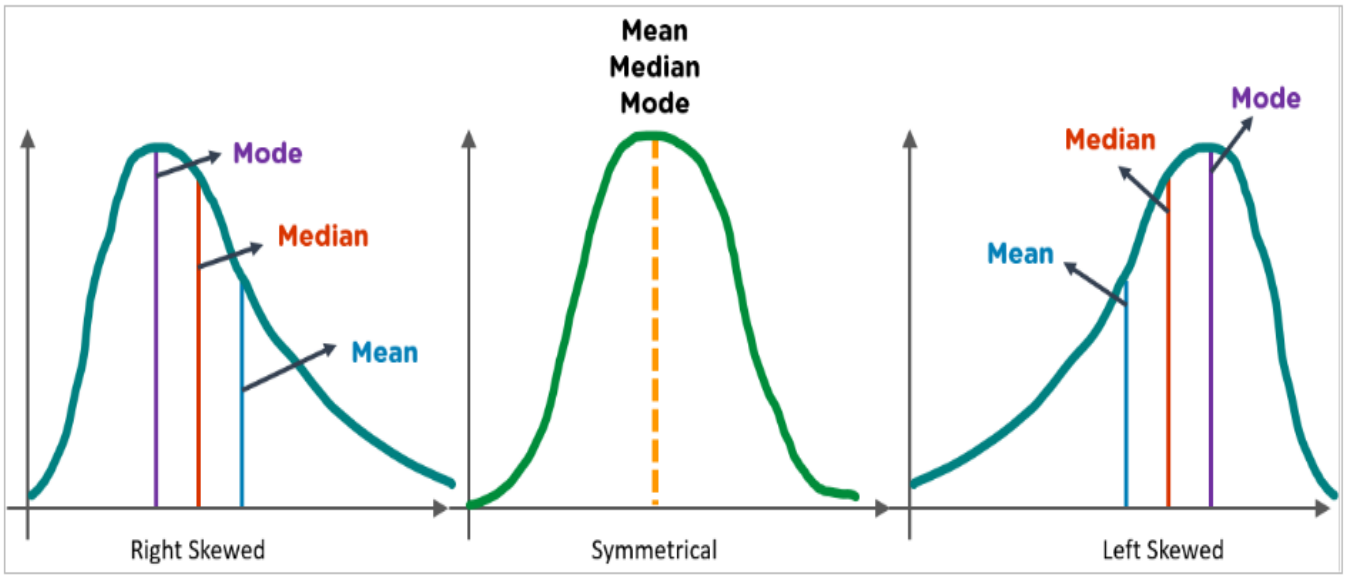
```
import numpy as np

data = [10, 12, 13, 16, 16, 20, 25, 31, 34, 41, 43, 45, 50, 54, 60, 61, 68,
        75, 87, 91, 95]

# First quartile (Q1)
Q1 = np.percentile(data, 25, interpolation = 'midpoint')

# Third quartile (Q3)
Q3 = np.percentile(data, 75, interpolation = 'midpoint')

# Interquaritle range (IQR)
IQR = Q3 - Q1

print(IQR)
```

```
41.0
```

Skewness and Kurtosis

Skewness:

