Estimation and Statistics

Population is an aggregate of objects under study. The population may be finite or infinite.

A finite subset of statistical individuals in a population is called a **sample**, and no. of individuals in a sample is called the **sample size**.

Random Sampling: the sample units are selected at random from population, and each unit of population has an equal chance of being included in it. (ie $^{N}C_{n}$ N → population, n → sample size)

Simple sampling: It is a random sampling in which each unit of the population has an equal chance (say p), and this probability is independent of the previous drawings.
ie a series of n independent trials with constant prob. 'p' of success of each trial.

Note that if an urn (box) contains 'a' white balls, 'b' black balls, the prob. of drawing a white ball at first drawn is $\frac{a}{a+b} = p_1$ say, if ball is not replaced, the prob. of getting white ball is: $\frac{a-1}{a+b-1} = p_2 \neq p_1$,

the sampling is not simple.

but in the first drawn, white ball has the same chance.

in the second draw, again white ball has same chance of being drawn., the sample is random.

Parameters and Statistic: In order to avoid verbal confusion with the statistical constants of the population i.e. mean (μ), variance (σ^2) etc. which are usually referred to as parameters, statistical measures computed from sample observations alone:

mean ($\bar{x}$), variance (s^2) etc.

called statistics.

Statistics are based on sample, varies.

(estimate of parameter)

random variable

Unbiased estimate: A statistic $t = t(x_1, \dots x_n)$, a function of sample values $x_1, x_2, \dots x_n$ is an unbiased estimate of population parameter θ if $E(t) = \theta$.

ie $E(\text{statistic}) = \text{parameter}$.

Null hypothesis (H_0): a definite statement about the population parameter.

Alternative hypothesis (H_1): Complementary to the null hypothesis.

ie If $H_0: \mu = \mu_0$,

then H_1: (i) $\mu \neq \mu_0$ (two tailed)
(Could be) (ii) $\mu > \mu_0$ (right tailed)
(iii) $\mu < \mu_0$. (Left tailed)

Errors:

← Type-I: Rejected H_0 when it is true.

← Type-II: Accepted H_0 when it is wrong.

we write:

$P[\text{Reject } H_0 \text{ when it is true}]$

$= P[\text{Reject } H_0 | H_0] = \alpha$

$P[\text{Accept } H_0 \text{ when it is wrong}]$

$= P[\text{Accept } H_0 | H_1] = \beta.$

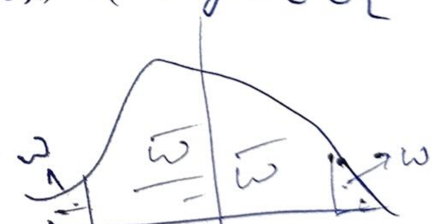
α, β are called sizes of type I, II errors resp.

ie $\alpha = P[\text{Reject a lot when it is good}]$

$\beta = P[\text{Accept a lot when it is bad}].$

Critical Region and Level of significance:

A region in the sample space S which amounts to rejection of H_0 is termed as critical region or region of rejection,



If w is the critical region and if $t = t(x_1, x_2, \ldots x_n)$ based on a sample of size n.

then $P(t \in w | H_0) = \alpha$

$P(t \in \bar{w} | H_1) = \beta.$

$\bar{w}$ is acceptance region

Note

Let, The probability 'α' that a random sample of the statistic t belongs to the critical region is known as the level of significance. ③

(prob. of type-I error)

Level of significance is always fixed in advance before collecting the sample information.

A test for testing the mean of the population:

$H_0: \mu = \mu_0$ against

$H_1: \mu > \mu_0$ (right tailed); the critical region lies entirely in the right tail of the sampling distribution of $\bar{x}$, (Similarly for left-tailed), is called single tailed test.

otherwise; if critical region lying in both the tails of the prob. curve of the test statistic, is known as two-tailed test.

