

Summary : Population, Sample, Sampling Distributions

- Random Experiment : E , Event Space of E : S
- Random Variable : $X: S \rightarrow \mathbb{R}$
- Population : observed values of X in infinite repetitions of E
 $x_1, x_2, \dots, x_n, \dots \dots \dots$
- Random Sample of size n : Observed values of X in n repetitions of E
 (x_1, x_2, \dots, x_n)
- Distribution of Population \equiv Distribution of the random variable X
 $F_X(x)$: d.f. of X ; $f_X(x)$: p.m.t. / p.d.f.

(x_1, x_2, \dots, x_n)

- Distribution of a Sample \equiv Distribution of the fake random variable \hat{X}

Spectrum of $\hat{X} = \{\underbrace{x_1, x_2, \dots, x_n}\}_{\text{spectrum}} \text{ with } P(\hat{X} = x_i) = \frac{1}{n}$

- Sample Characteristics: ✓ Sample mean: $\bar{x} = E(\hat{X}) = \frac{1}{n} \sum x_i$

$$\begin{aligned} \checkmark \text{Sample variance: } S^2 &= E\{(\hat{X} - \bar{x})^2\} \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

$\checkmark a_k, \checkmark m_k$ etc.

- Statistic: A real-valued function $\beta(x_1, x_2, \dots, x_n)$ of sample values (considering as variables).

$\left(\text{distri of f.v. corr. to statistics} \right)$

- "Random Variable" corresponding to a random sample of size n:

(X_1, X_2, \dots, X_n) where X_i corresponds to x_i
 \hookrightarrow (n -dimensional random variable) $i=1, 2, \dots, n.$

- Mathematically: $X : S \rightarrow \mathbb{R}$ (population r.v.)
- Define: $X_i : \underbrace{S \times S \times \dots \times S}_{S \times S \times \dots \times S} \rightarrow \mathbb{R}$ as $S = \{1, 2, \dots, 6\}$
- $\checkmark X_i(\omega_1, \omega_2, \dots, \omega_n) = X(\omega_i) = x_i \leftarrow i\text{-th coordinate of the random sample}$
- $P(X_i \leq 2) = P(X \leq 2)$

- By defⁿ: X_1, X_2, \dots, X_n are mutually independent
 - Each X_i has the same distribution as X :
- $\checkmark P(X_i \leq x) = P(X \leq x) \quad \forall x \in \mathbb{R}$

- Sampling Distribution \equiv Distribution of the random variable $\beta(x_1, x_2, \dots, x_n)$ corresponding to a statistic.
 (Required to measure goodness of a statistic)
- Important Statistics and Sampling Distributions:

- For a normal (m, σ) population,

$$\left\{ \begin{array}{l} \checkmark \cdot U = \frac{\bar{X} - m^*}{\sigma / \sqrt{n}} \sim N(0, 1) \\ \checkmark \cdot X^2 = \frac{n S^2}{\sigma^2} \sim \chi^2(n-1) \\ \checkmark \cdot t = \frac{\bar{X} - m^*}{S / \sqrt{n}} \sim t(n-1) \text{ where } S^2 = \frac{n}{n-1} S^2 \end{array} \right.$$

Estimation of Parameters

Problem: Let the functional form of the distribution function of the population r.v. X is known but contains a number of unknown parameters $\theta_1, \theta_2, \dots, \theta_K$, i.e.

population d.f. is $F_X(x; \theta_1, \theta_2, \dots, \theta_K)$.

To estimate these unknown parameters based on a random sample;

(x_1, x_2, \dots, x_n) .

V. Estimation of Parameters

- Point Estimate
- Interval Estimate

• Testing Hypothesis

Up Book: : .

S.K. Sen
S. De

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$\sigma > 0$$

Estimate and Estimator

Let α be an unknown population parameter and $a(x_1, x_2, \dots, x_n)$ be a statistic. If the probability mass in the sampling distribution of the statistic ' a ' is concentrated near the point ' α ' then ' a ' is called an estimate of ' α '.

Mathematically, $\sqrt{P}(\alpha - \epsilon < A < \alpha + \epsilon) \approx 1$

$$\text{or, } F_A(\alpha + \epsilon) - F_A(\alpha - \epsilon) \approx 1$$

for some $\epsilon > 0$. Small ϵ gives a good estimate.

$A = a(x_1, x_2, \dots, x_n)$ is called an estimator of α .

Unbiased Estimate

Let $A = a(x_1, x_2, \dots, x_n)$ be an estimator of an unknown population parameter α .

If $E(A) = \alpha$ then ' a ' is an unbiased estimate of α

If $E(A) \neq \alpha$, then ' a ' is a biased estimate and $E(A) - \alpha$ is called the bias.

Consistent Estimate:

Let $A = a(X_1, X_2, \dots, X_n)$ be an estimator of an unknown population parameter α .

