24 Apr

$X_1, \ldots, X_n$ are IID from an unknown distribution with CDF F. We want to estimate F.

The empirical CDF is defined as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x)$$

Properties

Lemma: For any fixed $x \in \mathbb{R}$.

$$\hat{F}_n(x) \xrightarrow{p} F(x)$$

as $n \to \infty$

Pf: Define

$$Y_i := \mathbb{I}(X_i \le x)$$

$$= \begin{cases} 1 & \text{if } X_i \le x \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{P}(Y_i = 1) = \mathbb{P}(X_i \le x) = F(x)$$

$$\mathbb{P}(Y_i = 0) = 1 - F(x)$$

So, $Y_i \overset{IID}{\sim} \text{Bernoulli}(F(x))$

Now

"WLLN for IID RVs ...
finite variance,

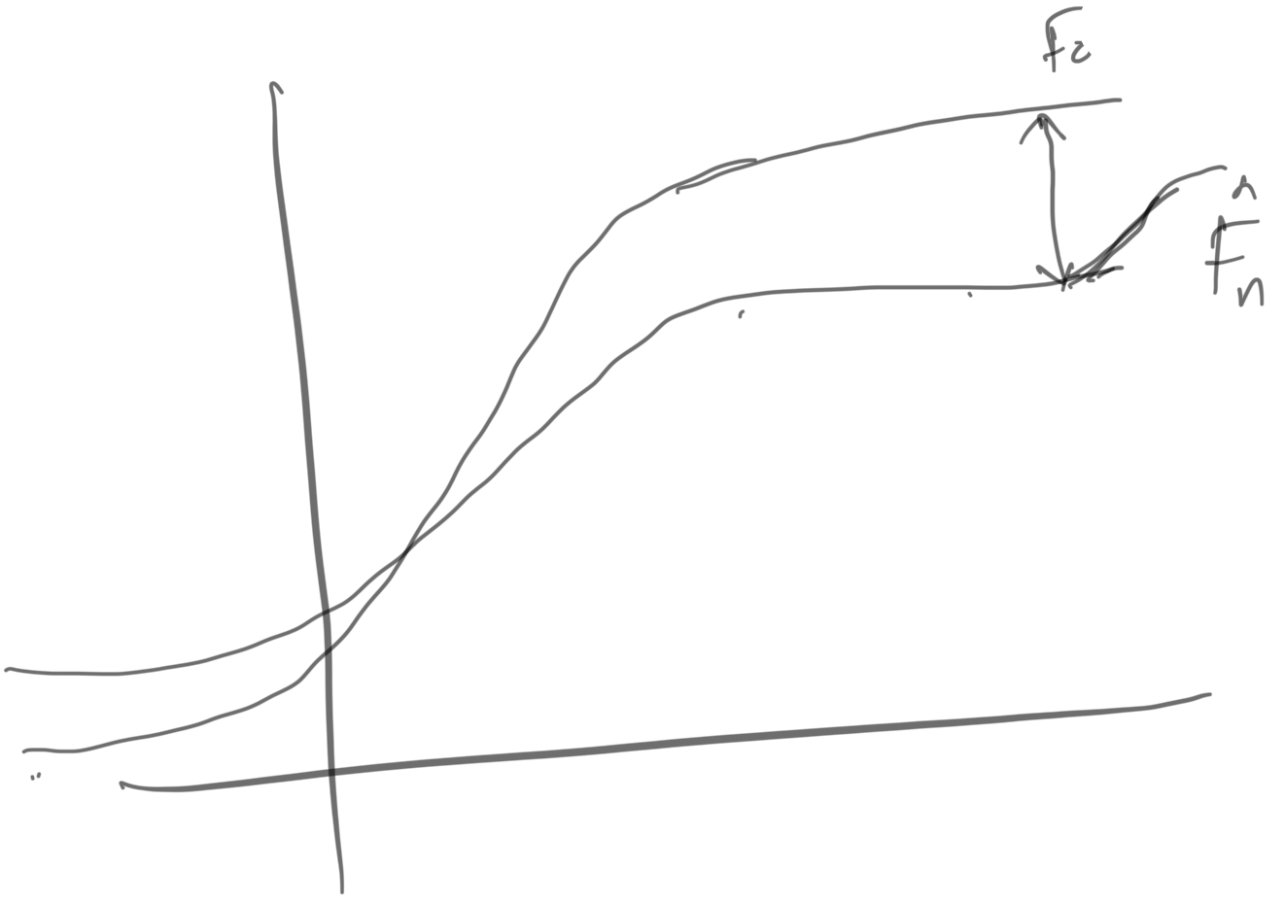
$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} E(Y_i) = F(x)$$