Two tailed eg: 1) (A) bulbs with standard process (μ_1)

(B) bulbs with new technique (μ_2)

$H_0: \mu_1 = \mu_2$ (if they are same)

$H_1 = \mu_1 \neq \mu_2$.

one tailed.

2) If B is better than A.

ie H_0: $\mu_1 = \mu_2$: $H_1: \mu_1 < \mu_2$.

etc.

The significant values of $Z \sim N(0,1)$ at 5%ile and 1% level of significance for two tailed tests are 1.96 and 2.58 resp.

$$\because \quad P(|z| > 1.96) = 1 - 0.95 = 0.05$$

$$P(|z| > 2.58) = 1 - 0.99 = 0.01.$$

<u>For single tailed</u> $P(|z| > 1.64) = 0.05$

$$P(|z| > 2.33) = 0.01.$$

<u>Eg:</u> A dice is thrown 9000 times and a throw of 3 or 4 is observed 3240 times. Show that the dice can't be regarded as an unbiased one and find the limits b/w which the probability of a throw of 3 or 4 lies.

<u>Sol:</u> Let coming of 3 or 4 is success.

$$n = 9000 \qquad X = 3240$$

H_0: dice is unbiased.

<u>ie</u> $P = 1/6 + 1/6 = 1/3$. (3 or 4)

H_1: $P \neq 1/3$ (dice is biased)

[It is proved that for a large n: binomial distribution tends to normal ie

$$X \sim N(nP, nPQ)$$

ie $$Z = \frac{X - nP}{\sqrt{nPQ}} \sim N(0,1)$$]

ie $z = \dfrac{3240 - 9000 \times 1/3}{\sqrt{9000 \times 1/3 \times 2/3}} = 5.36$

ie $|z| > 5.36$; Ho is rejected.

ie dice is certainly biased

(ii) Now the dice is unbiased.

ie $\hat{p} = \dfrac{3240}{9000} = 0.36$. $\hat{Q} = 0.64$.

new prob of success.

Prob limit is (for normal variate):

$E(X) \pm 3\sqrt{V(X)}$. ie

ie $\hat{p} \pm 3\sqrt{\hat{P}\hat{Q}/n}$. in our case

Note) In a sample of size n;

X be the no of persons possessing the given attribute. Then

Observed proportion of success $= X/n = \hat{p}$.

$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} \cdot nP = P$.

$V(\hat{p}) = \frac{1}{n^2} V(X) = \frac{nPQ}{n^2} = \frac{PQ}{n}$

ie $0.36 \pm 3\sqrt{\dfrac{0.36 \times 0.64}{9000}} = 0.36 \pm 0.015$

ie 0.345 and 0.375 } Prob. of getting 3o lies in:

In particular: 95% Confidence limits for P:

$$\hat{p} \pm 1.96\sqrt{\hat{p}\hat{Q}/n}$$

$$99\% \longrightarrow \quad \hat{p} \pm 2.58\sqrt{\hat{p}\hat{Q}/n}.$$

Exer: In a sample of 1000 people in Maharashtra; 540 are rice eaters, rest are wheat eaters. Can we conclude both rice and wheat are equally popular in this state with 1% of level of significance?

Hint: $n = 1000$, $X = 540$; $\hat{p} = \frac{540}{1000} = 0.54$.

H_0: both are equally popular

i.e. $P = 0.5$.

H_1: $P \neq 1/2$. (two tailed)

$$Z = \frac{\hat{p} - P}{\sqrt{PQ/n}} = \frac{0.54 - 0.5}{0.0138}$$

$$= 2.532.$$

$t = \hat{p} \rightarrow$ r.v

$\hat{p} \sim N(E(\hat{p}), V(\hat{p}))$

$Z = \frac{\hat{p} - E(\hat{p})}{\sqrt{V(\hat{p})}} \sim N(0,1)$

The significant or critical value of Z at 1% level of significance is 2.58 (for two tailed)

$$Z = 2.532 < 2.58$$

It is not significant. i.e. H_0 is accepted.

