Mastering the Fundamentals of Statistics for Data Science -Basic to Advanced-Part 2

Ritesh Gupta · Follow
Published in Artificial Intelligence in Plain English
13 min read · Jan 31

Listen Share More

Statistics *helps data scientists to understand patterns and trends in data, and make predictions based on that data. It allows data scientists to make inferences about a population based on a sample of data, and to test hypotheses about relationships between variables.*

In this article, we will cover intermediate statistics. If you missed Part 1, you can find it here.

Link:- Mastering the Fundamentals of Statistics for Data Science -Basic to Advance Level- Part 1

1. **Percentiles And Quantiles:**

Percentiles and quantiles are statistical terms that refer to the values that divide a set of observations into certain portions. Both terms describe the same concept, but percentiles are expressed as a percentage of a whole set, while quantiles are expressed as fractions of the same set.

Percentiles divide a set of observations into 100 equal parts, with each part represented as a percentage of the total set. For example, the 50th percentile is value that separates the lowest 50% of the observations from the highest 50%. T 50th percentile is commonly referred to as the median, which is the middle value of a set when the values are arranged in ascending or descending order.

Quantiles, on the other hand, divide a set of observations into n equal parts, where n is any number greater than 1. For example, a set of observations can be divided into 4 equal parts (quartiles), 10 equal parts (deciles), or 100 equal parts (percentiles).

To find percentiles or quantiles, the observations in a set must first be arranged in either ascending or descending order. The formula for finding the kth percentile (where k ranges from 1 to 100) is:

$P_k = (k/100) * (n + 1)$,

where n is the number of observations in the set.

In practice, finding percentiles and quantiles can be useful for data analysis, as it allows for the determination of the distribution of a set of observations. For instance, percentiles and quantiles can help to determine if a set of observations has a normal distribution, or if there are outliers that need to be addressed.

In conclusion, percentiles and quantiles are terms used in statistics to describe the values that divide a set of observations into certain portions. Percentiles are expressed as a percentage of a whole set, while quantiles are expressed as fractions of the same set. Understanding percentiles and quantiles can help in analyzing and interpreting data, making it an important concept in data science.

2. Five Number Summary:

The Five Number Summary is a summary statistic that provides a comprehensive description of the distribution of a set of numerical data. It consists of five values: the minimum value, the first quartile (Q1), the median, the third quartile (Q3), and the maximum value. These values are used to construct a box plot, which provides a visual representation of the distribution of the data.

To better understand the Five Number Summary, let's take a look at an example.

Suppose we have a set of observations, [3, 4, 5, 7, 8, 9, 10]. To find the Five Number Summary, we first need to arrange the data in ascending order, [3, 4, 5, 7, 8, 9, 1

The minimum value is the smallest observation in the set, which is 3.

To find Q1, we need to find the median of the lower half of the data, which is [3, 4, 5]. The median of this subgroup is 4.

