Recall:

$(\Omega, \mathcal{F}, \mathbb{P})$ - Probability space.

X - Discrete random variable

- Range(X) $\equiv$ finite / Countable
- $\mathbb{P}(X=x)$ for $x \in$ Range(X)
  (p.m.f.)

X - Continuous random variable

- Range(X) $\subseteq \mathbb{R}$
- $f: \mathbb{R} \to [0, \infty)$
  (p.d.f.)

$$\mathbb{P}(X \in (a,b)) = \int_a^b f(x)\,dx$$

Distribution function

$$F(x) = \mathbb{P}(X \leq x).$$

Expectation

$\sum_{k \in \text{Range}(X)} k\,\mathbb{P}(X=k)$ (Discrete)

$\int_{-\infty}^{\infty} x f(x)\,dx$ (Continuous)

Variance

$\sum_{k \in \text{Range}(X)} (k - E[X])^2\,\mathbb{P}(X=k)$ (Discrete)

$\int_{-\infty}^{\infty} (x - E[X])^2 f(x)\,dx$ (Continuous)

Empirical Distribution function

Assume they are chosen from a population X

Let $X_1, X_2, X_3 \ldots, X_n$ be random variables.

$\{x_1, x_2, \ldots, x_n\} \equiv R$

The "empirical distribution" is a discrete distribution on $\mathbb{R}$ with p.m.f given by

$$f(t) = \frac{1}{n} \mid \{i : X_i = t\} \mid .$$

Distribution gives weight $\frac{1}{n}$ to each x_i.

- Sample from Population — Relative Frequency
- Infer attributes from the sample
  - mean
  - variance
  - A - event $\mathbb{P}(A)$

Empirical Distribution

Random variable Y, that has the same distribution as the Empirical distribution is given by:

