(1) Statistics for Data Science- NTRODUCTION

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Amar Sahay

Statistics for Data Science, Analytics, and Machine Learning

Importance of Statistics in **Data Science**, **Analytics**, and **Machine Learning.**

This and subsequent sections explain the topics of **Statistics and Data Analysis** along with the statistical concepts essential to understanding and applying Statistics to Data Science, Analytics, and Machine Learning. Those working in these areas will find the information useful.

There are a number of learning material with background, concepts, examples, and computer applications with instructions. A step-by step explanation of the key statistical topics is provided in subsequent sections. These learning materials are designed to provide fundamental concepts to those who are working in the area of Data Science and Machine Learning. You should be able to download some of the key concepts for free or purchase some of them at a minimal cost.

Statistics is the **mathematical backbone** of Data Science, Analytics, and Machine Learning. These are data-driven decision-making process and they rely heavily on data. While dealing with data, computing power and applying models and algorithms (e.g., neural networks), it is statistics that provides the **framework for understanding data, making decisions under uncertainty, and validating results**.

Why Statistics is Critical in these areas?

The knowledge of Statistics is critical as it provides:

(a) Foundation for Data Understanding

Before building models, we must **understand the data** and answer the following questions:

- What is the distribution?
- Are there outliers?
- Is the data skewed or symmetric?

Statistics provides tools to **summarize and visualize data**, enabling meaningful insights. It also provides and lets us calculate the variability in the data.

Note that variation is integral part of data and all data have variation. We will discuss variability is detail in the subsequent sections. Statistics lets us quantify the variation. Knowledge of variability in data is critical in statistical applications and modelling,

Statistics lets us make decisions and draw conclusions using limited data (discussed in detail later).

(b) Statistical methods enable decision making under uncertainty.

Real-world data is noisy and incomplete. Statistical methods help:

- Quantify uncertainty
- Estimate confidence in predictions
- Make risk-aware decisions

Example:

Predicting customer churn with a probability (not certainty)

(c) Model Building and Validation

Most machine learning methods are rooted in statistical principles:

- Linear regression → based on statistical estimation
- Bayesian learning → probabilistic inference
- Loss functions → derived from likelihood concepts

Statistics helps:

- Avoid overfitting
- Evaluate model performance
- Compare models rigorously

(d) Data-Driven Inference

In analytics, we often ask:

- “Does this change improve performance?”
- “Is this pattern real or random?”

Statistical inference ensures conclusions are:

Valid
Generalizable
Not due to chance

(e) Experimental Design (A/B Testing)

Used heavily in business and ML:

- Compare two models or strategies
- Determine statistically significant improvements

A knowledge of statistics is essential in Data Science, Analytics and, Machine learning. We begin this journey by providing a deep understanding of Statistics and its importance in Data Science.