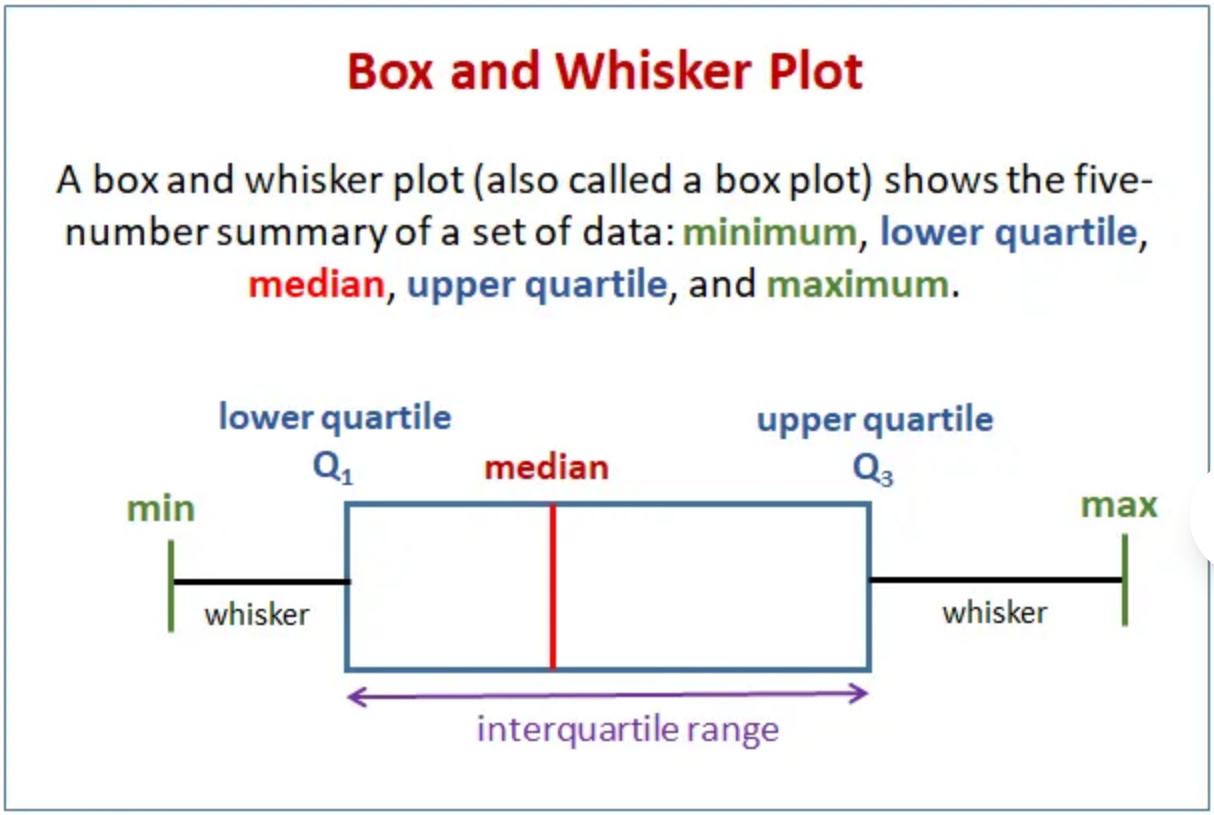
The median of the entire set, [3, 4, 5, 7, 8, 9, 10], is 7, as it separates the lowest 50% of the observations from the highest 50%.

To find Q3, we need to find the median of the upper half of the data, which is [7, 8, 9, 10]. The median of this subgroup is 8.5.

The maximum value is the largest observation in the set, which is 10.

So, the Five Number Summary of this set of observations is: [3, 4, 7, 8.5, 10].

This information can be used to create a box plot, which provides a visual representation of the distribution of the data. The box plot would show the median as a line within the box, Q1 and Q3 as the limits of the box, and the minimum and maximum values as whiskers outside the box. Outliers, if any, would be plotted as individual points outside the whiskers.

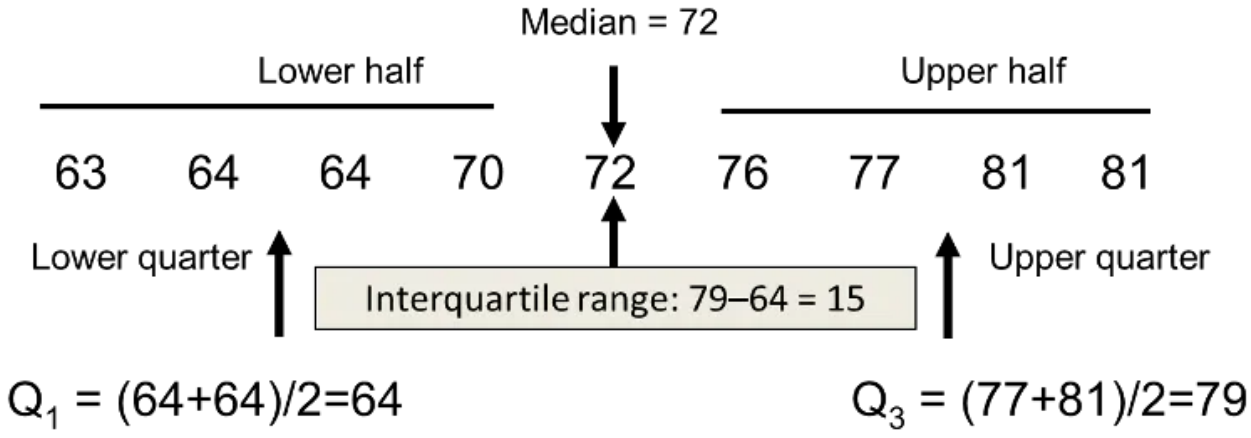


credit: onlinemathlearning

In conclusion, the Five Number Summary provides a comprehensive description of the distribution of a set of numerical data. It is a useful tool for summarizing and analyzing data, and can help in identifying patterns and outliers. Understanding the Five Number Summary can help in making informed decisions based on the data being analyzed.