- $\text{Range}(Y) = \{X_1, X_2, \ldots, X_n\}$,

- $P(Y = t) = \frac{1}{n} \mid \{i : X_i = t\} \mid$,

and

- $\hat{F}_n(x) = P(Y \leq x) = \frac{|\{i : X_i \leq x\}|}{n}$

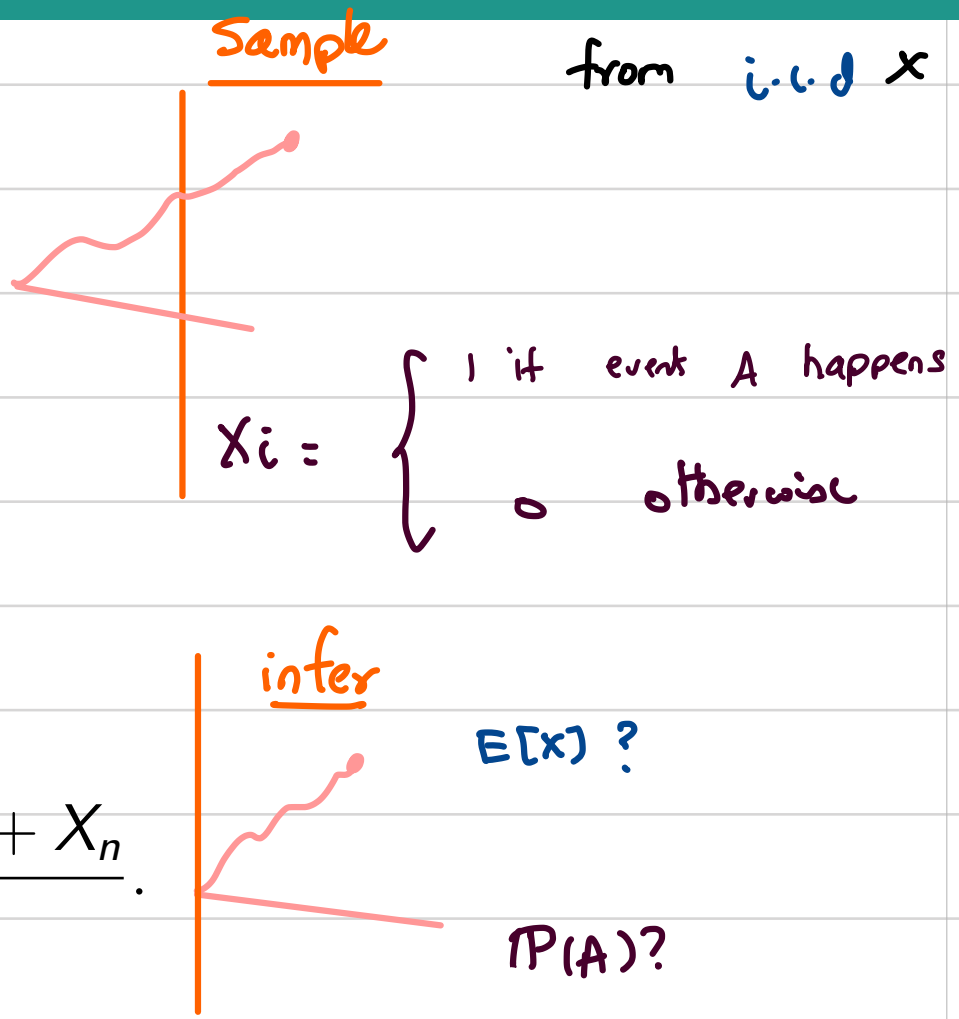
Empirical distribution function

Ex: $E[Y] = \bar{x}$ & $\text{var}[Y] = \frac{n-1}{n} S^2$

Sample Mean

Let $X_1, X_2, X_3 \ldots, X_n$ be i.i.d. random variables. The "sample mean"

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$



Unbiased Estimate and Variance Reduction

Theorem 1 :-

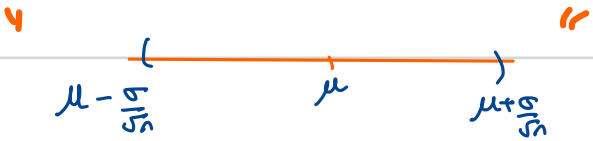
Let $X, X_1, X_2, X_3 \ldots, X_n$ be an i.i.d. sample of random variables.

$E[X] = \mu$ and $Var[X] = \sigma^2$.

Let $\bar{X}$ represent the sample mean. Then

- $E[\bar{X}] = \mu$ and $SD[\bar{X}] = \frac{\sigma}{\sqrt{n}}$.

Effective Range of $\bar{X}$ gets smaller as $n \to \infty$.



Proof:

$X_1, X_2, \ldots, X_n$ have same distribution as X

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- $E[\bar{X}] = E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right]$

(linearity) $= \frac{1}{n} \sum_{i=1}^{n} E[X_i]$

(identically distributed as X) $= \frac{1}{n} \cdot n \, E[X]$

$= E[X]$

- $\mathrm{Var}[\bar{X}] = \mathrm{Var}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right]$

(Scaling) $= \frac{1}{n^2} \mathrm{Var}[X_1 + X_2 + \cdots + X_n]$

See claim 1 below (independence) $= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}[X_i]$

$= \frac{n\sigma^2}{n^2}$

$= \frac{\sigma^2}{n}$

□

Claim 1: X, Y are independent

$\Rightarrow \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Proof: $\mathrm{Var}(X+Y)$

$$= E\left[\left(X+Y - E[X+Y]\right)^2\right]$$

$$= E\left[\left((X - E[X]) + (Y - E[Y])\right)^2\right]$$

$$= E(X - E[X])^2 + E(Y - E[Y])^2$$

$$+ 2\, E[(X - E[X])(Y - E[Y])]$$

(referred to as $\mathrm{Cov}(X, Y)$)

(linearity)

$$= \mathrm{Var}[X] + \mathrm{Var}[Y]$$

$$+ 2\left(E[XY] - E[X]E[Y]\right)$$

($= 0$ if independent)

(independence)

$$= \mathrm{Var}[X] + \mathrm{Var}[Y]$$

□

Sample Variance - function of the sample

Let $X_1, X_2, X_3 \ldots, X_n$ be i.i.d. random variables.