We may conclude they are equally popular.

Note

Let $x_1, x_2 \ldots x_n$ be a random sample of size 'n' from a large population of size N with mean μ and variance σ^2.

Then $\bar{x} = \frac{\sum x_i}{n}$; $s^2 = \frac{1}{n}\sum(x_i-\bar{x})^2$

$E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{n\mu}{n} = \mu.$

ie $\bar{x}$ is unbiased estimator of μ.

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

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

Also $V(x_i) = E\left[(x_i - E(x_i))^2\right]$

$= E\left[(x_i-\mu)^2\right] = \sigma^2$

$= E(x_i^2) - \mu^2$

$\Rightarrow E[x_i^2] = V(x_i) + \mu^2$

$= \sigma^2 + \mu^2 \longrightarrow (1)$

$\therefore V(\bar{x}) = V\left(\frac{\sum x_i}{n}\right) = \frac{1}{n^2}\sum V(x_i)$

$= \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{\sigma^2}{n}.$

and $E(\bar{x}^2) = V(\bar{x}) + (E(\bar{x}))^2$ (similarly)

Test of Significance for Sample Mean :

Let x_i $(i=1,2,\ldots n)$ is a random sample of size n from a normal population with mean μ and variance σ^2;

$$\Rightarrow \quad \bar{x} \sim N(\mu, \sigma^2/n)$$

ie
$$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}.$$

Under the null hypothesis, H_0 that the sample has been drawn from a population with mean μ and variance σ^2.

ie there is no significant difference b/w sample mean and population mean

Confidence limits: with 95% confidence interval for μ:

$$|Z| \leq 1.96 \quad \text{ie} \quad \left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| \leq 1.96$$

$$\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

Since $V(\bar{x}) = \sigma^2/n$.

Standard Error $= \dfrac{\sigma}{\sqrt{n}}$.

x ——— x

Eg: A Sample of 900 members has a mean 3.4 cms, and s.d 2.61 cms.
Is the sample from a large population of mean 3.25 and s.d. 2.61 cms?

Population is normal and its mean is unknown. Find 95%, 98% limits for true mean.

Sol: H_0: the sample is from population with $\mu = 3.25$ cm; $\sigma = 2.61$ cms.

H_1: $\mu \neq 3.25$. (two-tailed)

under H_0,

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

ie

$$Z = \frac{3.4 - 3.25}{2.61/\sqrt{900}} = \frac{0.15 \times 30}{2.61}$$

$$= 1.73.$$

$|Z| < 1.96$; we H_0 may be accepted at 5% level of Significance

$$E(\bar{x})^2 = \frac{\sigma^2}{n} + \mu^2.$$

$$\therefore \quad E(s^2) = \frac{1}{n}\sum_{i=1}^{n}\left[(\sigma^2+\mu^2) - \left(\frac{\sigma^2}{n}+\mu^2\right)\right]$$

$$= \frac{1}{\cancel{n}}\cancel{n}\cdot(\sigma^2+\mu^2) - \frac{1}{n}\sigma^2 - \mu^2.$$

$$= (1-1/n)\,\sigma^2$$

ie $\quad E(s^2) = \left(\frac{n-1}{n}\right)\sigma^2, \Rightarrow$ not unbiased estimator

where $s^2 = \frac{1}{n}\sum(x_i-\bar{x})^2.$

Put $S^2 = \frac{1}{(n-1)}\sum(x_i-\bar{x})^2.$

$$\Rightarrow \quad (S^2) = \left(\frac{n}{n-1}\right)\cdot\frac{1}{n}\left(\sum(x_i-\bar{x})^2\right)$$

$$= \left(\frac{n}{n-1}\right)(s^2)$$

ie $\quad (n-1)(S^2) = n(s^2) \longrightarrow \star$

and $\quad E(S^2) = \left(\cancel{\frac{n}{n-1}}\right)\cdot\left(\cancel{\frac{n-1}{n}}\right)\sigma^2$

$$= \sigma^2.$$

ie S^2 is an unbiased estimator for σ^2.