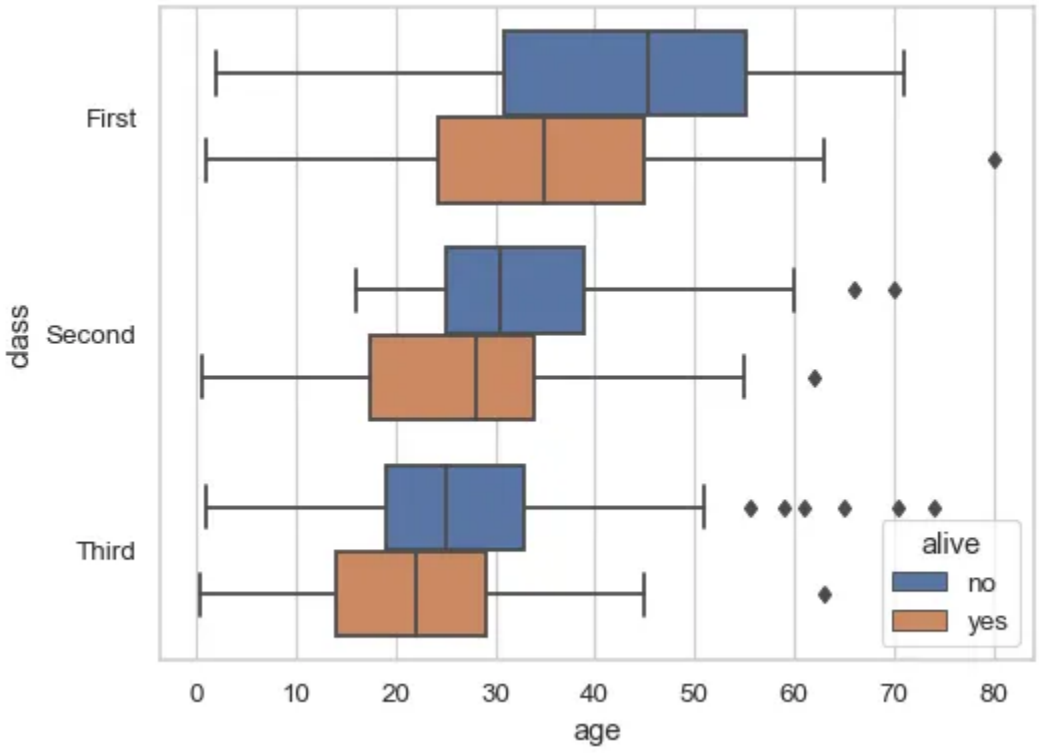
3. Inter Quartile Range(IQR)

The interquartile range (IQR) is a measure of the variability of a set of numerical data based on dividing a data set into quartiles. It is calculated as the difference between the third quartile (75th percentile) and the first quartile (25th percentile). It is used as a robust measure of dispersion as it is not affected by outliers in the data.



4. Boxplots

A boxplot, also known as a box-and-whisker plot, is a graphical representation of a set of numerical data based on its quartiles. It displays the minimum, first quartile, median, third quartile, and maximum values of a dataset as a box, with “whiskers” extending from the box to indicate the variability outside the upper and lower quartiles. Outliers can also be plotted individually as points outside the whiskers. Boxplots are useful for visualizing the distribution and skewness of a dataset and for comparing multiple sets of data.