The "sample variance" of these is

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}.$$

Theorem 2:

Suppose the distribution has finite expected value μ and finite variance σ^2 then

$$E[S^2] = \sigma^2.$$

Ex: $\mathrm{Var}[S^2] \longrightarrow 0$ as $n \to \infty$

Proof: $E[\bar{X}^2] = Var[\bar{X}] + (E[\bar{X}])^2$ (Fact)

(Theorem 1) $= \sigma^2/n + \mu^2$ — ①

$j = 1, \dots, n$ $\quad E[X_j^2] = Var[X_j] + (E[X_j])^2$

$= \sigma^2 + \mu^2$ — ②

$$E[\bar{S}^2] = E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right]$$

$$= E\left[\frac{\sum_{i=1}^{n}\left[X_i^2 - 2X_i\bar{X} + (\bar{X})^2\right]}{n-1}\right]$$

$$= \frac{\left(\sum_{i=1}^{n} EX_i^2\right) - 2E\left(\sum_{i=1}^{n} X_i\bar{X}\right) + nE(\bar{X})^2}{n-1}$$

$$\left(\text{As, } \sum_{i=1}^{n} X_i = \bar{X}\right) = \frac{\left(\sum_{i=1}^{n} EX_i^2\right) - 2nE(\bar{X})^2 + nE[\bar{X}]^2}{n-1}$$

$$= \frac{n\sigma^2 + n\mu^2 - n\left(\frac{\sigma^2}{n} + \mu^2\right)}{n-1}$$

$$= \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

□

Sample Proportion

Relationship between : Relative Frequency & Probability.

Let $X, X_1, X_2, X_3 \ldots, X_n$ be i.i.d. random variables.

Interested in: $P(X \in A)$ of any event A (value of p)

Estimate it by:

$$\hat{p}_n = \frac{|\{i : X_i \in A\}|}{n}$$

Is $\hat{p}_n$ a good estimate of p?

Theorem 3:- $E[\hat{p}_n] = p$, $Var[\hat{p}_n] = \frac{p(1-p)}{n}$

Proof: $\hat{p}_n = \frac{|\{i : X_i \in A\}|}{n} := \frac{\sum_{i=1}^{n} Z_i}{n}$

where $Z_i = \begin{cases} 1 & \text{if } X_i \in A \\ 0 & \text{otherwise} \end{cases}$

- $P(Z_i = 1) = P(X_i \in A) = p$
- (Fact) X_i are independent $\Rightarrow$ Z_i are independent
- $\sum_{i=1}^{n} Z_i \sim \text{Binomial}(n, p)$

- $E[\hat{p}_n] = E\left[\frac{\sum_{i=1}^{n} Z_i}{n}\right] = \frac{1}{n} \cdot np = p.$

- $Var[\hat{p}_n] = Var\left[\frac{\sum_{i=1}^{n} Z_i}{n}\right] = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}.$

□

Sampling distributions

$X_1, X_2, \ldots, X_n$ are i.i.d X ; $E[X] = \mu < \infty$

$Var[X] = \sigma^2 < \infty$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad ; \quad S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

Lemma 1

X_1, X_2 are i.i.d Normal $(0,1)$

$$Y = X_1 + X_2 \Rightarrow Y \overset{d}{=} \text{Normal}(0,2)$$

Proof: $y \in \mathbb{R}$

$$\mathbb{P}(Y \le y) = \mathbb{P}(X_1 + X_2 \le y)$$

$$= \iint_{\{(x_1,x_2) : x_1 + x_2 \le y\}} f(x_1, x_2)\, dx_1\, dx_2$$

(independence) $$= \int_{-\infty}^{\infty} \int_{-\infty}^{y - x_2} f_{X_1}(x_1)\, f_{X_2}(x_2)\, dx_1\, dx_2$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{y-x_2} \frac{e^{-\frac{x_1^2}{2}}}{\sqrt{2\pi}} \frac{e^{-\frac{x_2^2}{2}}}{\sqrt{2\pi}} \, dx_1 \, dx_2$$

