

Introduction to SDA

1 Probability vs Statistics

Probability theory and statistics deal with the same mathematical objects but solve inverse problems regarding the flow of information.

- **Probability:**

- **Given:** A fully defined probabilistic model (distribution $F(x)$, parameters θ).
- **Goal:** Deduce the likelihood of observing specific outcomes (data).
- **Direction:** General → Specific.

- **Statistics:**

- **Given:** Observed data (a specific realization of a random process).
- **Goal:** Infer the properties of the underlying model (distribution $F(x)$, parameters θ).
- **Direction:** Specific → General.

2 Foundations of sampling

To apply mathematical methods to real-world data, we treat observations as realizations of random variables.

2.1 The Concept of a Sample

In mathematical statistics, we bridge the gap between abstract probability theory and real-world observations.

Definition 1 (Sample). A **sample** of size n is a sequence of observations (x_1, x_2, \dots, x_n) obtained from a specific process or phenomenon. Formally, we view these data points through two lenses:

1. **Theoretical (Random Sample):** A vector of random variables $X^n = (X_1, \dots, X_n)$. This represents the data *before* it is collected.
2. **Empirical (Realization):** A vector of specific real numbers $x^n = (x_1, \dots, x_n) \in \mathbb{R}^n$. This is the actual data *after* the experiment.

Context	Random Variable (X_i)	Realization (x_i)
Meteorology	Daily temperature in the campus at 12:00 PM.	22.5°C, 21.0°C, ...
Retail Analytics	Number of customers entering a store between 6:00–7:00 PM.	45, 12, 38, 55, ...
Campus Life	Number of students connected to the campus Wi-Fi at noon.	1250, 1180, 1310, ...
Quality Control	Lifespan of a battery produced on a specific line (in hours).	120.5, 118.2, 122.1, ...

Table 1: Examples

2.2 Real-World Examples of Samples

To explain the concept of a sample to students, we can map physical observations to mathematical notation. In each case, we assume that the process is observed n times under similar conditions.

- **Interpreting the i.i.d. assumption in these cases:**

- *Identically Distributed*: We assume the underlying “mechanism” (the weather patterns, the store’s popularity, the class schedule) remains constant during the observation period.
- *Independent*: We assume that the number of customers today does not influence the number of customers tomorrow.

Note for students: In reality, the i.i.d. assumption is often an approximation. For example, daily temperature is often *dependent* (if today is extremely hot, tomorrow is likely to be warm too).

2.3 Fundamental Assumptions (i.i.d.)

To make the model mathematically tractable, we usually adopt the **i.i.d. assumption**. We assume that the sample is “born” from a sequence of random variables that are:

1. **Independent**: The outcome of one observation does not affect another.
2. **Identically Distributed**: Every observation is generated by the same underlying probabilistic law (F).

Intuition for students: Imagine the same experiment being performed n times under identical, independent conditions. Each x_i is just one of many possible outcomes that the random variable X_i could have taken.

2.4 Simple random sample (i.i.d.)

To allow for tractable inference, we often impose structure on the joint distribution of X^n .

Definition 2 (i.i.d.). The variables X_1, \dots, X_n are **independent and identically distributed (i.i.d.)** if:

1. They are mutually independent: $P(X_1 \leq x_1, \dots, X_n \leq x_n) = \prod_{i=1}^n P(X_i \leq x_i)$.
2. They follow the same distribution: $F_{X_i}(x) = F(x)$ for all i .

A sample satisfying these conditions is called a **Simple Random Sample**.

2.5 Representativeness

Representativeness is the property that the sample accurately reflects the population. Mathematically, this implies that the sampling mechanism is unbiased (every element of the target population has a non-zero (often equal) probability of being included). If this condition is violated, $\hat{\theta}$ may not converge to θ .

3 Statistical model

A statistical model formalizes the assumptions we make about the data generation process.

Definition 3 (Statistical Model). A statistical model is a pair $(\mathcal{X}^n, \mathcal{P})$, where:

- \mathcal{X}^n is the sample space (the set of all possible observations).
- $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a family of probability distributions on \mathcal{X}^n , parameterized by θ .

Our goal is to identify which $P_\theta \in \mathcal{P}$ best describes the observed data x^n .

4 Order statistics

Often the order in which data arrives is irrelevant, and only the values matter. Sorting the sample allows us to study its distribution more effectively.

Definition 4 (Order Statistics). Let (X_1, \dots, X_n) be a random sample. If we sort these variables in non-decreasing order:

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

the resulting sequence is called the **order statistics** of the sample.

- $X_{(1)} = \min(X_1, \dots, X_n)$ is the **first order statistic** (minimum).
- $X_{(n)} = \max(X_1, \dots, X_n)$ is the **n-th order statistic** (maximum).
- The value $X_{(k)}$ is called the **k-th order statistic**.

5 Statistics and estimators

Definition 5 (Statistic). A statistic $T(X^n)$ is any measurable function of the sample X^n that does not depend on unknown parameters θ .

Important Note. *Is a statistic a number or a random variable?* The answer depends on whether the data has already been collected:

- **Before the experiment ($T(X^n)$):** The statistic is a **random variable**. Since the observations X_1, \dots, X_n are random, any function of them is also random. It has its own probability distribution, which we call the **sampling distribution**.
- **After the experiment ($T(x^n)$):** Once we replace the variables with actual measurements (e.g., $x_1 = 22.5, x_2 = 21.0$), the statistic becomes a **fixed number** (a realization).

Example 1. The sample mean $\bar{X} = \frac{1}{n} \sum X_i$ is a **random variable** (it could have been different if we had picked a different day). But the value $\bar{x} = 21.75$ calculated from today's temperatures is just a **number**.

Characteristic	Population Parameter	Sample Estimator
Mean	$\mathbb{E}X = \int x dF(x)$	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
Variance	$\mathbb{D}X = \mathbb{E}(X - \mathbb{E}X)^2$	$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
Quantile	$x_\alpha = F^{-1}(\alpha)$	$X_{([\alpha n])}$ or interpolated $X_{(k)}$
Median	$F^{-1}(0.5)$	$m = \begin{cases} X_{((n+1)/2)} & n \text{ odd} \\ \frac{1}{2}(X_{(n/2)} + X_{(n/2+1)}) & n \text{ even} \end{cases}$
Skewness	$\gamma_1 = \mathbb{E} \left[\left(\frac{X - \mu}{\sigma} \right)^3 \right]$	$g_1 = \frac{\sqrt{n} \sum (X_i - \bar{X})^3}{(\sum (X_i - \bar{X})^2)^{3/2}}$
Mode	$\operatorname{argmax}_x f(x)$	Mode of the histogram / KDE

Table 2: Correspondence between theoretical distribution characteristics and sample statistics.