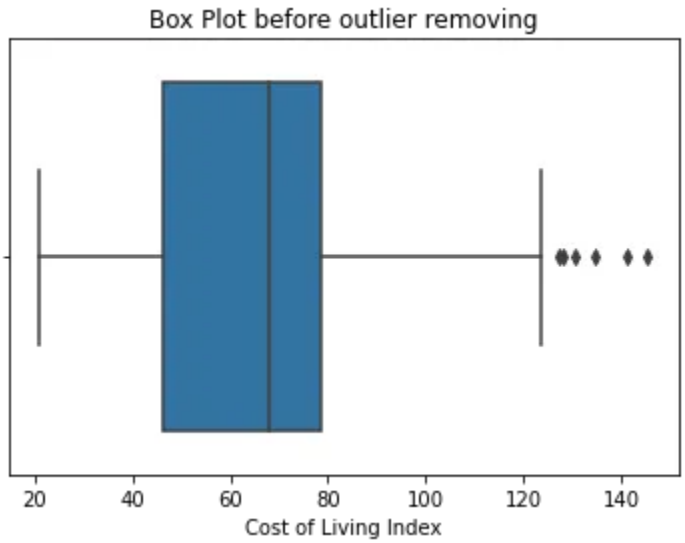
5. Effect Of Outliers And It’s Removal

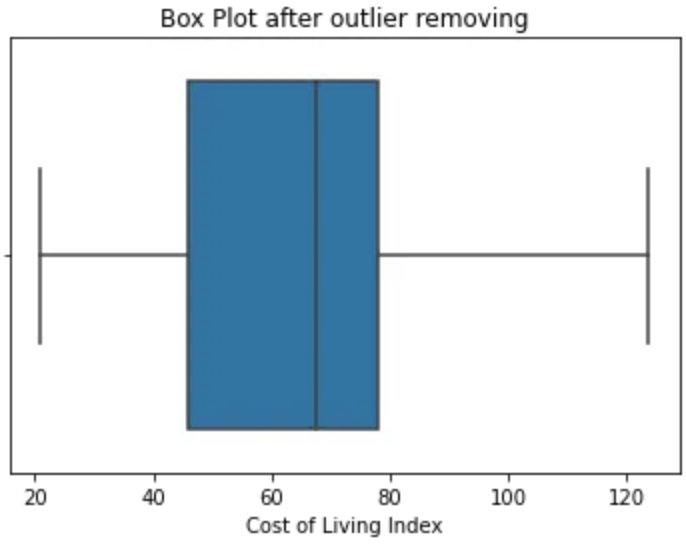
Outliers can significantly impact the results of statistical analyses. They can skew the mean, increase variance, and affect the distribution of data. Outlier removal can improve the accuracy of statistical models, but it should be done carefully as it can also remove important information. The decision to remove outliers should be based on a thorough understanding of the data and the reasons for the outliers. Some common methods for outlier removal include the use of statistical tests and setting upper and lower bounds based on the distribution of the data.

For example, consider a dataset of exam scores for 100 students. If there is one student who scored significantly higher than the rest of the students, this stude score could be considered an outlier. If this outlier is not removed, it could greatly impact the mean of the data, making it seem as though the average exam score is higher than it actually is.

When removing outliers, it's important to use appropriate methods that take into account the specific characteristics of the data. One common method is the Z-score method, which calculates the number of standard deviations a data point is from the mean. Data points that are more than a certain number of standard deviations away from the mean can be considered outliers and removed.

In this example, if the Z-score method is used and a threshold of 3 standard deviations is set, the student with the outlier score could be removed from the dataset. This would reduce the impact of the outlier on the mean exam score and provide a more accurate representation of the data. However, it's important to consider whether this student's score is truly an outlier or if it could represent an important piece of information that should be kept in the analysis.

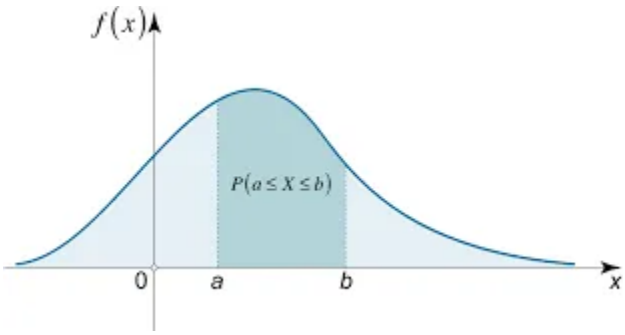




Another method for outlier removal is the IQR (Interquartile Range) method, which is based on the distribution of the data. The IQR is calculated as the difference between the 75th and 25th percentile of the data. Outliers are considered any data points that are outside of the range of 1.5 times the IQR from the 25th or 75th percentile.

6. Probability Density Function

Probability density function (pdf) is a mathematical function used in probability theory and statistics to describe the likelihood of a random variable taking on a particular value. It gives the probability per unit value for a continuous random variable, and the area under the curve represents the total probability of the variable falling within a certain range. For a discrete random variable, it gives the probability of the variable taking on a particular value.



credit: math24

Example 1: If the probability density function is given as:

$$f(x) = \begin{cases} x(x-1) & 0 \le x < 3 \\ x & x \ge 3 \end{cases}$$

Find $P(1< X < 2)$.

