

Statistics for Data Science

Statistics for Data Science is an important field that is used to extract meaningful insights from vast amounts of data. It involves the use of mathematical techniques and tools to analyze, interpret, and draw conclusions from data.

1. Different Types of Statistical Inference

Statistical inference is a very important tool that allows data scientists to make informed decisions and solve complex problems. It is some set of techniques used to accurately make important decisions about datasets based on samples of it. This is achieved by studying the sample and drawing conclusions about the whole population.

Statistical Inference has several different types in data science, including:

- **Hypothesis Testing**

Hypothesis testing involves using data to make decisions. We need to characterize the conclusions we can draw. Classical hypothesis testing focuses on deciding between two options: the null hypothesis (H_0) represents the status quo and is assumed true by default, while the alternative or research hypothesis (H_a or H_1) requires evidence to be concluded. So, we need statistical evidence to reject the null hypothesis in favor of the research hypothesis.

The null hypothesis is usually an equality statement about population parameters, such as the population mean being equal to zero. The alternative hypothesis is the opposite of the null hypothesis, indicating that the population mean is not equal to zero. These two hypotheses are mutually exclusive, and only one can be true. However, one of them will always be correct.

There can only be four possible outcomes here, which are:

Truth	Decision	Result
H_0	H_0	Correct, accept H_0
H_0	H_a	Type I error
H_a	H_a	Correct, reject H_0
H_a	H_0	Type II error

There are two more important factors to consider, which are:

- **The level of significance α** is the level of confidence needed to accept or reject a hypothesis. Usually, a level of significance of 5% is used, which means that the output should be 95% confident to be considered reliable.
- **The p-value**, or calculated probability, is the likelihood of obtaining the observed results if the null hypothesis is true. If the p-value is less than the chosen level of significance, then the null hypothesis is rejected and the alternative hypothesis is accepted.

Example:

Given a coin and it is unknown whether that is a real coin, or a gimmick coin so let's decide null and alternate hypothesis.

Null Hypothesis(H_0): The coin is a real coin.

Alternative Hypothesis(H_1): The coin is a gimmick coin.

$\alpha = 0.05$

Now let's toss the coin and calculate the p-value.

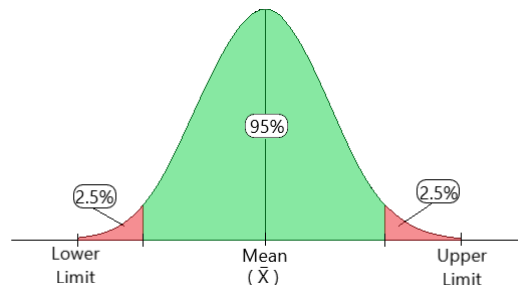
1st toss assuming that the result is head, **p-value = 50%**

2nd toss assuming that the result is also head, **p-value = 25%**

Similarly till the 6th Toss, now **p-value = 1.5%** which is less than $\alpha = 5\%$, so the null hypothesis is rejected. Suggesting that the coin is in fact a gimmick coin, which makes sense given that it landed heads 6 times in a row.

• Confidence Intervals

A confidence interval is a range of values in statistics that helps estimate an unknown parameter based on observed data. It's like a "best guess" with a margin of error. The confidence level, usually 95% or 99%, tells us how confident we are in our estimate. It's a way to show how uncertain or certain we are about the parameter being estimated.



Confidence level is like a measure of how sure we are about the results we get from sampling. It tells us how likely it is that our sample captures the true population value we're interested in. For example, if we set a 95% confidence level, it means that 95 out of 100 times our sample results would include the true population value.

Imagine you're surveying the average height of men in a city, and you set a 95% confidence level. Your calculated confidence interval might be (168,182) cm. This means that if you did the survey many times, in 95 out of 100 cases, the true average height of men in the city would fall between 168 cm and 182 cm.

• Regression Analysis

Regression analysis is a way to figure out how different things are related to each other using math. It helps us understand how one thing changes when another thing changes. For example, a gym supplement company might use it to see how prices and ads affect how much of their supplements people buy. It's like connecting the dots to see patterns and make predictions.

