Sampling Distributions & Sampling Theory

(i). **Population**:

The entire set of elements for which the experiments and estimations are to be done.

(ii) **Sample**:

A subset of population. All the estimations about the population are made based on the observations of the sample. (also called sample set)

Eg $P = \{a, b, c, d, e, f\}$

Then $S_1 = \{a, b, c\}$

$S_2 = \{a, d, e\}$

$S_3 = \{b, e, f, d\}$

are examples of samples.

(iii). **The population distribution and parameters**:

The population is considered to a random variable. The distribution of this R.V. is the distribution of the population. The parameters of this distribution are called the parameters of the population.

Examples

1. Let $P = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
Let $S = \{1, 6, 9, 10\}$ be a sample

Now

Population mean $= (1+2+3+4+5+6+7+8+9+10)/10$

$= \frac{55}{10} = 5.5$

Sample mean $= \frac{1+6+9+10}{4}$

$= \frac{27}{4} = 6.75$.

Clearly S.M. $\neq$ P.M. (This estimation is however close to P.M.)

Let $S_1 = \{1, 10\}$

$SM = 11/2 = 5.5$!!

But usually we do not know the population mean. Thus we dont know which sample mean is correct ~~esm~~ estimate of P.M !

The objective of studying sampling theory is to estimate characteristics of the population based on the sample values. Also we need to make formal definitions about the "correctness" of the estimate as we dont know the characteristics of the population.

Notes: (i) Different samples from the same population will result in different estimates ~~esto~~ of the parameters. These estimates are ~~themselves~~ themselves treated as R.V. and hence follow some probability distributions. Such distributions are called sampling distributions.

(ii). If the distribution of the population is known then only the parameters of this distribution is to be estimated from the observed sample. The confidence in such estimates is to be evaluated.

(iii) Hypothesis testing & Goodness of fit

Assume that the distribution of the population is not known. Then we may try to fit some known family of distribution f^n. This fit needs to be evaluated for correctness. This evaluation is called goodness of fit.

However in some cases we may only be interested in testing a hypothesis regarding some property of the population. Based on the observed sample and tests of the sample we may accept or reject the hypothesis.

Examples:

(1). Suppose we want to evaluate an algorithm for execution time. Naturally we can't give all possible input (population). So we evaluate on a sample and calculate $E(T)$ and $Var(T)$, T = execution time. Here the

execution time, T, is a R.V. Hence E(T) and V(T) make sense.

(2) Jobs come at a call center. The interarrival time is, say, exponentially distributed with parameter λ which is unknown.

Here we know the type of distribution but not the parameter λ. We now estimate λ. For this, we take some samples. (i.e. observe the interarrival time for a finite period). Then we obtain an estimate of λ (w/a $\hat{\lambda}$).

Sometimes we are not interested in λ (or $\hat{\lambda}$) but we are only interested if λ is greater than or some threshold value. i.e

$$\lambda = \lambda_0$$
$$\lambda > \lambda_0$$
or $$\lambda < \lambda_0.$$

If we make a hypothesis that

$H_1: \lambda > \lambda_0$ ($\Rightarrow$ No more resources needed)

$H_2: \lambda < \lambda_0$

$\Rightarrow$ (more resources needed)

If H_1 is accepted then resources sufficient so we proceed to test the hypotheses H_1 & H_2 for acceptance or rejection.

Parameter Estimation

Let P be the population. Suppose the probability distribution, f(x), of P is known. Only the parameter, θ, of X is not known. This parameter can be:

θ = Mean of population i.e. E(x)
or θ = Variance " " i.e. V(x).

Procedure:

Collect 'n' experimental outcomes:

$x_1, x_2, \ldots, x_n$.

Each experimental outcome, x_i, is a value of a r.v. X_i. Then the set of random variables

$X_1, X_2, \ldots, X_n$ of 'X'

is called a sample of size 'n'

Note that $x_1, x_2, \ldots x_n$ is (a) sample and this sample can have so many values from the set P.

Ex Suppose $P = \{1, 2, 3, \ldots, N\}$.
we can have

~~$x_1 = 1$, or 2 or 3, ... or N~~
~~x_2~~

X_1 can take value = $1, 2, \ldots, N = x_1$
X_2 " " " = $1, 2, \ldots N = x_2$
X_n : : = $1, 2, \ldots N = x_n$

Random Sample

The set of random variables $X_1, X_2, \ldots X_n$ is called a random sample of size n from the population, P, with distribution of the R.V X and $F(x)$. provided that $X_1, X_2, \ldots X_n$ are pairwise independent with same distribution f^n i.e

$$F_i(x) = F(x) \quad \forall\ i \text{ & } x.$$

This holds if the samples are drawn with replacement. If replacement is not done the outcome of R.V X_2 will be affected by the outcome of X_1 & so on. In that case

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

$$= \frac{1}{N} \cdot \frac{1}{(N-1)} \cdots \frac{1}{N-n+1}$$

$$= \frac{(N-n)!}{N!}$$

Statistic

Any function $T(X_1, X_2, \ldots X_n)$ of the observations $x_1, x_2, \ldots, x_n$ is called Statistic Since it is a f^n of r.v, a statistic is itself a R.V. and its distribution is called the sampling distribution of T.