<u>For $n\to\infty$:</u> $\quad s^2 = (1-1/n)S^2 = S^2$; Large <u>Sample</u>

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 3.5705 \text{ and } 3.2295$$

98%

$$\bar{x} \pm 2.33 \frac{\sigma}{\sqrt{n}} = 3.6027 \text{ and } 3.1973.$$

Eg: A normal population has a mean of 0.1 and s.d of 2.1. find the prob. that the mean of sample of size 900 will be negative.

Sol: $X \sim N(\mu, \sigma^2)$.

Sample mean $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$.

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 0.1}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - 0.1}{0.07}.$$

ie $\bar{x} = 0.1 + 0.07Z$; where $Z \sim N(0,1)$

Now let $p = P(\bar{x} < 0) = P(0.1 + 0.07Z < 0)$

$$= P\left(Z < \frac{-0.1}{0.07}\right) = P(Z < -1.43)$$

$$= P(Z > 1.43)$$

$$= 0.5 - P(0 < Z < 1.43)$$

$$= 0.5 - 0.4236$$

$$= 0.0764.$$

Eg: The guaranteed average life of a certain type of bulb is 1000 hrs. with sd → 125hrs. It is decided to sample the output so as to ensure that 90% of the bulbs do not fall short of guaranteed average by more than 2.5%. What must be the minimum size of the sample?

Sol: $\mu = 1000$; $\sigma = 125$.

We should have:

$$\bar{x} > 1000 - 2.5\% \text{ of } 1000 = 975$$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

ie $$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} > \frac{975 - 1000}{\frac{125}{\sqrt{n}}} = -\frac{\sqrt{n}}{5}$$

ie $$P\left(Z > -\frac{\sqrt{n}}{5}\right) = 0.90$$

ie $$P\left(0 < Z < \frac{\sqrt{n}}{5}\right) = 0.40$$

ie $$\frac{\sqrt{n}}{5} = 1.28 \text{ (by table)}$$

$$\Rightarrow n = 25 \times (1.28)^2 = 41 \text{ (approx)}.$$

Estimator: Consider a general family of distributions;

$$\{ f(x, \theta_1, \theta_2, \ldots \theta_k) : \theta_i \in \Theta ; i = 1, 2, \ldots k \}$$

In Particular; the parameter space

$$\Theta = \{ (\mu, \sigma^2) ; -\infty < \mu < \infty, 0 < \sigma < \infty \} \text{ if } X \sim N(\mu, \sigma^2).$$

For a sample $x_1, x_2, \ldots x_n$ of size n from a population with pdf $f(x, \theta_1, \theta_2, \ldots \theta_k)$; where θ_i's are unknown population parameter, there are infinite statistics which may be proposed as estimator of one or more parameters.

We wish to determine

$$T_1 = \hat{\theta}_1 (x_1, \ldots x_n), \; T_2 = \hat{\theta}_2 (x_1, \ldots x_n) \text{ etc}$$

s.t their distribution is concentrated as closely as possible near the true value of the parameter.

Characteristic of Estimator:

(i) Consistency (ii) Unbiasedness

(iii) Efficiency.

1) An estimator $T_n = T(x_1, \ldots x_n)$ is said to be Consistent estimator of $\gamma(\theta)$; $\theta \in \Theta$; if

T_n Converges to $\gamma(\theta)$ in probability.

ie if $T_n \xrightarrow{p} \gamma(\theta)$ as $n \to \infty$.

ie $\forall \epsilon > 0$; $\exists \eta > 0$ st $\forall n \geq m(\epsilon, \eta)$ ↑ depends

st

$$P[|T_n - \gamma(\theta)| < \epsilon] \longrightarrow 1 \quad \text{as } n \to \infty$$

OR

$$P[|T_n - \gamma(\theta)| < \epsilon] > 1 - \eta \quad \forall n \geq m(\text{large}).$$

Eg:

$$\overline{X}_n = \frac{\sum X_i}{n} \xrightarrow{p} E(X_i) = \mu \quad \text{as } n \to \infty$$

ie sample mean is always a consistent estimator of the population mean μ.