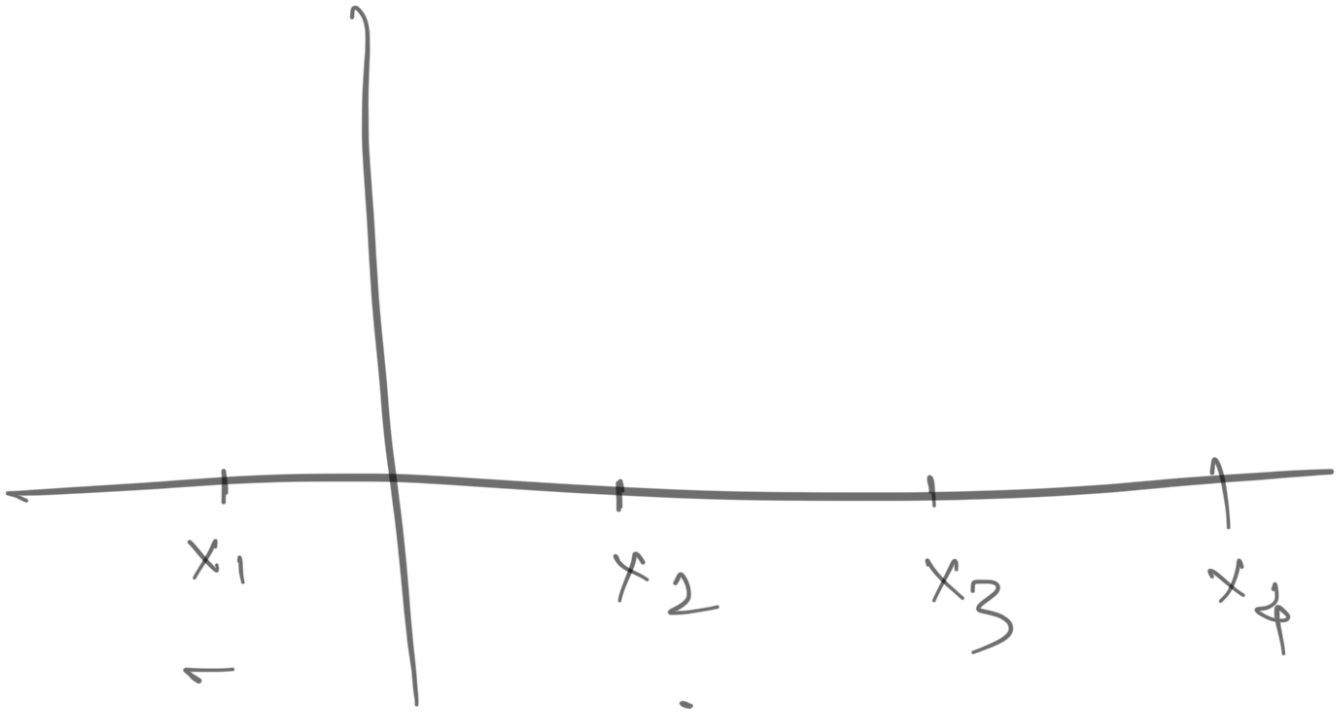
$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \le x) \xrightarrow{p} F(x)$$