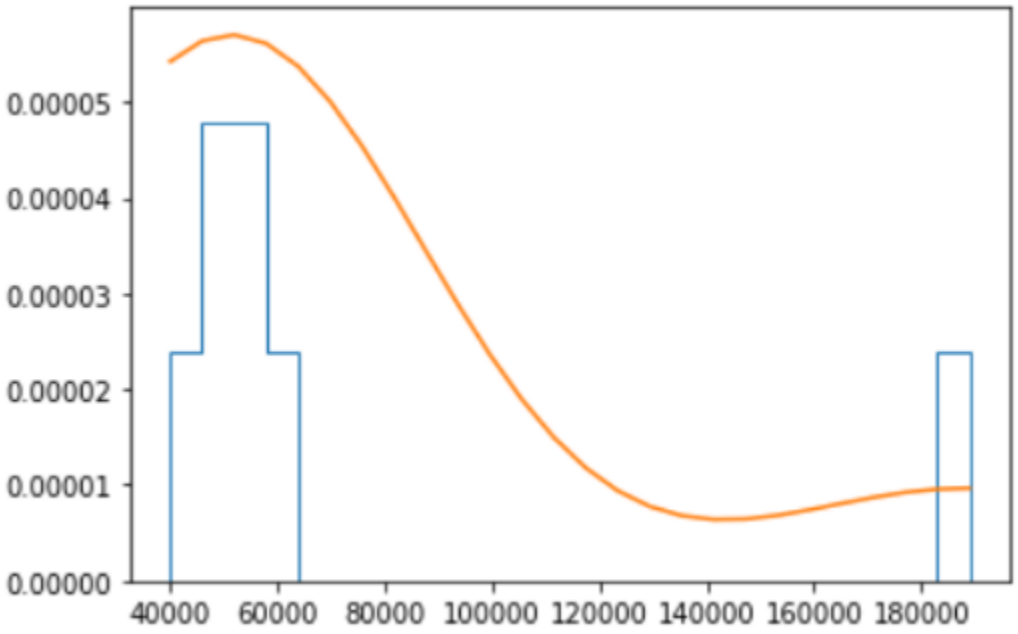
Skewness measures the shape of the distribution. A distribution is symmetrical when the proportion of data at an equal distance from the mean (or median) is equal. If the values extend to the right, it is right-skewed, and if the values extend left, it is left-skewed.



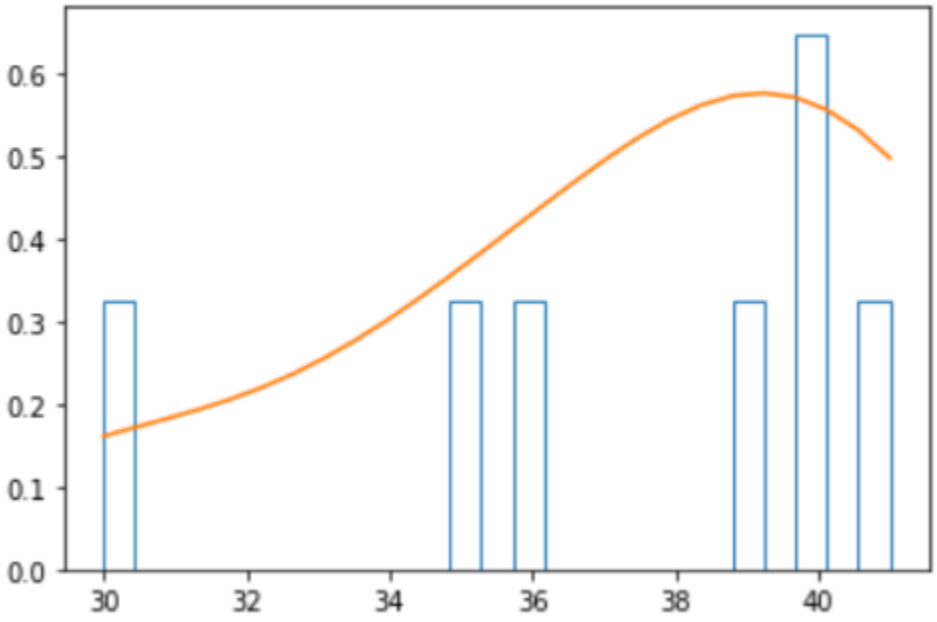
Kurtosis:

Kurtosis in statistics is used to check whether the tails of a given distribution have extreme values. It also represents the shape of a probability distribution.