Note: ~~For~~ a given Consistent estimator T_n; of $\gamma(\theta)$

$$T_n' = \left(\frac{n-a}{n-b}\right) T_n = \left(\frac{1-a/n}{1-b/n}\right) T_n \xrightarrow{p} \gamma(\theta) \text{ as } n \to \infty$$

infinite such estimator can be Constructed.

(for different a, b)

2) On the other hand, unbiasedness is associated with finite n.

i.e. $T_n = T(x_1, x_2 - x_n)$ is unbiased estimator for $\gamma(\theta)$ if $E(T_n) = \gamma(\theta)$ $\forall$ $\theta \in \Theta$

eg

$E(\bar{x}) = \mu$ but $E(s^2) \neq \sigma^2$ and $E(S^2) = \sigma^2$.

Example: Show that $t = 1/n \sum x_i^2$; where $x_1, - x_n$ is a random sample from $N(\mu, 1)$, is an unbiased estimator of $\mu^2 + 1$.

Sol:

$$E(x_i) = \mu; \quad V(x_i) = 1.$$

$$E(x_i^2) = V(x_i) + (E(x_i))^2 = 1 + \mu^2$$

$$\therefore E(t) = \frac{1}{n} \sum E(x_i)^2 = \frac{1}{n}(n \cdot (1+\mu^2)) = 1 + \mu^2.$$

E2 If T is unbiased estimator for θ, then T^2 is a biased estimator for θ^2.

Hint

$$E(T) = \theta$$

$$Var(T) = E(T^2) - E(T))^2$$
$$= E(T^2) - \theta^2$$

i.e. $E(T^2) = Var(T) + \theta^2$. $(Var\,T > 0)$

$$\neq \theta^2.$$

Exer: Show that $\frac{\sum x_i(\sum x_i - 1)}{n(n-1)}$ is an unbiased estimator of θ^2; $x_1, x_2 - x_n$ drawn on X which takes 0 or 1 with prob $(1-\theta)$ and θ.

Sol: $T = \sum_{i=1}^{n} x_i \sim B(n,\theta)$ [Bernoulli population]

$E(T) = n\theta$; $var(T) = n\theta(1-\theta)$

Then

$$E\left[\frac{T(T-1)}{n(n-1)}\right] = \theta^2 \text{ (check it)}$$

Sufficient Condition:

Thm: Let $\{T_n\}$ be a sequence of estimators s.t $\forall\ \theta \in \Theta$;

(i) $E_\theta(T_n) \longrightarrow \gamma(\theta)$ as $n \to \infty$

(ii) $Var_\theta(T_n) \longrightarrow 0$ as $n \to \infty$

Then T_n is a Consistent estimator of $\gamma(\theta)$.

Eg: $\bar{x} = \frac{\sum x_i}{n}$; from $N(\mu, \sigma^2)$.

$E(\bar{x}) = \mu$; $V(\bar{x}) = \sigma^2/n \longrightarrow 0$ as $n \to \infty$

$\Rightarrow \bar{x}$ is a Consistent estimator for μ.

Exer: Show that $x_1, x_2, \ldots x_n$ are random observation on a Bernoulli variate X taking the value 1 with prob 'p' and the value 0 with prob $(1-p)$ show that:

$$\frac{\sum x_i}{n}\left(1-\frac{\sum x_i}{n}\right)$$ is a Consistent estimator for $p(1-p)$.