Regression analysis helps organizations to understand what their data points mean and to use them carefully with business analysis techniques to arrive at better decisions. It showcases how dependent variables vary when one of the independent variables is varied and the other independent variables remain unchanged. It acts as a tool to help business analysts and data experts pick significant variables and delete unwanted ones. There are several types of regression analysis such as:

Regression Type	Description	Advantages	Disadvantages	Use Cases	Graph
Linear Regression	Models the linear relationship between a dependent variable and one or more independent variables	<ul style="list-style-type: none">Simple and interpretableFast to computeDoes not require complex parameter tuning	<ul style="list-style-type: none">Assumes linear relationship between variables.Sensitive to outliersMay not perform well with non-linear data	<ul style="list-style-type: none">Predicting numeric valuesIdentifying significant predictorsAssessing strength and direction of relationship	
Logistic Regression	Models the probability of an event occurring based on predictor variables	<ul style="list-style-type: none">Interpretable and provides probability estimates.Handles binary or ordinal outcomes.Can handle multiple predictors	<ul style="list-style-type: none">Assumes linearity of logitMay not perform well with non-linear relationships.Assumes independence of errorsMay suffer from multicollinearity	Predicting binary outcomes, such as yes/no or 0/1 events	
Polynomial Regression	Models a curvilinear relationship between a dependent variable and one or more independent variables	<ul style="list-style-type: none">Can capture non-linear patterns.Can fit complex data with curved relationships	<ul style="list-style-type: none">May overfit the dataMay have issues with extrapolation beyond the observed range	<ul style="list-style-type: none">Capturing non-linear patternsFitting complex data with curved relationships	

2. Different Types of Probability Distributions

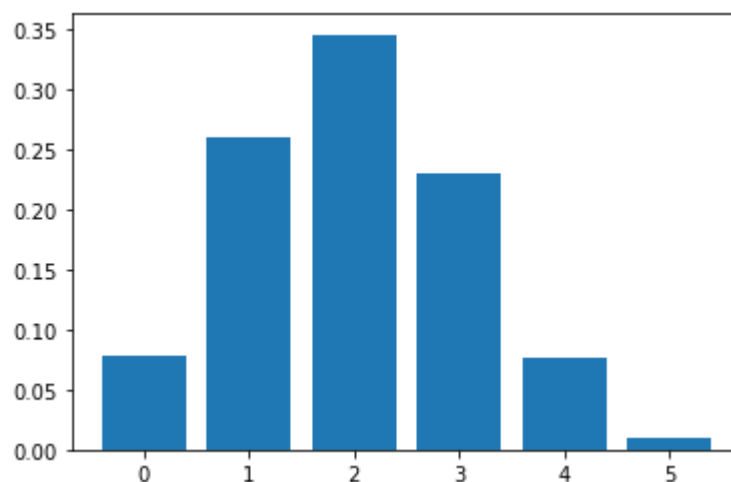
A probability distribution is a statistical tool used to describe all the possible values and probabilities of a random variable within a given range. The range of the distribution is determined by the minimum and maximum possible values, while the placement of possible values on the distribution is determined by factors such as the mean (average), standard deviation, skewness, and kurtosis.

Types of Probability Distribution:

- Binomial Distribution

The binomial distribution is a discrete distribution with a finite number of possibilities. When observing a series of what are known as Bernoulli trials, the binomial distribution emerges. A Bernoulli trial is a scientific experiment with only two outcomes: success or failure.