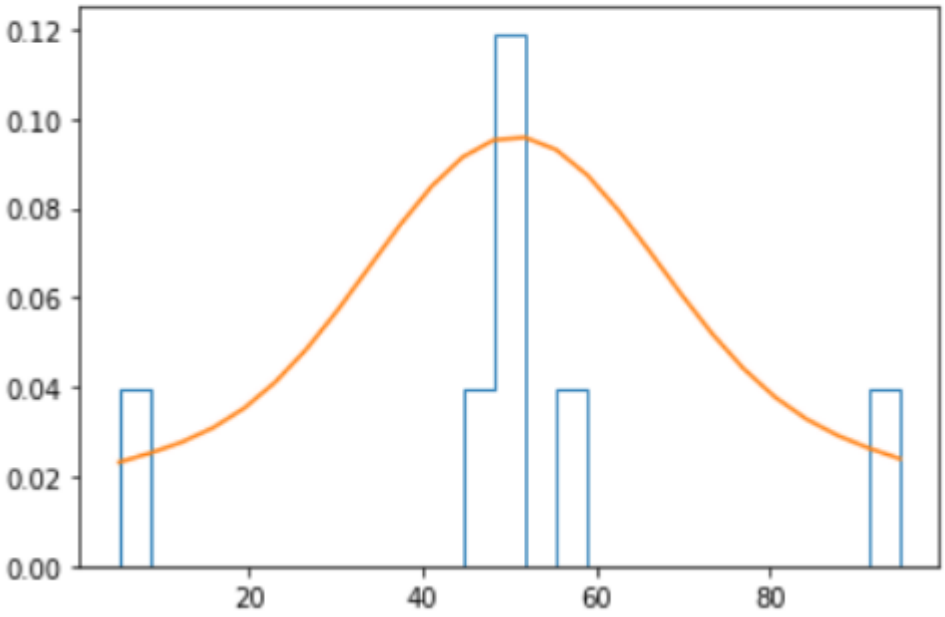
```
# Skewness and Kurtosis

%matplotlib inline
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import scipy.stats as stats

df = pd.DataFrame({'Name': ['Jane', 'Michael', 'Willian', 'Rosy', 'Hana', 'Ferdie', 'Graeme'],
                   'Salary':[50000,54000,50000,189000,55000,40000,59000],
                   'Hours':[41,40,36,30,35,39,40],
                   'Grade':[50,50,46,95,50,5,57]})

numcols = ['Salary', 'Hours', 'Grade']
for col in numcols:
    print(df[col].name + ' skewness: ' + str(df[col].skew()))
    print(df[col].name + ' kurtosis: ' + str(df[col].kurt()))
    density = stats.gaussian_kde(df[col])
    n, x, _ = plt.hist(df[col], histtype='step', normed=True, bins=25)
    plt.plot(x, density(x)*6)
    plt.show()
    print('\n')
```




```
Hours skewness: -1.194570307262883
Hours kurtosis: 0.9412265624999989
```