Solution: Integrating the function,

$$\int_1^2 x(x-1)dx = \int_1^2 (x^2 - x)dx$$

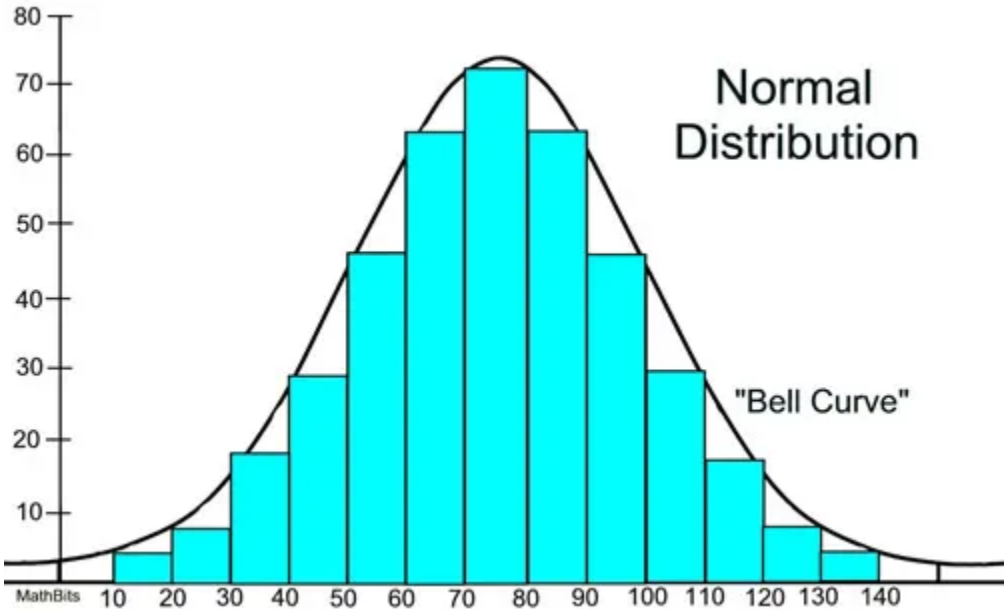
$$= \left[\frac{x^3}{3} - \frac{x^2}{2}\right]_1^2$$

$$= 5 / 6$$

Answer: $P(1< X < 2) = 5 / 6$

7. **Normal Distribution or Gaussian Distribution And Emperical Formula**

Normal distribution, also known as Gaussian distribution, is a widely used probability distribution in statistics and is named after the mathematician Carl Friedrich Gauss. It is a continuous probability distribution that describes the distribution of a random variable with a symmetrical bell-shaped curve centered around its mean value.



credit: mathbitsnotebook

The normal distribution is defined by two parameters, the mean (μ) and the standard deviation (σ). The mean represents the center of the distribution, and the

standard deviation describes the spread of the data. A normal distribution with a small standard deviation will have a tighter clustering of values around the mean, while a normal distribution with a larger standard deviation will have a wider spread of values.

In a normal distribution, approximately 68% of the data falls within one standard deviation of the mean, 95% of the data falls within two standard deviations, and 99.7% falls within three standard deviations. This property of normal distribution makes it a useful tool for modeling the distribution of various real-world phenomena, such as height, weight, intelligence, etc.

Empirical formula is a formula that is derived from the observation of data and not from any theoretical principle. It provides a mathematical representation of the relationship between variables in a given dataset. In other words, it represents the pattern that can be observed in the data, without making any assumptions about the underlying cause of the pattern.

One of the simplest and widely used empirical formulas is the linear regression formula, which can be used to model the relationship between two variables. The linear regression formula is derived by finding the line of best fit that minimizes the sum of the squared differences between the observed values and the values predicted by the formula. The resulting formula can then be used to make predictions about the values of the dependent variable based on the values of the independent variable.

8. Z-score

A Z-score is a numerical measurement that describes the number of standard deviations a value is away from the mean of a set of data. It is used to determine how many standard deviations a particular value is from the mean of the data. The Z-score is calculated as the difference between the value of interest and the mean of the data, divided by the standard deviation of the data.

Here's an example to illustrate the concept of Z-score:

Suppose we have a set of exam scores for a class of 20 students, with a mean score of 75 and a standard deviation of 5. Let's say we want to find the Z-score for a student who scored 85 on the exam.

The Z-score can be calculated as follows:

Z = (85–75) / 5 = 2

This means that the student who scored 85 on the exam is 2 standard deviations above the mean. In other words, their score is 2 standard deviations away from the average score of the class.