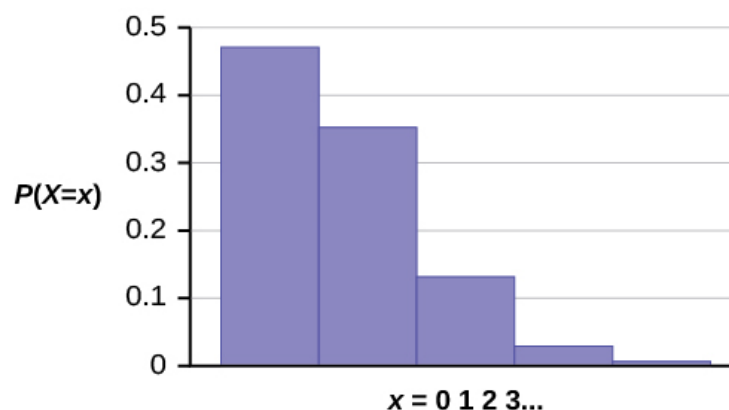
Consider a random experiment in which you toss a biased coin six times with a 0.4 chance of getting head. If 'getting a head' is considered a 'success', the binomial distribution will show the probability of r successes for each value of r . The binomial random variable represents the number of successes (r) in n consecutive independent Bernoulli trials.



The binomial distribution is used in modeling of binary outcomes, such as success/failure, yes/no, etc. and in hypothesis testing for proportions.

- Poisson Distribution

A Poisson distribution is a probability distribution commonly employed in statistics to model the occurrence of events within a specific timeframe. In other words, it is a type of count distribution. Poisson distributions are often used to analyze the frequency of independent events that occur at a consistent rate over a designated time period.

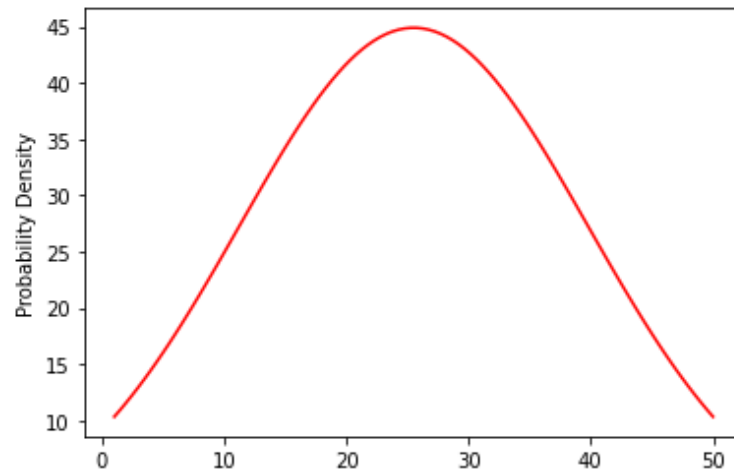


The Poisson distribution is used in modeling of rare events, such as the number of customer arrivals, number of defects, etc. and in survival analysis and epidemiological studies.

- Normal Distribution

The Normal Distribution, which is also named as the Gaussian Distribution. It is a fundamental type of continuous probability distribution. It presents symmetry around its mean value, indicating that data close to the mean occurs more often than data that is further away from it. This statistical distribution is widely used in various fields due to its versatility and ubiquity in modeling real-world phenomena.

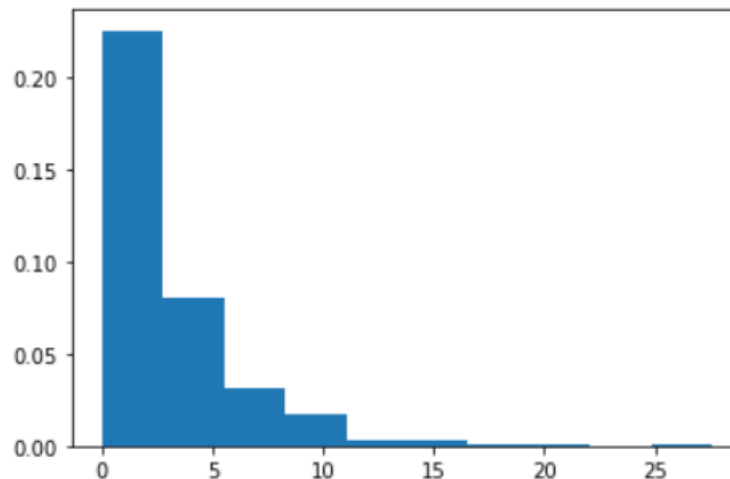
In the example, you generated 100 random variables ranging from 1 to 50. After that, you created a function to define the normal distribution formula to calculate the probability density function. Then, you have plotted the data points and probability density function against X-axis and Y-axis, respectively.



