Summary : Population, Sample, Sampling Distributions

- **Random Experiment** : E , Event Space of E : S
- **Random Variable** : $X : S \to \mathbb{R}$
- **Population** : observed values of X in infinite repetitions of E

  $x_1, x_2, \ldots, x_n, \ldots\ldots$
- **Random Sample of size n** : observed values of X in n repetitions of E

  $(x_1, x_2, \ldots, x_n)$
- **Distribution of Population** $\equiv$ Distribution of the random variable X

  $F_X(x)$: d.f. of X ; $f_X(x)$: p.m.f. / p.d.f.

$(x_1, x_2, \ldots, x_n)$

- **Distribution of a Sample** $\equiv$ Distribution of the Fake random variable $\mathring{X}$

  Spectrum of $\mathring{X} = \{x_1, x_2, \ldots, x_n\}$ with $P(\mathring{X} = x_i) = \frac{1}{n}$ ✓✓

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

  ✓ Sample variance: $S^2 = E\{(\mathring{X} - \bar{x})^2\}$

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

  ✓ a_k, ✓ m_k etc.

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

  ⇓ (distribution of r.v. corr. to statistics)

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

$(X_1, X_2, \ldots, X_n)$ where X_i corresponds to x_i

$\hookrightarrow$ (n-dimensional random variable) $i = 1, 2, \ldots, n$.

- Mathematically: $X : S \to \mathbb{R}$ (population r.v.)

Define: $X_i : S \times S \times \cdots \times S \to \mathbb{R}$ as

$S = \{1, 2, \ldots, 6\}$

$S \times S \times \cdots \times S$

$X_i(\omega_1, \omega_2, \ldots \omega_i, \ldots, \omega_n) = X(\omega_i) = x_i$ ← i-th coordinate of the random sample $(x_1, x_2, \ldots, x_n)$

r.v.s

$P(X_i \leq 2) = P(X \leq 2)$

- By def^n: $X_1, X_2, \ldots, X_n$ are mutually independent
- Each X_i has the same distribution as X:

$P(X_i \leq x) = P(X \leq x) \quad \forall x \in \mathbb{R}$

- **Sampling Distribution** ≡ Distribution of the random variable $S(X_1, X_2, \ldots, X_n)$ corresponding to a statistic.
  (Required to measure goodness of a statistic)
- **Important Statistics and Sampling Distributions:**
  - For a normal (m, σ) population,

✓ • $U = \dfrac{\bar{X} - m^*}{\sigma/\sqrt{n}} \sim N(0,1)$

✓ • $\chi^2 = \dfrac{nS^2}{*\sigma^2} \sim \chi^2(n-1)$

✓ • $t = \dfrac{\bar{X} - m^*}{S/\sqrt{n}} \sim t(n-1)$ where $S^2 = \dfrac{n}{n-1} S^2$

Estimation of Parameters

- Estimation of Parameter
  - Point Estimate
  - Interval Estimate
- Testing Hypothesis

Book:
S.K. Sen
S. De

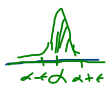
Problem: Let the functional form of the distribution function $F_X(x)$ of the population r.v. X is known but contains a number of unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$, i.e. population d.f. is $F_X(x; \theta_1, \theta_2, \ldots, \theta_k)$. To estimate these unknown parameters based on a random sample;

$(x_1, x_2, \ldots, x_n)$.

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

$\sigma > 0$

Estimate and Estimator



Let α be an unknown population parameter and $a(x_1, x_2, \ldots, x_n)$ 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, $P(\alpha - \epsilon < A < \alpha + \epsilon) \approx 1$

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

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

$A = a(X_1, X_2, \ldots, X_n)$ is called an estimator of 'α'.

Unbiased Estimate

Let $A = a(X_1, X_2, \ldots, X_n)$ 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.

