**Statistical concepts that every Data Scientist should know👨🏻‍💻👨🏻‍🎓!!**

[[](https://medium.com/@dhilip_maharish?source=post_page-----478b90a997ad--------------------------------)](https://medium.com/@dhilip_maharish?source=post_page-----478b90a997ad--------------------------------)

[[AI Mind](https://pub.aimind.so/?source=post_page-----478b90a997ad--------------------------------)](https://pub.aimind.so/?source=post_page-----478b90a997ad--------------------------------)

[Dhilip Maharish](https://medium.com/@dhilip_maharish?source=post_page-----478b90a997ad--------------------------------)

·

[Follow](https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fsubscribe%2Fuser%2F9c6dfb453a3&operation=register&redirect=https%3A%2F%2Fpub.aimind.so%2Fstatistical-concepts-that-every-data-scientist-should-know-478b90a997ad&user=Dhilip+Maharish&userId=9c6dfb453a3&source=post_page-9c6dfb453a3----478b90a997ad---------------------post_header-----------)

Published in

[AI Mind](https://pub.aimind.so/?source=post_page-----478b90a997ad--------------------------------)

·

7 min read

·

Aug 26

2.6K

22

**Essential Statistics concepts to build basic foundation for Modern Data Scientists**📊



**Source: Pixels images**

Inthe world of Data Science, there are some important ideas that makes efficient progress in workflow and also as super tool. These ideas help data scientists make sense of all the information they work in it.

Yes, it is none other than **Statistics.**The basics foundational concepts that build the process in data science.

In this article, we are going to explore how statistical concepts contribute to data science. Whether you’re new to data science or have been doing it for a while, these ideas are like a guidebook. They help you understand numbers better and use them to make smart decisions.

So, let’s deep dive into these essential statistical ideas that make data science so powerful.

First, we can get clear on this **what data science is?**

The title itself explains you, taking Data and applying scientifical concepts like statistics, probability and calculus to derive the meaningful insights out of it.

***Data Science is understanding Past information and predicting future information.***



**Source: Pixels Images**

Examples:

Data science helps us predict the future, like a weather forecast telling us if it will rain tomorrow. It is not a magic it uses number and machine learning. It’s about finding the truth in data. It helps us answer questions and solve problems.

Now we can get into **Why statistics is needed in data science and how it contributes in it?**

***Statistics is the backbone of data science.***

It provides the necessary tools, methods, and principles for data scientists to explore, analyze, and extract valuable insights from data. Without statistics, data science would lack the rigor and reliability needed to make data-driven decisions and solve complex problems.

It contributes to every process in Data science such as

✅**Data Exploration and Summarization**

✅**Data Cleaning and Preprocessing**

✅**Inferential Analysis**

✅**Predictive Modeling**

✅**Feature Selection**

✅**Model Evaluation**

✅**Time Series Analysis**



**Source: Pixels Images**

In statistics, it is broadly classified into various types which applies in Data science are listed below.

1. **Descriptive Statistics**
2. **Inferential Statistics**
3. **Regression Analysis**
4. **Data Sampling**
5. **Feature Selection**
6. **Statistical Evaluation on Model**

**1. Descriptive Statistics**

Descriptive statistics is a branch of statistics that deals with the **presentation** and **summary of data**. Its primary goal is to provide a clear and concise overview of data, allowing for easier interpretation and understanding.

It involves various concepts to make understanding data easier. They are

✅**Mean (Average)**- Measure the average value in the distribution of numerical data.

✅**Median**- Provide the average information with more efficient way compared to Mean and it is not affected by outlier in data.

✅**Variance**- Measure the Spread in data.

✅**Standard Deviation** — The square root of the variance, providing a more interpretable measure of data variability.

✅**Percentile**- It is a measure that indicated the percentage of data points that are equal to or below a specific value in a dataset.

✅**IQR (Interquartile range)**- It is the measure of range between first quartile and third quartile which helps to identify middle of 50 % of data.