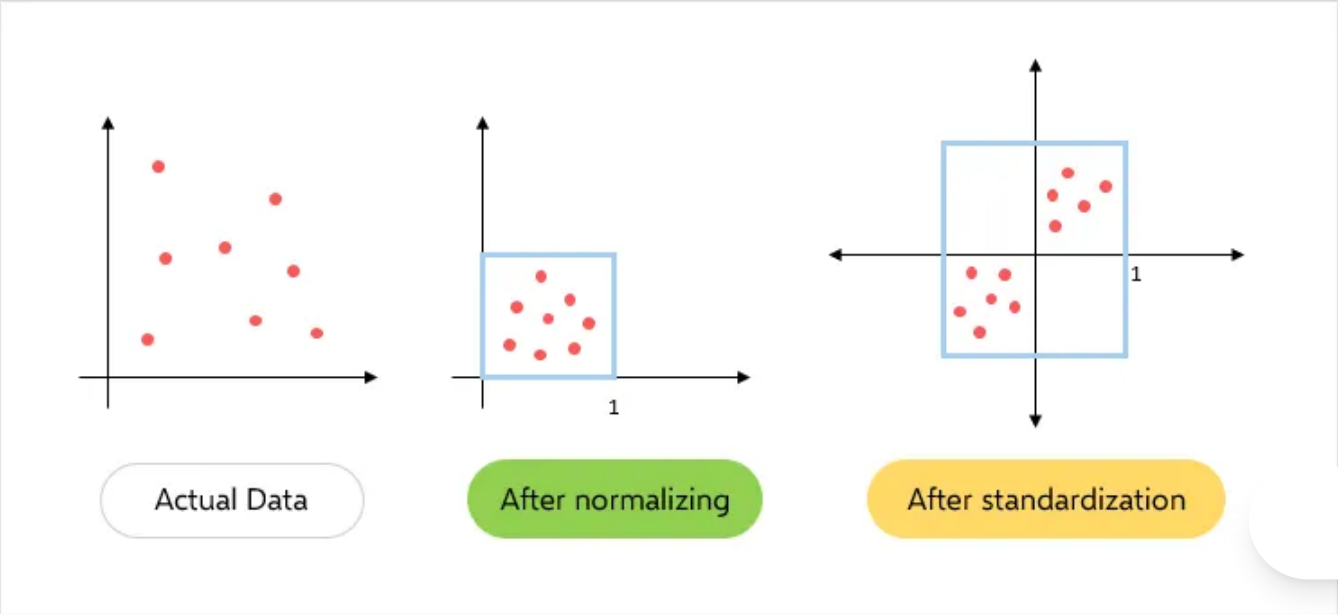
The Z-score provides a standardized way of comparing values in a set of data, regardless of the scale of the data. This makes it a useful tool for comparing valu across different data sets and for determining the statistical significance of valu a data set.

9. Standardization Vs Normalization

Standardization and normalization are two common techniques used in preprocessing of data in machine learning and statistics. They are used to transform variables to a common scale, so that they can be compared and analyzed.

Standardization involves transforming a variable to have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean of the variable and dividing the result by its standard deviation. This makes it easy to compare variables with different means and standard deviations, as they are all transformed to the same scale.