S^2, s^2

Consistent Estimate:

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

If $A \xrightarrow[\text{in } P]{} \alpha$ as $n \to \infty$

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

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

PS-②

Prob ① Prove that sample mean is consistent and unbiased estimate of the population mean.

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

Sample mean $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$

r.v. of $\bar{x}$: $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ where X_is are i.i.d as X

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

Again $\bar{X} \xrightarrow[\text{inp}]{} m$ as $n \to \infty$ (using LLN) (check!).

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

(*) Prob: $A_k \xrightarrow{in\ P} \alpha_k$

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

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

r.v. corr. to the sample variance:

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

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

$S^2 = \frac{1}{n}\sum_{i=1}^{n}\{x_i^2 - 2x_i\bar{x} + \bar{x}^2\} = a_2 - \bar{x}^2$

$S^2 = A_2 - \bar{X}^2 \xrightarrow{in\ P} \alpha_2 - m^2 = \sigma^2$

$S^2 \xrightarrow{in\ P} \sigma^2$ (check!)

Prob: $A_k \xrightarrow{\text{in } P} \alpha_k$

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \quad A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

$$\alpha_k = E(X^k) = E(X_i^k), \quad i = 1, 2, \ldots, n.$$

Using T.I.

$$P\left(|A_k - E(A_k)| \geq \epsilon\right) \leq \frac{Var(A_k)}{\epsilon^2}, \text{ for any } \epsilon > 0.$$

$$Var(A_k) = E\left\{\left(A_k - E(A_k)\right)^2\right\}$$

$$= E\left\{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k - \alpha_k\right)^2\right\}$$

$$= \frac{1}{n^2} E\left\{\left(\sum_{i=1}^{n}\left(X_i^k - \alpha_k\right)\right)^2\right\} = \frac{1}{n^2} E\left\{\sum_{i=1}^{n}\left(X_i^k - \alpha_k\right)^2 + 2\sum_{1 \leq i < j \leq n}\left(X_i^k - \alpha_k\right)\left(X_j^k - \alpha_k\right)\right\}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} E\left\{\left(X_i^k - \alpha_k\right)^2\right\} + \frac{2}{n^2}\sum_{1 \leq i < j \leq n} E\left(X_i^k - \alpha_k\right) E\left(X_j^k - \alpha_k\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i^k\right) = \frac{n}{n^2} Var\left(X^k\right) = \frac{1}{n} Var\left(X^k\right) \to 0$$

$Var(X^k) = E\left\{\left(X^k - \alpha_k\right)^2\right\}$ is finite.

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

$$= \ldots \ldots$$

$$= \sigma^2 - E\left\{(\bar{X} - m)^2\right\}$$

$$E\left\{(\bar{X} - m)^2\right\} = E\left\{\left(\frac{X_1 + X_2 + \cdots + X_n}{n} - m\right)^2\right\} = \ldots = \frac{\sigma^2}{n}$$

(check!)

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

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

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

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

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

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

Q: How to compute an Estimate?

The method of maximum likelihood

(introduced by R.A. Fischer, 1912)

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

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

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

for the discrete case (i.e. if X is discrete)

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

$$= P(X = x_1)\, P(X = x_2) \ldots P(X = x_n)$$

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

For the continuous case:

$$P(x_1 < X_1 \le x_1 + dx_1, \ldots\ldots, x_n < X_n \le x_n + dx_n)$$

$$= P(x_1 < X_1 \le x_1 + dx_1) \ldots\ldots P(x_n < X_n \le x_n + dx_n)$$

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

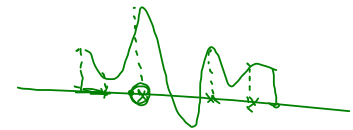
$$dx_1\, dx_2 \ldots dx_n$$

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

$$= f_X(x_1; \theta_1, \theta_2 \ldots, \theta_k) \ldots\ldots f_X(x_n, \theta_1, \theta_2, \ldots, \theta_k)$$

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

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



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

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

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

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

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

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

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