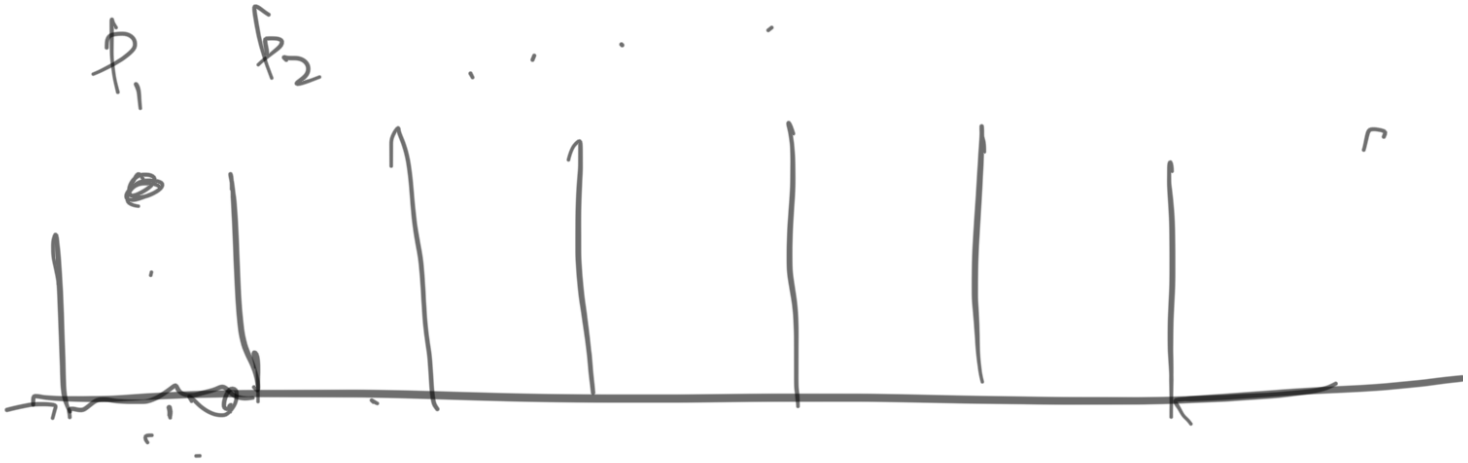
$$\hat{F}_n(x)$$

Lemma: we have CLT for $\hat{F}_n(x)$ as well.

Lemma: $E\left(\hat{F}_n(x)\right) = F(x)$


f_2
$\hat{f}_n$


x_1
x_2
x_3
x_4


f_1
f_2

MTL108
Introduction to Nonparametric Inference (Optional)

Rahul Singh

Nonparametric inference is a branch of statistics in which the data is not assumed to come from a probability distribution that can be completely defined by a finite number of parameters.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with an unknown cumulative distribution function (CDF), $F(x)$.

- In a **parametric** model, we assume that $F(x)$ belongs to a specific family of distributions entirely indexed by a finite-dimensional parameter vector θ. That is, $F \in \{F_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$.
  Example: If we assume the data is Normal, $\theta = (\mu, \sigma^2)$. Once we estimate these two parameters, the entire distribution curve is perfectly known.
- In a **nonparametric** model, the family of possible distributions $\mathcal{F}$ cannot be indexed by a finite-dimensional parameter vector. The parameter space is effectively infinite-dimensional. The only assumptions made are typically very weak and general, such as assuming the distribution is continuous or symmetric.

Note: Nonparametric tests are frequently referred to as **distribution-free tests** because their validity does not depend on the specific shape (e.g., normal, exponential) of the underlying population distribution.

Parametric vs. Nonparametric Inference

The choice between parametric and nonparametric methods represents a fundamental trade-off in statistics between **Power** and **Robustness**.