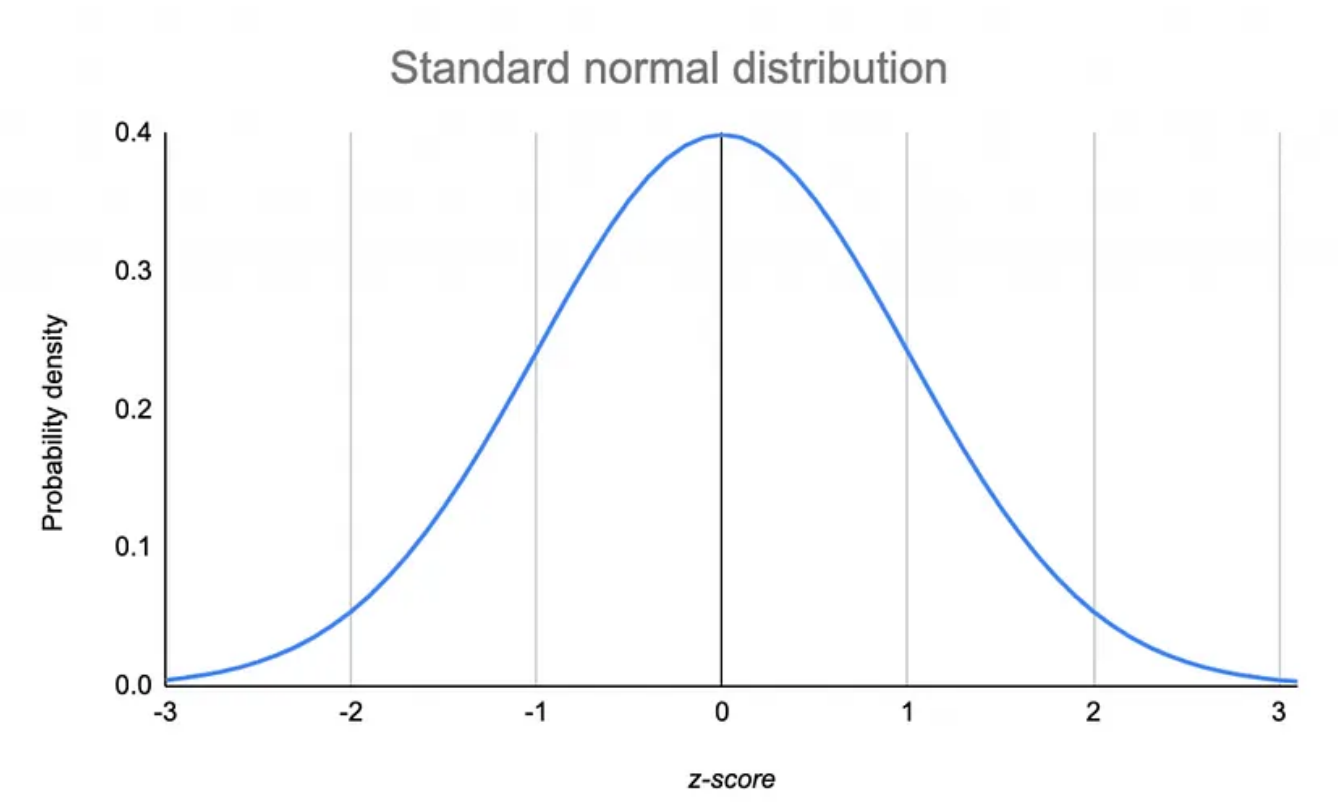
Normalization, on the other hand, involves transforming a variable to have a specific range, typically between 0 and 1. This is done by subtracting the minimum value of the variable and dividing the result by the range of the variable. Normalization is often used for variables that have a skewed distribution or a large range of values, as it transforms them to a common scale that makes it easier to compare the values.



credit: someka

10. Standard Normal Distribution

The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1. It is also known as the Z-distribution and is widely used in statistical analysis and hypothesis testing. The standard normal distribution is important because any normal distribution can be converted to a standard normal distribution through the use of standardization.



credit: scribbr

11. Central Limit Theorem

The Central Limit Theorem states that the distribution of the sum (or average) of a large number of independent and identically distributed random variables approaches a normal distribution, regardless of the shape of the original distribution. This means that as the sample size increases, the distribution of the sample mean becomes increasingly close to a normal distribution with a mean equal to the population mean and a standard deviation equal to the population standard deviation divided by the square root of the sample size. The Central Limit Theorem is a fundamental result in probability theory and has important implications for statistical inference.

For example, let's say we have a population of heights of all individuals in a certain country. The distribution of heights is not necessarily a normal distribution, but might have a skewed or irregular shape. However, if we take a sample of n individuals from this population and calculate the mean height of this sample, the distribution of the sample means will approach a normal distribution as n increases.

Suppose we take a sample of n=30 individuals and repeat this process 100 times, each time calculating the mean height of the sample. The central limit theorem tells us that the distribution of these 100 sample means will approximate a normal distribution with a mean equal to the population mean height and a standard deviation equal to the population standard deviation divided by the square root of the sample size (n=30). This is an important result because it allows us to make inferences about the population mean height based on the sample mean height, and to quantify the uncertainty of these inferences through the use of confidence intervals and hypothesis tests.