$$= \int_{-\infty}^{\infty} \frac{e^{-\frac{x_2^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{y-x_2} \frac{e^{-\frac{x_1^2}{2}}}{\sqrt{2\pi}} \, dx_1 \, dx_2$$

$$f(y) = \frac{d}{dy} \mathbb{P}(Y \le y) \overset{!}{=} \text{(Ex.)} \int_{-\infty}^{\infty} \frac{e^{-\frac{x_2^2}{2}}}{\sqrt{2\pi}} \frac{d}{dy} \left[\int_{-\infty}^{y-x_2} \frac{e^{-\frac{x_1^2}{2}}}{\sqrt{2\pi}} \, dx_1 \right] dx_2$$

$$= \int_{-\infty}^{\infty} \frac{e^{-\frac{x_2^2}{2}}}{\sqrt{2\pi}} \frac{e^{-\frac{(y-x_2)^2}{2}}}{\sqrt{2\pi}} \, dx_2$$

$$= \int_{-\infty}^{\infty} \frac{e^{-\frac{x_2^2}{2}} e^{-\frac{(y^2 - 2yx_2 + x_2^2)}{2}}}{2\pi} \, dx_2$$

$$= \frac{1}{2\pi} e^{-\frac{y^2}{2}} \int_{-\infty}^{\infty} e^{-(x_2^2 - yx_2)} \, dx_2$$

$$= \frac{1}{2\pi} e^{-\frac{y^2}{2}} \int_{-\infty}^{\infty} e^{-(x_2 - \frac{y}{2})^2 + \frac{y^2}{4}} \, dx_2$$

$$= \frac{e^{-y^2/4}}{2\pi} \frac{\sqrt{2\pi}}{\sqrt{2}} \underbrace{\frac{1}{\sqrt{2\pi}\sqrt{\frac{1}{2}}} \int_{-\infty}^{\infty} e^{-\frac{(x_2 - \frac{y}{2})^2}{2 \cdot \frac{1}{2}}} \, dx_2}_{\left(\text{Normal}\left(\frac{y}{2}, \frac{1}{2}\right) \text{ density integrates to } 1\right)}$$

$$= \frac{e^{-y^2/4}}{\sqrt{2}\sqrt{2\pi}} \cdot 1$$

$\square$

Ex: $X_1 \overset{d}{=} \text{Normal}(\mu_1, \sigma_1^2) \perp X_2 \overset{d}{=} \text{Normal}(\mu_2, \sigma_2^2)$

$X_1 + X_2 \overset{d}{=} \text{Normal}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$

$X_i \overset{d}{=} \text{Normal}(\mu_i, \sigma_i^2) \quad 1 \le i \le n \quad X_i \perp X_j \ (i \neq j)$

$$\sum_{i=1}^{n} \alpha_i X_i \overset{d}{=} \text{Normal}\left(\sum_{i=1}^{n} \alpha_i \mu_i , \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2 \right)$$

Using the above

$\bar{X} = \frac{X_1 + \cdots + X_n}{n} \qquad X_i \overset{d}{=} \text{Normal}(\mu, \sigma^2)$ i.i.d

$\Rightarrow \bar{X} \overset{d}{=} \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$

Theorem 4: $X_1, X_2, \ldots, X_n$ i.i.d Normal(μ, σ^2)

(i) $\bar{X} \overset{d}{=}$ Normal$\left(\mu, \frac{\sigma^2}{n}\right)$

(ii) $\frac{(n-1)S^2}{\sigma^2} := \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \overset{d}{=} \chi^2_{n-1}$

(iii) $\bar{X}$ and S^2 are independent

Corollary 1: (i) $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{d}{=}$ Normal$(0, 1)$

(ii) $$= \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}}} \overset{d}{=} t_{n-1}$$

t_{n-1} → t-distribution

Proof: In due course ☺