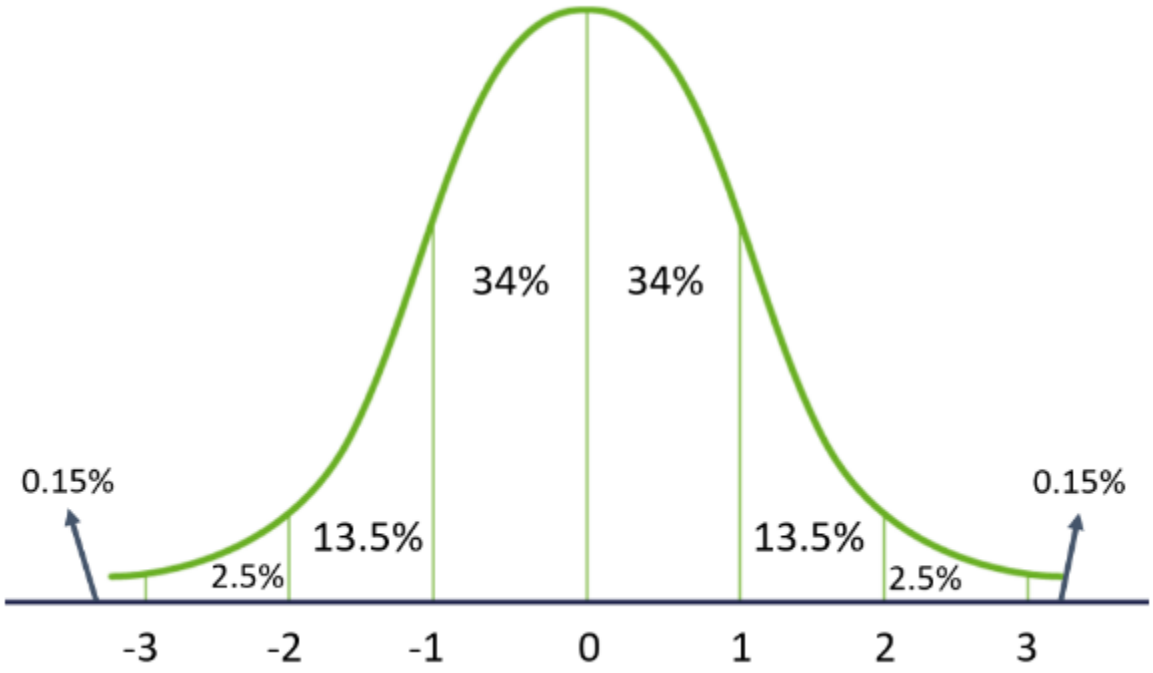
```
Grade skewness: -0.06512433009682762
Grade kurtosis: 2.7484764913773034
```



Now, it’s time to discuss a very popular distribution in statistics for machine learning,

Gaussian Distribution

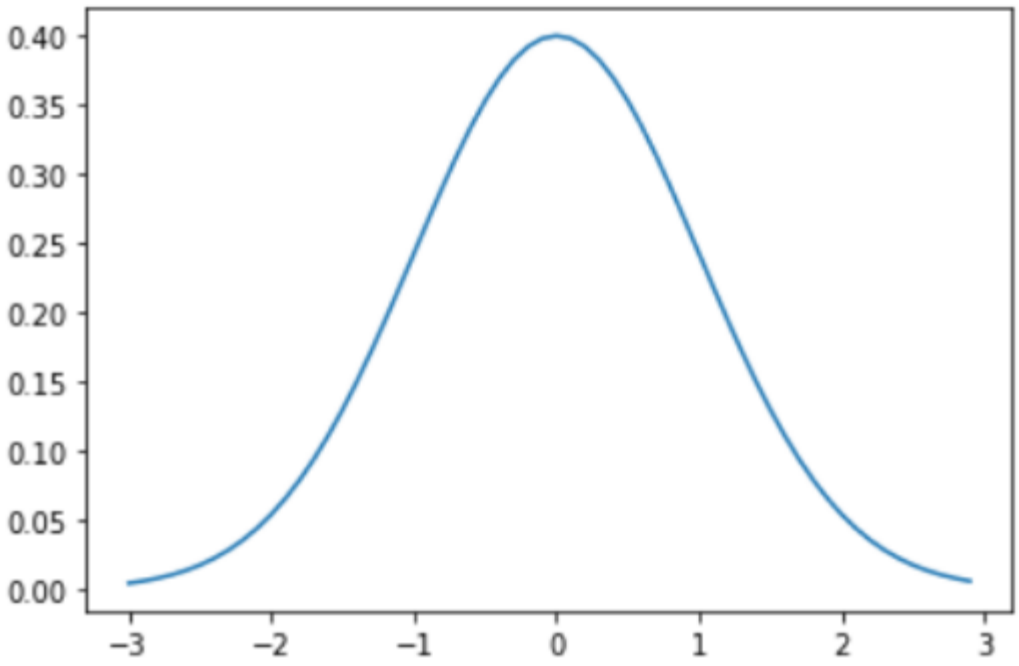
In statistics and probability, Gaussian (normal) distribution is a popular continuous probability distribution for any random variable. It is characterized by 2 parameters (mean μ and standard deviation σ). Many natural phenomena follow a normal distribution, such as the heights of people and IQ scores.