If $A \xrightarrow{\text{in P}} \alpha$ as $n \rightarrow \infty$

then 'a' is called a consistent estimate of α .

$$\Leftrightarrow \lim_{n \rightarrow \infty} P(|A - \alpha| \geq \epsilon) = 0 \quad \forall \epsilon > 0.$$

PS-②

Prob ①

Prove that sample mean is consistent and unbiased estimate of the population mean.

Sol. Let, (x_1, x_2, \dots, x_n) be a random sample of size n from a population of r.v. X .

Sample $\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$

r.v. of \bar{X} : $\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$ when x_i 's are i.i.d as X

$E(\bar{X}) = E(X) = m$. when m is the population mean

Again $\bar{X} \xrightarrow{\text{inP}} m$ as $n \rightarrow \infty$ (using LLN) (check!).

Prob 4: Prove that the sample variance is a consistent, but biased estimate of population variance.

Sol.

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

r.v. corr. to the sample variance:

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

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

Let population mean is m and population variance is σ^2 .

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

$$= \dots$$

$$= \sigma^2 - E\{(\bar{x} - m)^2\}$$

$$E\{(\bar{x} - m)^2\} = E\left\{ \left(\frac{x_1 + x_2 + \dots + x_n}{n} - m \right)^2 \right\} = \dots = \frac{\sigma^2}{n}$$

$$E(S^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2 \quad (\text{check!})$$

$$\Rightarrow E(S^2) \neq \sigma^2$$

$\Rightarrow S^2$ is a biased estimator of σ^2

We define an unbiased estimate of population variance σ^2 as :

*
$$\beta^2 = \frac{n}{n-1} S^2$$

Since $E(\beta^2) = \frac{n}{n-1} E(S^2) = \frac{n}{n-1} \frac{n-1}{n} \sigma^2 = \sigma^2$

Q: How to compute an Estimate?

The method of maximum likelihood

(introduced by R.A. Fisher, 1912)

Let, $\theta_1, \theta_2, \dots, \theta_k$ be the unknown population parameters and let (x_1, x_2, \dots, x_n) be a random sample of size n from the given population. We define a likelihood function of the ^{random} sample (x_1, x_2, \dots, x_n) as:

$$L(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_k)$$

$$= P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \quad \text{for } k \text{ discrete}$$

case
(i.e. if X is
discrete)

$$= P(X_1 = x_1) P(X_2 = x_2) \dots P(X_n = x_n)$$

$$= P(X = x_1) P(X = x_2) \dots P(X = x_n)$$

$$= f_{x_1}(\theta_1, \theta_2, \dots, \theta_k) f_{x_2}(\theta_1, \theta_2, \dots, \theta_k) \dots f_{x_n}(\theta_1, \theta_2, \dots, \theta_k)$$

For the continuous Case:

$$P(x_1 < X_1 \leq x_1 + dx_1, \dots, x_n < X_n \leq x_n + dx_n)$$

$$= P(x_1 < X_1 \leq x_1 + dx_1) \dots P(x_n < X_n \leq x_n + dx_n)$$

$$= f_X(x_1; \theta_1, \theta_2, \dots, \theta_n) \dots f_X(x_n; \theta_1, \theta_2, \dots, \theta_n)$$

$$dx_1 dx_2 \dots dx_n$$

$$L(\underbrace{x_1, x_2, \dots, x_n}_{\text{ }}; \theta_1, \theta_2, \dots, \theta_n)$$

$$= f_X(x_1; \theta_1, \theta_2, \dots, \theta_n) \dots f_X(x_n; \theta_1, \theta_2, \dots, \theta_n)$$

$$L = \begin{cases} f_{x_1}(\theta_1, \theta_2, \dots, \theta_k) \dots \dots f_{x_n}(\theta_1, \theta_2, \dots, \theta_k) & \text{if } X \text{ is discrete} \\ f_X(x_1; \theta_1, \theta_2, \dots, \theta_k) \dots \dots f_X(x_n; \theta_1, \theta_2, \dots, \theta_k) & \text{if } X \text{ is cont} \end{cases}$$

* 

Principle of MLE method of maximum likelihood is finding the values of $\theta_1, \theta_2, \dots, \theta_k$ for a fixed sample (x_1, x_2, \dots, x_n) such that L becomes  globally maximum for these values of $\theta_1, \theta_2, \dots, \theta_n$.

$$\text{If } \hat{\theta}_1 = \hat{\theta}_1(x_1, x_2, \dots, x_n)$$

$$\hat{\theta}_2 = \hat{\theta}_2(x_1, x_2, \dots, x_n)$$

$$\vdots$$
$$\hat{\theta}_k = \hat{\theta}_k(x_1, x_2, \dots, x_n)$$

be the values of $\theta_1, \theta_2, \dots, \theta_k$ for which L is globally maximum for a fixed sample, i.e.

$$L(x_1, x_2, \dots, x_n; \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k) \geq L(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_k)$$

for all admissible values of $\theta_1, \theta_2, \dots, \theta_k$

then $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ are called the maximum likelihood estimates (MLE) of $\theta_1, \theta_2, \dots, \theta_k$ respectively.