If the assumptions of a parametric test are perfectly met (e.g., the data is truly Normal), the parametric test will always be more powerful (higher probability of rejecting a false null hypothesis). However, if the assumptions are violated, parametric tests can yield completely invalid results, whereas nonparametric tests remain reliable.

Key Differences

Feature	Parametric Inference	Nonparametric Inference
Underlying Assumption	Data follows a specific, known distribution (usually Normal).	Distribution-free; no specific functional form is assumed.
Parameter Focus	Focuses on estimating specific parameters (e.g., mean μ, variance σ^2).	Focuses on empirical distributions, ranks, or the median.
Data Level	Requires quantitative data (Interval or Ratio scales).	Can handle continuous, Ordinal (ranked), or Nominal data.
Sample Size (n)	Can be highly accurate for small n, *only if* the distribution assumption is strictly true.	Highly preferred for small n when the true distribution shape is unknown or highly skewed.
Outlier Sensitivity	Highly sensitive (outliers drastically shift the mean and variance).	Highly robust (outliers do not drastically shift ranks or medians).

Analogous Statistical Tests

For nearly every standard parametric test, there exists a nonparametric equivalent that relies on sorting and ranking the data rather than calculating raw sums and averages.

- **Two Independent Samples:**

  - *Parametric:* Independent Two-Sample Student's t-test.
  - *Nonparametric:* Wilcoxon Rank-Sum Test (or Mann-Whitney U Test).

- **Paired Samples:**
  - *Parametric:* Paired t-test.
  - *Nonparametric:* Wilcoxon Signed-Rank Test.

- **Comparing Three or More Groups:**
  - *Parametric:* One-Way ANOVA (F-test).
  - *Nonparametric:* Kruskal-Wallis Test.

- **Correlation:**
  - *Parametric:* Pearson Correlation Coefficient (r).
  - *Nonparametric:* Spearman's Rank Correlation Coefficient (ρ).

The Empirical Distribution Function (EDF)

In practice, the true Cumulative Distribution Function (CDF), $F(x) = P(X \leq x)$, is usually unknown. We estimate it directly from the sample data without assuming any parametric form.

Definition: Let $X_1, X_2, \ldots, X_n$ be an independent and identically distributed (i.i.d.) random sample from a population with CDF $F(x)$. The **Empirical Distribution Function (EDF)**, denoted as $\hat{F}_n(x)$, is the proportion of observations in the sample that are less than or equal to x:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

where $I(\cdot)$ is the indicator function defined as:

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x \end{cases}$$

Because $\sum I(X_i \leq x)$ follows a Binomial distribution with parameters n and $p = F(x)$, the expected value and variance of the EDF for any fixed x are:

$$E[\hat{F}_n(x)] = F(x) \quad \text{(unbiased)}$$
$$\text{Var}(\hat{F}_n(x)) = \frac{F(x)(1 - F(x))}{n}$$

Convergence Properties of the EDF

The foundational power of the EDF lies in its convergence to the true CDF as the sample size grows.

Pointwise Convergence

By the Weak Law of Large Numbers (WLLN), for any fixed value of x, the EDF converges in probability to the true CDF:

$$\hat{F}_n(x) \xrightarrow{p} F(x) \quad \text{as } n \to \infty$$

Furthermore, by the Central Limit Theorem (CLT), the standardized EDF converges in distribution to a Normal random variable:

$$\sqrt{n}\left(\hat{F}_n(x) - F(x)\right) \xrightarrow{d} N(0,\, F(x)(1-F(x)))$$

Uniform Convergence (Glivenko-Cantelli Theorem)

Pointwise convergence only guarantees convergence at individual points. The **Glivenko-Cantelli Theorem** (often called the Fundamental Theorem of Statistics) provides a much stronger result. It states that the EDF converges to the true CDF *uniformly* over all possible values of x, with probability 1 (almost surely):

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0 \quad \text{as } n \to \infty$$

This means the maximum vertical distance between our step function and the true continuous curve shrinks to zero everywhere simultaneously.