Def: (Efficiency) If T_1 is the most efficient estimator with variance V_1; T_2 is any other estimator with variance V_2; then the efficiency E of T_2 is defined as:

$$E = \frac{V_1}{V_2}$$; (E Can't exceed unity.)

Eg: A random sample of size 5; from normal population with unknown μ.

Consider $T_1 = \frac{X_1 + \ldots X_5}{5}$; $T_2 = \frac{X_1 + X_2}{2} + X_3$

$$T_3 = \frac{2X_1 + X_2 + \lambda X_3}{3}.$$

(i) find λ, s.t T_3 is unbiased estimator of μ.

(ii) who is best among T_1, T_2, T_3.

Sol: $E(T_1) = \mu \Rightarrow T_1$ is unbiased estimator of μ

$E(T_2) = 2\mu \Rightarrow T_2$ is not unbiased.

$E(T_3) = \mu \Rightarrow \lambda = 0.$

$V(T_1) = \frac{1}{5}\sigma^2 \qquad V(T_3) = \frac{5}{9}\sigma^2.$

$V(T_2) = \frac{3}{2}\sigma^2$

$V(T_1)$ is least ; T_1 is best estimator (in the sense of efficiency) of μ.

Maximum likelihood Estimator.

Def: [Likelihood Function]

Let $x_1, x_2 \ldots x_n$ be a random sample of size 'n' from a population with density function $f(x, \theta)$. Then the likelihood func of the sample values is given by:

$$L = L(\theta) = f(x_1, \theta) \ldots f(x_n, \theta).$$
$$= \prod_{i=1}^{n} f(x_i, \theta)$$

we wish to find $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_n)$: $\theta = (\theta_1, \theta_2$

st $L(\hat{\theta}) > L(\theta) \ \forall \ \theta$ in Θ.

ie $L(\hat{\theta}) = \sup L(\theta) ; \ \theta \in \Theta.$

$\hookrightarrow$ Maximum likelihood Estimator.

Eg! In a random sampling from $N(\mu, \sigma^2)$ find MLE for

(i) μ, when σ^2 is known

(ii) σ^2, when μ is known

(iii) Simultaneous of μ and σ^2.

Sol! $X \sim N(\mu, \sigma^2)$.

$$L = \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \right)$$

$$\log L = -\frac{1}{2} \cdot n \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2$$

Case (i) when σ^2 is known.

$$\frac{\partial}{\partial \mu}(\log L) = 0$$

$$\Rightarrow \frac{-1}{2\sigma^2} \sum 2(x_i-\mu)(-1) = 0$$

$$\Rightarrow \sum(x_i-\mu) = 0$$

$$\Rightarrow \sum x_i = n\mu$$

i.e. $\hat{\mu} = \frac{\sum x_i}{n} = \bar{x}$

Hence MLE for μ is $\bar{x}$.

Case-2. μ is known.

$$\frac{\partial}{\partial\sigma^2}(\log L) = 0 \Rightarrow -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum(x_i-\mu)^2 = 0$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum(x_i-\mu)^2.$$

Case-3 $\frac{\partial}{\partial\mu}(\log L) = 0$ and $\frac{\partial}{\partial\sigma^2}(\log L) = 0$

ie $\hat{\mu} = \bar{x}$ $\therefore$ $\hat{\sigma}^2 = \frac{1}{n}\sum(x_i-\hat{\mu})^2$

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

Sample variance

Exer: Prove that MLE of parameter 'α' of population with pdf: $\frac{2}{\alpha^2}(\alpha-x)$; $0<x<\alpha$. for a sample of Unit size is $2x$; x being the sample value. Show also estimator is biased.

Hint: $\frac{d}{d\alpha}(\log L) = 0 \Rightarrow \alpha = 2x.$

ie $\hat{\alpha} = 2x.$

$$E(\hat{\alpha}) = \frac{2}{3}\alpha \neq \alpha.$$