12. Chebyshevs Inequality

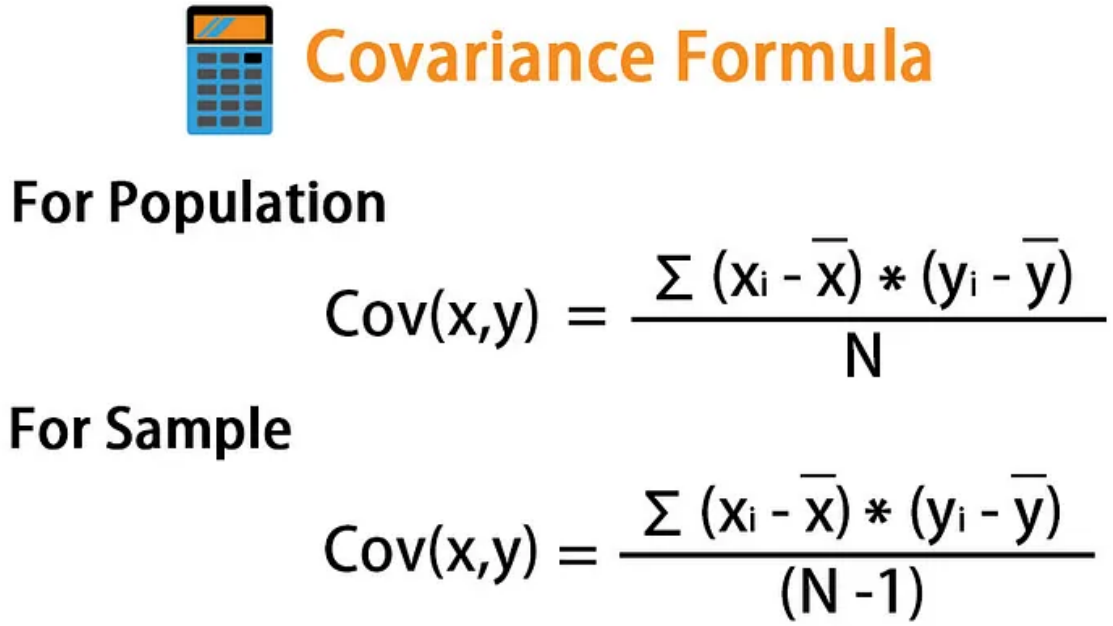
Chebyshev's Inequality states that for any random variable and for any positive number k, the proportion of values that are more than k standard deviations away from the mean is at most $1/k^2$. In other words, the inequality bounds the amount of the population that lies outside k standard deviations from the mean, regardless of the shape of the underlying distribution.

The inequality states that: $P(|X-\mu| \geq k\sigma) \leq 1/k^2$

where X is the random variable, μ is the mean, σ is the standard deviation, and k is a positive number. This inequality provides a rough estimate of how much of the population lies within a certain range of the mean and can be useful in practice for characterizing the spread of a distribution.

13. Covariance

Covariance is a measure of the linear association between two random variables. It reflects how two variables change with respect to each other. A positive covariance indicates that the variables are positively related (i.e., they tend to increase or decrease together), while a negative covariance indicates that the variables are inversely related (i.e., one variable increases while the other decreases).



credi: educba

Covariance is a useful measure for understanding the linear relationship between two variables, but it has some limitations. For example, covariance does not distinguish between the strength or direction of the relationship, and it is not standardized, making it difficult to compare covariance values between variables with different units or scales. The Pearson correlation coefficient is often used to overcome these limitations and to provide a standardized measure of the linear association between two variables.

14. Pearson Correlation Coefficient

The Pearson correlation coefficient (also known as Pearson's r) is a measure of the linear association between two continuous variables. It ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship. The Pearson correlation coefficient is calculated as the covariance between two variables divided by the product of their standard deviations.

The formula for the Pearson correlation coefficient is given by:

$$r = Cov(X, Y) / (\sigma X * \sigma Y)$$

where X and Y are the two variables, Cov(X, Y) is the covariance between X and Y, and σX and σY are the standard deviations of X and Y, respectively.

The Pearson correlation coefficient provides a standardized measure of the linear association between two variables and is widely used in fields such as psychology, sociology, and finance to quantify the strength and direction of the relationship between variables. It is important to note that while the Pearson correlation coefficient measures linear association, it may not capture non-linear or more complex relationships between variables.

To be Continued.....!

> **Follow me** *for Part 3 of Statistics for Data Science -Basic to Advance*
>
> *Link:- Mastering the Fundamentals of Statistics for Data Science -Basic to Advance Level- Part 1*

Thanks for Reading!

If you enjoyed this, follow me to never miss another article on data science guides, tricks and tips, life lessons, and more!

More content at ***PlainEnglish.io****. Sign up for our* ***free weekly newsletter****. Follow us on* ***Twitter****,* ***LinkedIn****,* ***YouTube****, and* ***Discord****.*

Interested in scaling your software startup*? Check out* ***Circuit****.*

Statistics | Data Science | Mathematics | Machine Learning | Artificial Intelligence