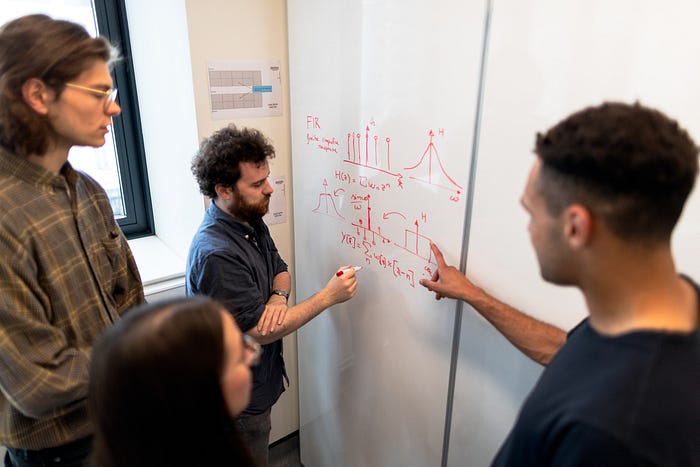
✅**Histogram**- It is the measure of frequency or count of data points falling into specific intervals (bins) along the horizontal axis.

✅**PDF (Probability Density Function)**-It is a statistical function that describes the likelihood of a continuous random variable taking on a specific value within a given range.

✅**CDF (Cumulative Density Function)**- It is a statistical function that gives the cumulative probability that a random variable is less than or equal to a specific value.

✅**Skewness**- It describes the asymmetry in the distribution of data.

✅**Kurtosis**- It measures the tailedness of the data distribution.



**Source: Pixels Images**

**2.Inferential Statistics**

Inferential statistics is a branch of statistics involves data to make **inferences**, **predictions**, or **generalizations** about populations based on sample data. It helps us to draw conclusions or make statements about a larger group (population) by analyzing a smaller, representative subset of that group (sample).

✅**Hypothesis Testing**- It formulate hypotheses about population parameters (e.g., population mean) and use sample data to test whether these hypotheses are supported or refuted.

✅**Estimation**- It estimate population parameters based on sample data.

✅**Confidence Interval**- It provide a range of values within which a population parameter is likely to fall.

✅**Statistical Tests-** A wide range of statistical tests, such as t-tests, chi-squared tests, ANOVA, and regression analysis, are used in inferential statistics to compare groups, assess relationships, and make predictions.

✅**Level of Significance-** It often denoted by α, which represents the probability of making a Type I error ie., incorrectly rejecting a true null hypothesis.



**Source: Pixels Images**

**3. Regression Analysis**

Regression analysis is the statistical technique used in Data science which quantify the **relationship between** one or more **independent variables** (predictors) and a **dependent variable** (outcome) in order to make predictions or understand the impact of the predictors on the outcome.

✅**Linear Regression**- It makes relationship between a dependent variable and one or more independent variables by fitting a linear equation to the data.

✅**Multiple Regression**- It incorporate two or more independent variables to predict a single dependent variable.

✅**Polynomial Regression**- It make relationship between variables appears to be nonlinear, this model fits a polynomial (e.g., quadratic or cubic) equation to the data.

✅**Ridge Regression and Lasso Regression-** Variations of linear regression that incorporate regularization techniques to handle multicollinearity and prevent overfitting.

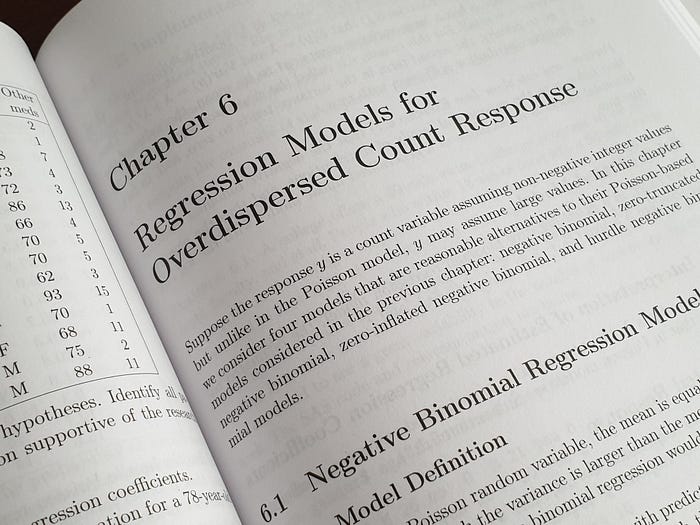


Photo by [Enayet Raheem](https://unsplash.com/@raheemsphoto?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com/?utm_source=medium&utm_medium=referral)

**4. Data Sampling**

Data sampling is a statistical technique used in data science to select a **subset** of data points from a larger dataset. The purpose of sampling is to make data analysis more manageable, cost-effective, and practical, especially when working with large or extensive datasets.