Properties of Gaussian Distribution:

- The mean, median, and mode are the same
- It has a symmetrical bell shape
- 68% data lies within 1 standard deviation of the mean
- 95% data lie within 2 standard deviations of the mean
- 99.7% of the data lie within 3 standard deviations of the mean

```
%matplotlib inline
from numpy import arange
from matplotlib import pyplot
from scipy.stats import norm

# x-axis
x_axis = arange(-3, 3, 0.1)
# y-axis
y_axis = norm.pdf(x_axis, 0, 1)
# plot data
pyplot.plot(x_axis, y_axis)
pyplot.show()
```



Central Limit Theorem

According to the central limit theorem, given a population with mean as μ and standard deviation as σ, if you take large random samples from the population, then the distribution of the sample means will be roughly normally distributed, irrespective of the original population distribution.

Rule of Thumb: For the central limit theorem to hold true, the sample size should be greater than or equal to 30.

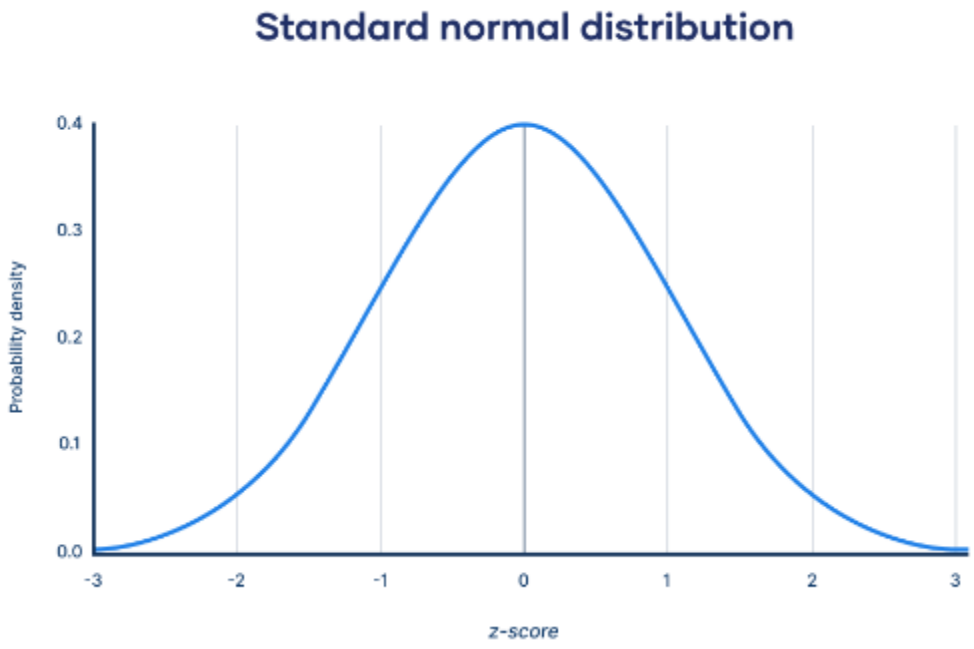
$$x_i \longrightarrow N\left(\mu, \sigma^2/n\right)$$

Now, you will learn a very critical concept in statistics for machine learning, i.e., Hypothesis testing.

The Standard Normal Distribution

The **standard normal distribution**, also called the ***z*-distribution**, is a special normal distribution where the mean is 0 and the standard deviation is 1.

Any normal distribution can be standardized by converting its values into *z* scores. Z scores tell you how many standard deviations from the mean each value lies.



Converting a normal distribution into a *z*-distribution allows you to calculate the probability of certain values occurring and to compare different data sets.