(i) Sample Mean:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Although x is the distribution RV of the population we write $\bar{x}$ as sample mean to understand that it represents joint DF of $(x_1, x_2, \ldots, x_n)$

(ii) Sample Variance:

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

Estimator:

Let θ be a parameter of the population. $g(t)$ is statistic (i.e. fn of RV x) is used to estimate θ, called $\hat{\theta}$. is called estimator of θ.

i.e. $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$

is called the estimator of θ

A specific value of the estimator $\hat{\theta}$ is called an estimate of θ.

Unbiased Estimator

A statistic $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ is called an unbiased estimator of θ if

$$\boxed{E[\hat{\theta}(x_1, x_2, \ldots, x_n)] = \theta}$$

Estimate

Let $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ be an estimator of θ. A specific value of $\hat{\theta}$ obtained by substituting observations of $x_1, x_2, \ldots, x_n$ is called estimate of θ.

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

Ex

The sample mean is an unbiased estimator of the population mean μ.

For $E[\bar{X}] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right]$

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

$$= \frac{1}{n} \left[E[X_1] + E[X_2] + \cdots + E[X_n]\right] \quad (*)$$

We know that DF of each X_i is the DF of X i.e.

$F_{X_i}(x) = F(x) \quad \forall\ i \text{ & } x$

$\therefore E[X_i] = E[X]$

$\therefore E[\bar{X}] = \frac{1}{n} [\mu + \mu + \cdots + \mu]$

$= \frac{1}{n} [n\mu]$

$= \mu$

$\therefore$ The estimator is an unbiased estimator of population mean.

Computing Variance of the estimator sample mean:

$$V[\bar{X}] = \sum_{i=1}^{n} \text{Var}\left[\frac{X_i}{n}\right]$$

Note (*) Consider Rolling of die 3 times

$X_1 \equiv$ first outcome, $X_2 \equiv 2^{nd}$ & $X_3 = 3^{rd}$ outcome

$$E(\bar{X}) = \frac{1}{n} E\left[\sum_{i=1}^{3} X_i\right] = \frac{1}{n} \sum_{i=1}^{3} E[X_i]$$

$$= \frac{1}{n} \left[E[X_1] + E[X_2] + E[X_3]\right]$$

$$= \frac{1}{n} \left[3.5 + 3.5 + 3.5\right] = 3.5 = \mu.$$

$$= Var\left[\sum_{i=1}^{n} \frac{x_i}{n}\right]$$

We know for independent RV x_1 & x_2

$$Var[x_1 + x_2] = Var(x_1) + Var(x_2)$$

$$\therefore V[\bar{x}] = V\left[\sum_{i=1}^{n} \frac{x_i}{n}\right]$$

$$= V\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]$$

$$= \frac{1}{n} V\left[\sum_{i=1}^{n} x_i\right]$$

$$= \frac{1}{n}\left[V(x_1) + V(x_2) + \cdots + V(x_n)\right]$$

$$= \frac{n}{n^2}\left[\sum_{i=1}^{n} V[x_i]\right]$$

For independent RV x_1 & x_2, we have

$$V[a_1 x_1 + a_2 x_2] = a_1^2 V[x_1] + a_2^2 V[x_2]$$

Generalizing and applying here

$$V[\bar{x}] = V\left[\sum_{i=1}^{n} \frac{x_i}{n}\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} V[x_i]$$

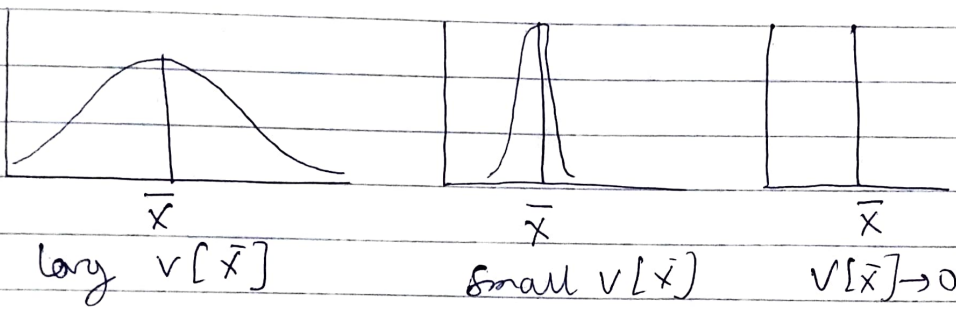
$$= \frac{1}{n^2}\sum_{i=1}^{n} V[x]$$

$$= \frac{1}{n^2} [n\sigma^2]$$

$$\Rightarrow \boxed{V[\bar{X}] = \sigma^2/n}$$

This means that $\bar{X}$ as estimate of μ will have less variance when n is large. Symbolically

$$\lim_{n \to \infty} V[\bar{X}] = 0.$$



Larg $V[\bar{X}]$ Small $V[\bar{X}]$ $V[\bar{X}] \to 0$

∴ As $n \to \infty$, The accuracy of $\bar{X}$ as an estimate of population mean is excellent.