✅**Random Sampling**- In this method, every item or member in the population has an equal chance of being selected for the sample. It reduces bias and ensures that the sample is representative of the population.

✅**Stratified Sampling-** The population is divided into subgroups or strata based on certain characteristics (e.g., age, gender, location). Then, random sampling is performed within each stratum to ensure representation of all groups.

✅**Systematic Sampling-** The starting point is randomly chosen, and then every “kth” item is included in the sample. It’s simple and often more efficient than simple random sampling.



**Source: Pixels Images**

**5.Feature Selection**

It the Statistical techniques which guides in selection of relevant features (variables) for predictive modeling. Techniques like **feature importance** and **correlation analysis** help data scientists choose the most influential factors.

✅**Correlation-Based Feature Selection-** Selects features based on their correlation with the target variable, removing redundant or highly correlated features.

✅**Tree-Based Feature Importance-** Decision tree and ensemble models (e.g., Random Forest, Gradient Boosting) can provide feature importance scores, which can be used to select the most important features.

✅**Mutual Information-** Measures the dependency between features and the target variable, selecting features with high mutual information.

✅**L1 Regularization (Lasso)**- Encourages sparsity in the model by penalizing the absolute values of feature coefficients, effectively selecting a subset of features.



**Source: Pixels Images**

**6.Statistical Evaluation on Model**

It involves various statistical metrics and tests to quantitatively measure how well the model performs.

✅**Accuracy-**Accuracy measures the proportion of correctly classified instances in a classification model.

✅**Mean Absolute Error (MAE)-**MAE measures the average absolute difference between the predicted values and the actual values.

✅**Mean Squared Error (MSE)-**MSE calculates the average of the squared differences between predicted and actual values.

✅**Root Mean Squared Error (RMSE)**- RMSE is the square root of MSE, providing an interpretable metric in the same units as the target variable.

✅**R-squared (R²) or Coefficient of Determination**- R² measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model.

✅**Area Under the Receiver Operating Characteristic (ROC AUC)-**Itmeasures the area under the receiver operating characteristic curve, which plots the trade-off between true positive rate (recall) and false positive rate at various thresholds.

✅**Confusion Matrix-**A table that shows the number of true positives, true negatives, false positives, and false negatives, providing detailed insights into the performance of a classification model.

✅**Precision-**Measures the ratio of true positive predictions to the total positive predictions, emphasizing the model’s ability to avoid false positives.

✅**Recall-** Measures the ratio of true positives to the total actual positives, emphasizing the model’s ability to find all relevant instances.

✅**F1-Score-** The harmonic mean of precision and recall, offering a balance between the two metrics.



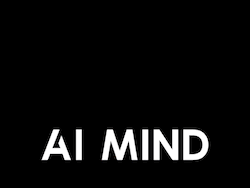
Photo by [ThisisEngineering RAEng](https://unsplash.com/@thisisengineering?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com/?utm_source=medium&utm_medium=referral)

If you like the above article kindly provide claps and

Follow me in medium: [Dhilip Maharish — Medium](https://medium.com/@dhilip_maharish)

Follow me in Linkedin: [Dhilip Maharish | LinkedIn](https://www.linkedin.com/in/dhilip-kumar-ds/)

**A Message from AI Mind**



Thanks for being a part of our community! Before you go:

* 👏 Clap for the story and follow the author 👉
* 📰 View more content in the [AI Mind Publication](https://pub.aimind.so/)
* 🧠 Improve your [AI prompts effortlessly and FREE](https://www.aimind.so/prompt-generator?utm_source=pub&utm_medium=message)
* **🧰 Discover**[**Intuitive AI Tools**](https://www.aimind.so/?utm_source=pub&utm_medium=message)