The random variable of a standard normal distribution is known as the **standard score or a z-score**. It is possible to transform every normal random variable X into a z score using the following formula:

$z = (X – \mu) / \sigma$

where X is a normal random variable, μ is the mean of X, and σ is the standard deviation of X. You can also find the normal distribution formula here. In probability theory, the normal or Gaussian distribution is a very common continuous probability distribution.

How to calculate a *z* score

To standardize a value from a normal distribution, convert the individual value into a *z*-score:

1. Subtract the mean from your individual value.
2. Divide the difference by the standard deviation.

***Z*-score formula**	**Explanation**
$Z=(x- \mu)/ \sigma$	• x = individual value • μ = mean • σ = standard deviation

Example: Finding a *z* scoreYou collect SAT scores from students in a new test preparation course. The data follows a normal distribution with a mean score (*M*) of 1150 and a standard deviation (*SD*) of 150. You want to find the probability that SAT scores in your sample exceed 1380.
To standardize your data, you first find the *z* score for 1380. The *z* score tells you how many standard deviations away 1380 is from the mean.

Step 1: Subtract the mean from the *x* value.	x = 1380 M = 1150 $x - M$ = 1380 − 1150 = 230
Step 2: Divide the difference by the standard deviation.	SD = 150 z = 230 ÷ 150 = 1.53

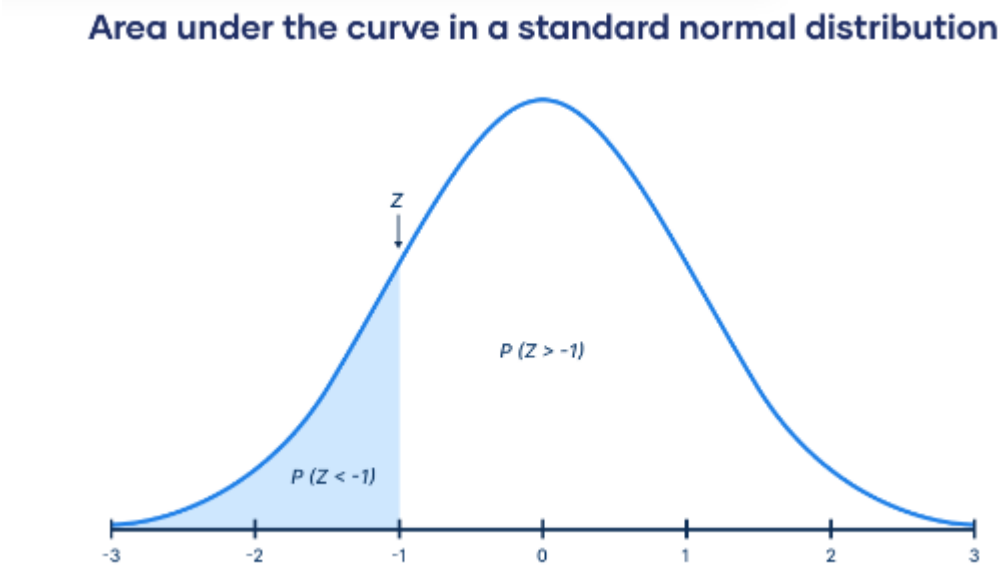
The *z* score for a value of 1380 is **1.53**. That means 1380 is 1.53 standard deviations from the mean of your distribution.

Next, we can find the probability of this score using a *z* table.

Use the standard normal distribution to find probability

The standard normal distribution is a **probability distribution**, so the area under the curve between two points tells you the probability of variables taking on a range of values. The total area under the curve is 1 or 100%.

Every *z* score has an associated *p* value that tells you the probability of all values below or above that *z* score occurring. This is the area under the curve left or right of that *z* score.



P-values

The ***p* value** is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true.

P values are used in hypothesis testing to help decide whether to reject the null hypothesis. The smaller the *p* value, the more likely you are to reject the null hypothesis.