Ex

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

is an unbiased estimator of the population variance.

Note: Be careful that $E[\bar{X}] = \mu$ and not $\bar{X} = \mu$. This means taking many samples & finding $\bar{X}$ will lead to different $\bar{X}$ however, the average of all $\bar{X}$ will be μ.

Proof:

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

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

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

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

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

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

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

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

---------- (2)

But $E[x_i^2] = V[x_i] + (E[x_i])^2$

$= \sigma^2 + \mu^2$ ---------- (3)

& $E[\bar{x}^2] = V[\bar{x}] + (E[\bar{x}])^2$

$= \frac{\sigma^2}{n} + \mu^2$ ——— (4)

Substituting (3) & (4) in (2)

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

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

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

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

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

$= \sigma^2$

$\Rightarrow \boxed{E[s^2] = \sigma^2}$

$\therefore s^2$ is an unbiased estimator of the population variance σ^2.

~~Ex~~ Ex

Let the population be "heights of 20 students in a class" normally distributed with

$$\mu = 175 \text{ cm}$$

& $\sigma = 5 \text{ cm.} \Rightarrow \sigma^2 = 25$

$P = \{163, 169.5, 159.14, \ldots, 186.41, 178.99, 185.2\}$
$\quad\quad x_1 \quad\quad x_2 \quad\quad x_3 \quad\quad\quad\quad x_{18} \quad\quad x_{19} \quad x_{20}$

Let us calculate a sample mean $\bar{x}$ with sample size = 3.

Let $S_1 = \{x_1, x_2, x_3\}$

$S_2 = \{x_{18}, x_{19}, x_{20}\}$

There are 6840 samples is possible.

Sample mean for $S_1 = \bar{x}_1 = \dfrac{163 + 169.5 + 159.14}{3}$

$$\Rightarrow \boxed{\bar{x}_1 = 160.5 \text{ cm.}}$$

Sample mean for $S_2 = \bar{x}_2 = \dfrac{186.41 + 178.99 + 185.2}{3}$

$$\Rightarrow \boxed{\bar{x}_2 = 183.5 \text{ cm}}$$

Sampling distribution of s.m

$\underbrace{\bar{x}_1, \bar{x}_2, \ldots \bar{x}_i}_{\bar{X}}$

Number of samples ≠ Sample size.

We know ~~for normal distribution~~

$E[\bar{x}] = \mu$ & ~~$E(\bar{x}) = \frac{\sigma^2}{n}$~~

$Var[\bar{x}] = \sigma^2/n$

If the population is $N(\mu, \sigma^2)$

then distribution of $\bar{x}$ is $N(\mu, \sigma^2/n)$

If x represents height of class then

$\mu_{\bar{x}} = 175$ cm & $\sigma_{\bar{x}} = \frac{5}{\sqrt{5}} = 2.9$ cm

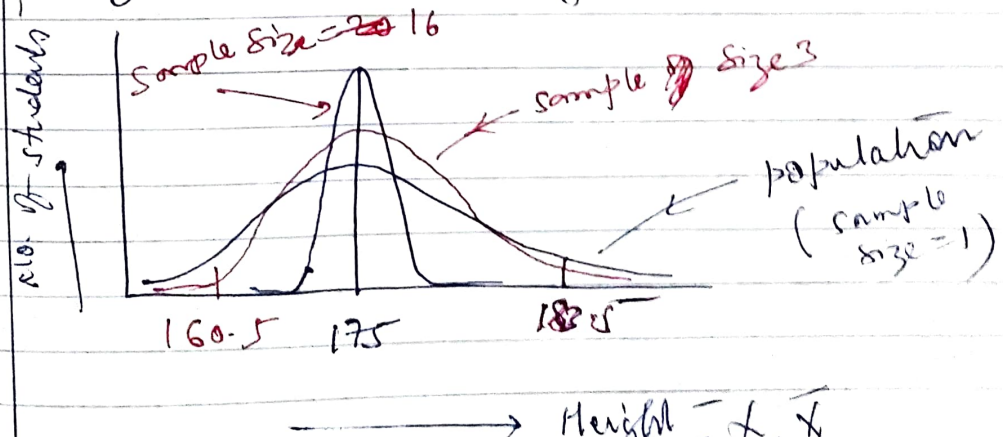
$\therefore \bar{x}$ has distribution ~~of~~ $N(\mu, 25/5)$

Now we see Sample mean variance $= 25/5 = \sigma^2/n$

As sample size increases $\& n \to 20$

then ~~samp~~ $V[\bar{x}] \to 0$ i.e. $\mu_{\bar{x}} = \mu = 175$

which is obvious because there is only one sample of size 20.



Note Population distribution is a special case of sampling dist with $n = 1$