Goodness of Fit: The Kolmogorov-Smirnov Test

We utilize the Glivenko-Cantelli distance to test whether a sample originates from a specific continuous theoretical distribution.

Hypotheses:

- H_0: The sample comes from the specific distribution $F_0(x)$.
- H_1: The sample does not come from $F_0(x)$.

The Test Statistic (D_n)

The K-S test statistic is the supremum (maximum absolute distance) between the EDF and the theoretical CDF:

$$D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$

Computational Formula: Because $\hat{F}_n(x)$ is a step function that jumps at every observed data point, the maximum distance must occur either immediately before or exactly at one of the data points. Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ be the ordered sample statistics. The exact test statistic is calculated as:

$$D_n = \max_{1 \leq i \leq n} \left\{ \max \left(\left| \frac{i}{n} - F_0(X_{(i)}) \right|, \left| \frac{i-1}{n} - F_0(X_{(i)}) \right| \right) \right\}$$

Example 1. Test if the sample $\{0.2, 0.5, 0.9\}$ comes from a Standard Uniform distribution $U(0,1)$, where $F_0(x) = x$. $n = 3$.

- **At $X_{(1)} = 0.2$:** $F_0(0.2) = 0.2$.
  Upper distance: $|1/3 - 0.2| = 0.133$.
  Lower distance: $|0 - 0.2| = \mathbf{0.200}$.
- **At $X_{(2)} = 0.5$:** $F_0(0.5) = 0.5$.
  Upper distance: $|2/3 - 0.5| = \mathbf{0.167}$.
  Lower distance: $|1/3 - 0.5| = \mathbf{0.167}$.
- **At $X_{(3)} = 0.9$:** $F_0(0.9) = 0.9$.
  Upper distance: $|1 - 0.9| = 0.100$.
  Lower distance: $|2/3 - 0.9| = \mathbf{0.233}$.

The K-S statistic is the absolute maximum of these distances: $D_n = \max(0.200, 0.167, 0.233) = \mathbf{0.233}$. We compare this to the K-S critical values table for $n = 3$ to find the p-value.

Pearson's Chi-Square Goodness of Fit Test

While K-S compares continuous CDFs, Pearson's χ^2 test compares discrete categorical frequencies by mapping the data to a Multinomial distribution.

Hypotheses:

- H_0: $p_1 = p_{1,0}, \ldots, p_k = p_{k,0}$ (The observed category frequencies match theoretical expectations).
- H_1: $p_i \neq p_{i,0}$ for at least one i.

Mathematical Details

Let the data be classified into k mutually exclusive categories. Let O_i be the observed frequency in category i, where $\sum_{i=1}^{k} O_i = n$. Under H_0, the expected frequency is $E_i = n \cdot p_{i,0}$.

The test statistic sums the squared standardized residuals:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

By multivariate normal approximation of the multinomial distribution, as $n \to \infty$, this statistic converges in distribution to a Chi-Square distribution:

$$\chi^2 \xrightarrow{d} \chi^2_{k-1-m}$$

where m is the number of parameters estimated from the sample to generate the expected probabilities.

Example 2. A casino tests a six-sided die by rolling it $n = 60$ times. If fair (H_0), $E_i = 60(1/6) = 10$ for all faces. $k = 6$. Observed frequencies: $\{8, 12, 10, 15, 5, 10\}$.

$$\chi^2 = \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(15-10)^2}{10} + \frac{(5-10)^2}{10} + \frac{(10-10)^2}{10}$$

$$\chi^2 = 0.4 + 0.4 + 0 + 2.5 + 2.5 + 0 = 5.8$$

For $df = 6 - 1 = 5$, the critical value at $\alpha = 0.05$ is $\chi^2_{0.05,5} = 11.07$. Since $5.8 < 11.07$, we fail to reject H_0. There is insufficient statistical evidence to claim the die is loaded.