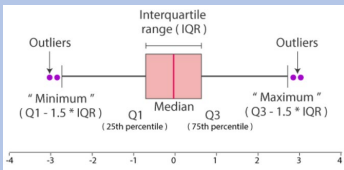
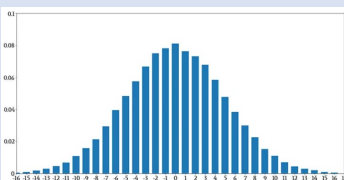
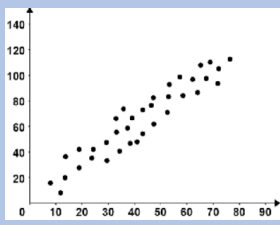
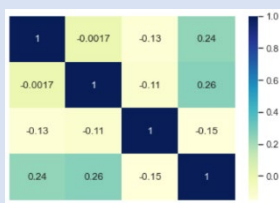
The binomial distribution is used in modeling of continuous variables such as heights, weights, test scores, etc. and in statistical process control and regression analysis.

- Exponential Distribution

In a Poisson process, an exponential distribution is a continuous probability distribution that describes the time between events (success, failure, arrival, etc.).



3. Different Methods for Data Exploration and Visualization

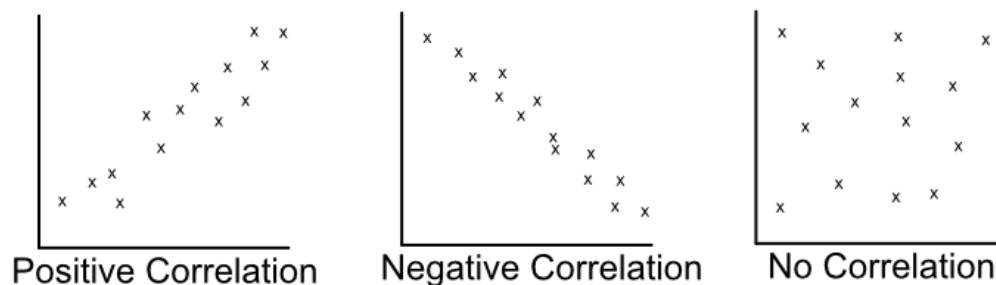
Visualization Method	Description	Advantages	Disadvantages	Use Cases	Graphs
Box Plot	Displays the distribution of data using a box with whiskers	<ul style="list-style-type: none"> Provides a visual summary of the median, quartiles, and outliers of the data. Can easily identify potential outliers and skewness in data 	<ul style="list-style-type: none"> Limited in showing detailed information of the data distribution. May not be suitable for displaying data with multiple modes or complex distributions. 	<ul style="list-style-type: none"> Comparing distributions of multiple variables. Detecting outliers Identifying skewness in data 	
Histogram	Displays the distribution of data using bins or bars	<ul style="list-style-type: none"> Provides a visual representation of data distribution. Shows the frequency or count of data in each bin 	<ul style="list-style-type: none"> May lose information due to binning. Limited in showing detailed information about central tendency and variability 	<ul style="list-style-type: none"> Analyzing the shape of data distribution Identifying data skewness and modes Checking for data outliers 	
Scatter Plot	Displays the relationship between two variables using points on a Cartesian plane	<ul style="list-style-type: none"> Provides visual representation of how two variables interact. Identify patterns, trends, and outliers in data. 	<ul style="list-style-type: none"> Limited to visualizing only two variables. May not be suitable for displaying data with multiple variables 	<ul style="list-style-type: none"> Exploring relationships between two variables Identifying patterns and trends Detecting outliers 	
Heatmap	Displays data in a tabular format using colors to represent values	<ul style="list-style-type: none"> Provides a visual representation of data in a tabular format. Can identify patterns and trends in data. 	<ul style="list-style-type: none"> Limited in displaying detailed information about individual data points. May not be suitable for displaying large datasets 	<ul style="list-style-type: none"> Identifying patterns or trends in tabular data Analyzing relationships between variables 	