Hypothesis Testing

Hypothesis testing is a statistical analysis to make decisions using experimental data. It allows you to statistically back up some findings you have made in looking at the data. In hypothesis testing, you make a claim and the claim is usually about population parameters such as mean, median, standard deviation, etc.

- The assumption made for a statistical test is called the null hypothesis (H0).

- The Alternative hypothesis (H1) contradicts the null hypothesis stating that the assumptions do not hold true at some level of significance.

Hypothesis testing lets you decide to either reject or retain a null hypothesis.

Example: H0: The average BMI of boys and girls in a class is the same

H1: The average BMI of boys and girls in a class is not the same

To determine whether a finding is statistically significant, you need to interpret the p-value. It is common to compare the **p-value** to a threshold value called the significance level.

It often sets the level of significance to 5% or 0.05.

If the p-value > 0.05 - Accept the null hypothesis.

If the p-value < 0.05 - Reject the null hypothesis.

Some popular hypothesis tests are:

- Chi-square test
- T-test
- Z-test
- Analysis of Variance (ANOVA)

How to use a z table

Once you have a *z* score, you can look up the corresponding probability in a ***z* table**.

In a *z* table, the area under the curve is reported for every *z* value between -3.4 and 3.4 at intervals of 0.01.

There are a few different formats for the *z* table. Here, we use a portion of the cumulative table. This table tells you the total area under the curve up to a given *z* score—this area is equal to the probability of values below that *z* score occurring.

The first column of a *z* table contains the *z* score up to the first decimal place. The top row of the table gives the second decimal place.

To find the corresponding area under the curve (probability) for a z score:

1. Go down to the row with the first two digits of your *z* score.
2. Go across to the column with the same third digit as your *z* score.
3. Find the value at the intersection of the row and column from the previous steps.

Example

Let's walk through an invented research example to better understand how the standard normal distribution works.

As a sleep researcher, you're curious about how sleep habits changed during COVID-19 lockdowns. You collect sleep duration data from a sample during a full lockdown.

Before the lockdown, the population mean was 6.5 hours of sleep. The lockdown sample mean is 7.62.

To assess whether your sample mean significantly differs from the pre-lockdown population mean, you perform a ***z* test**:

1. First, you calculate a *z* score for the sample mean value.
2. Then, you find the *p* value for your *z* score using a *z* table.

Step 1: Calculate a *z*-score

To compare sleep duration during and before the lockdown, you convert your lockdown sample mean into a *z* score using the pre-lockdown population mean and standard deviation.

Formula	Explanation	Calculation
Z=(x- μ)/ σ	*x* = sample mean μ = population mean σ = population standard deviation	$x = 7.62$ $\mu = 6.5$ $\sigma = 0.5$ $z = \frac{7.62 - 6.5}{0.5} = 2.24$

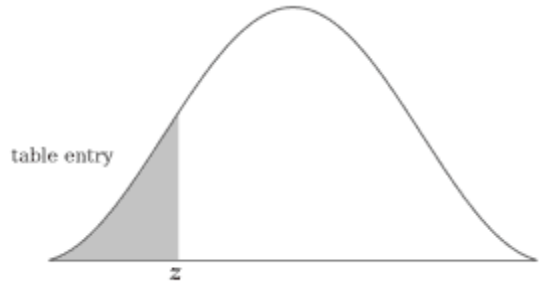
A *z* score of 2.24 means that your sample mean is 2.24 standard deviations greater than the population mean.

Step 2: Find the *p* value

To find the probability of your sample mean *z* score of 2.24 or less occurring, you use the *z* table to find the value at the intersection of row 2.2 and column +0.04.

Z TABLE

Negative Z score table



Use the negative Z score table below to find values on the left of the mean as can be seen in the graph alongside. Corresponding values which are less than the mean are marked with a negative score in the z-table and represent the area under the bell curve to the left of z.