4. Correlation and Causation

Correlation and causation are two related but distinct concepts in data science. Correlation is a measure of the strength and direction of the relationship between two variables, while causation is the process of establishing a cause-and-effect relationship between two variables. Let's dive deeper:

- Correlation

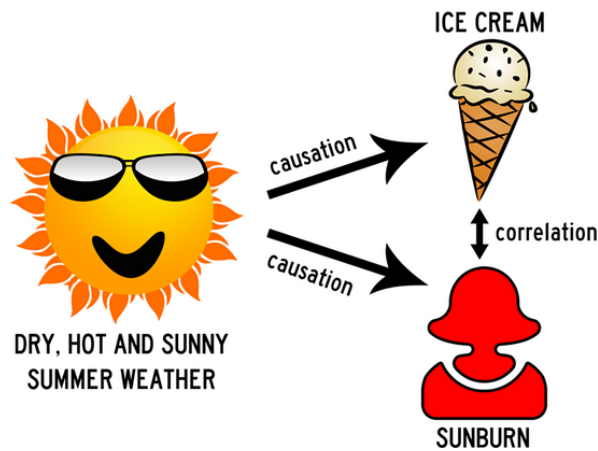
Correlation refers to the connection and interdependence between variables, where changes in one variable are associated with changes in another variable. For instance, as the temperature rises during hot weather, there is typically an increase in ice-cream sales.



A positive correlation implies that the variables exhibit movement in the same direction (as depicted in the left plot), while a negative correlation suggests that the variables move in opposite directions (as shown in the middle plot). The rightmost plot illustrates a lack of correlation between the variables.

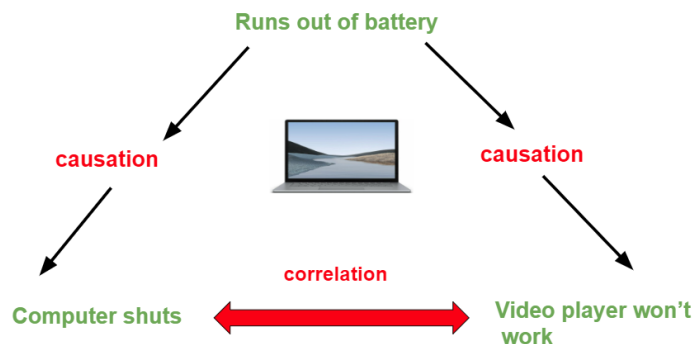
- Causation

Causation refers to the phenomenon where changes in one variable lead to changes in another, indicating a dependent relationship between the two. It is commonly known as cause and effect. For instance, when the weather gets hotter, it is observed that people tend to experience an increase in sunburn cases. This implies that the weather serves as the causal factor that influences the occurrence of sunburns.



- Difference between them

Let's consider another scenario using this visualization. Imagine your laptop running out of battery, resulting in an automatic shutdown. As a consequence, the video player you were using also shuts down. While there is a correlation between the events of the computer and video player shutting down, the underlying cause is the depletion of battery power.



- In Data Science

How often have you come across research that suggests a causal relationship between two variables, such as asserting that going to the gym leads to increased productivity and focus? It's crucial as a data scientist to avoid being swayed solely by correlations, as this can result in biased feature engineering and inaccurate conclusions.

Correlation does not imply causation.

Drawing conclusions about causality solely based on correlations can be misleading. When constructing a machine learning model to explore the connection between gym attendance and productivity, it is crucial to consider the underlying factors that truly drive high performance, such as hard work, perseverance, and consistent routines. By prioritizing these causative elements, the model can accurately validate cause-and-effect relationships, rather than relying solely on correlated features like gym attendance which may not necessarily imply causation.