Negative Z Table:

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44034	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08692	.08534	.08379	.08226
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-2	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-3	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-4	.00003	.00003	.00003	.00003	.00003	.00003	.00002	.00002	.00002	.00002

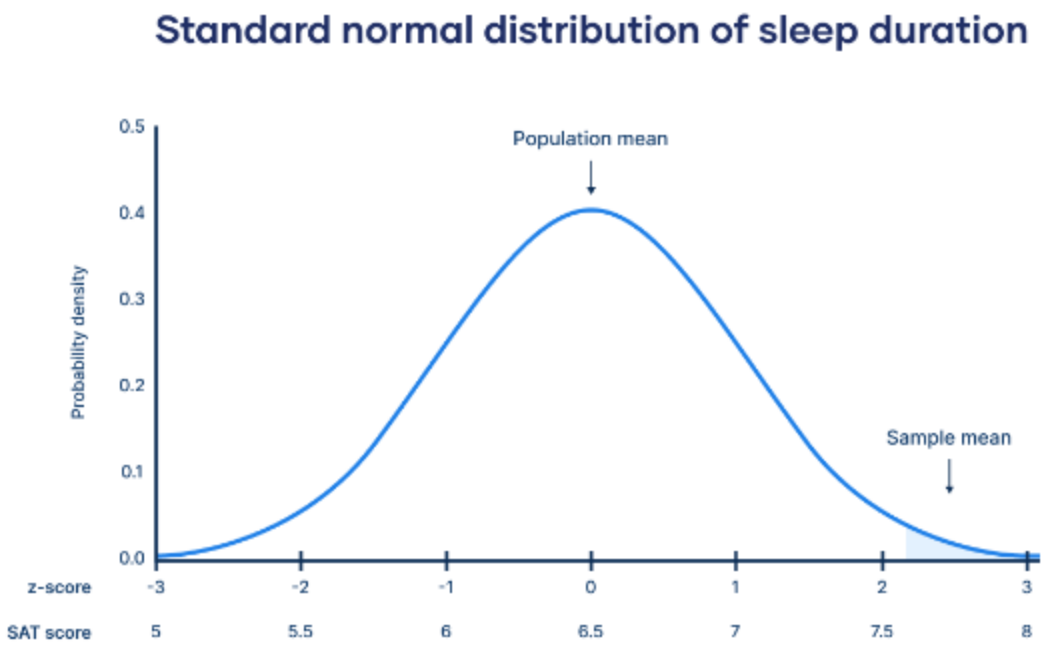
Positive Z score table

Use the positive Z score table below to find values on the right of the mean as can be seen in the graph alongside. Corresponding values which are greater than the mean are marked with a positive score in the z-table and represent the area under the bell curve to the left of z.

Positive Z-table:

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
+0.1	.53983	.54380	.54776	.55172	.55567	.55966	.56360	.56749	.57142	.57535
+0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
+0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
+0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
+0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
+0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
+0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
+0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
+0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
+1	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
+1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
+1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
+1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91308	.91466	.91621	.91774
+1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
+1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
+1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
+1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
+1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
+1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
+2	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
+2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
+2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
+2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
+2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
+2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
+2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
+2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
+2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
+2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
+3	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
+3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
+3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
+3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
+3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
+3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
+3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
+3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
+3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
+3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997
+4	.99997	.99997	.99997	.99997	.99997	.99997	.99998	.99998	.99998	.99998

The table tells you that the area under the curve up to or below your *z* score is 0.9874. This means that your sample's mean sleep duration is higher than about 98.74% of the population's mean sleep duration pre-lockdown.



To find the *p* value to assess whether the sample differs from the population, you calculate the area under the curve above or to the right of your z score. Since the total area under the curve is 1, you subtract the area under the curve below your *z* score from 1.

A *p* value of less than 0.05 or 5% means that the sample significantly differs from the population.

Probability of $z > 2.24 = 1 - 0.9874 =$ **0.0126** or **1.26%**

With a *p* value of less than 0.05, you can conclude that average sleep duration in the COVID-19 lockdown was significantly higher than the pre-lockdown average.