Sampling

1. Our aim is to draw a genuine conclusion about a large group of objects/individuals (population). Instead of examining the entire group (may be difficult/impossible) we only inspect only a small part of this population, called a sample. The process of obtaining sample is sampling.

i) Suppose we would like to draw conclusions about the heights of 10K students (popl^n) be examining only, say 100 students (a sample) selected from the population.

ii) We may wish to draw conclusions about the colours of 1K marbles (popl^n) in a bag, (by sample) selecting a sample of 20 marbles from the bag (each marble selected is return-ed after its colour is observed → WR)

2. i) "Population" → used to denote the observations/measurements (rather than individuals/objects)

ii) Population → finite / infinite. The no. is called the size of population (N). The no. in the sample (n) is called the sample size.

iii) N = 10k, 1k. n = 100, 20. (above)

iv) Suppose we would like to draw conclusions about the fairness (unbiased) of a specific coin by tossing it repeatedly.

Population → all possible tosses of the coin (N: infinite)

A sample → may be obtained by examining, say 70 tosses of the coin and noting the % of heads/tails (n: 70)

3. Sampling → WR, WOR

i) After drawing an object from a bag we may or may not replace same into the bag before we draw again.

a particular object comes up repeatedly (WR)

the object comes up once (WOR)

ii) A finite pop^n that is sampled with replacement can theoretically be considered infinite since samples of any size can be drawn without exhausting the population.

4. How to choose a sample?

One way to do this for finite populations is to make sure that each member of the pop^n has equal chance of being included in the sample, which is then called a random sample.

As inference from sample to pop^n can't be certain, we use the language of probability in any statement of conclusions.

5. Parameter (of population)

Population $\xrightarrow{\text{known}}$ if its prob dist^n $f(x)$ of the associated r.v. X is known.

If $X \sim$ binomial dist^n we say that the pop^n is binomially distributed or that we have a binomial pop^n.

There will be certain quantities that appear in $f(x)$, e.g. μ, σ (for normal) & p, n (for binomial). Other quantities moments, skewness etc. can be determined in terms of μ, σ etc. All such quantities are often called population parameters.

For a given/known population $f(x)$, population parameters are also known.

If prob. distn $f(x)$ of the population is not known precisely, although we may have some idea of, or at least be able to make some hypothesis concerning, the general behaviour of $f(x)$.

Ex: We may have some reason to suppose that a particular population is normally distributed. In that case we may not know one or both of the values μ and σ and so we might wish to draw statistical inferences about them.

6. Statistics (of sample)

X : values are various heights

X_1 : height of the 1st individual, x_1 be its value

X_2 : height of the 2nd individual, x_2 be its value.

In general, a sample of size n described by the values $x_1, x_2, \ldots, x_n$ of the random variables $X_1, X_2, \ldots, X_n$

SRSWR : $X_1, X_2, \ldots X_n \rightarrow$ iid having prob. distn $f(x)$. Their joint pdf

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

Sample statistic or Statistic

Any quantity obtained from a sample for the purpose of estimating a popl$\frac{n}{-}$ parameter.

A statistic for a sample of size n can be defined as a function of r.v. $x_1, x_2, \dots, x_n$

i.e., $g(X_1, X_2, \dots, X_n) \longrightarrow$ another random variable

Its values are represented by $g(x_1, x_2, \dots, x_n)$.

Corresponding to each popl$\frac{n}{-}$ parameter $\exists$ a statistic to be computed from the sample.

One of the imp. probs of sampling This is to decide how to form the statistic that will best estimate a given popl$\frac{n}{-}$ parameter.

popl$\frac{n}{-}$ parameters : μ, σ

sample statistics : m, s etc.

7. Sampling Distribution

The prob. dist$\frac{n}{-}$ of a sample statistic $g(X_1, X_2, \dots, X_n)$ is called the sampling dist$\frac{n}{-}$ of the statistic.

Alt. Consider all possible samples of size n to be drawn from the popl$\frac{n}{-}$. For each sample we compute the statistic. In this way, we obtain the dist$\frac{n}{-}$ of the statistic ——— called sampling dist$\frac{n}{-}$.

Mainly, we are interested to compute sampling dist$\frac{n}{-}$ of sample mean, sample variance etc.

8. Sampling Dist$\frac{n}{-}$ of Sample Mean

$X_1, X_2, \dots X_n \rightarrow$ iid r.v. for a random sample of size n.

$$\bar{X} = \frac{1}{n}\sum X_i \rightarrow \text{sample mean.}$$

If $x_1, x_2, \dots, x_n$ denote values obtained in a particular sample of size n, then the mean for that sample $\bar{x} = \frac{1}{n}\sum_1^n x_i$.

Let $f(x)$: prob distⁿ with mean μ and variance σ^2 of some given popl$\frac{n}{\cdot\cdot}$, from which we draw a sample of size n.

i) the mean of the sampling distⁿ of means, denoted by $\mu_{\bar{x}} = E(\bar{X}) = \mu$, the popl$\frac{n}{\cdot\cdot}$ mean

ii) If the popl$\frac{n}{\cdot\cdot}$ is infinite and the sampling is random or if the popl$\frac{n}{\cdot\cdot}$ is finite and sampling is with replacement,

the variance of the sampling distⁿ of means, denoted by $\sigma_{\bar{X}}^2 = E[\bar{X}-\mu]^2 = \frac{\sigma^2}{n}$,

σ : the popl$\frac{n}{\cdot\cdot}$ variance.

iii) If the popl$\frac{n}{\cdot\cdot}$ is finite (N) and the sampling is without replacement (WOR), then

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \longrightarrow \frac{\sigma^2}{n}, \text{ as } N \to \infty.$$

iv) If the popl$\frac{n}{\cdot\cdot}$ $X \sim N(\mu, \sigma^2)$, then the sample mean $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

v) If the popl$\frac{n}{\cdot\cdot}$ from which samples are taken, which has a prob. distⁿ with mean μ and variance σ^2 (not necessarily normal distribution), then

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \text{ is asymptotically normal,}$$

i.e., $$\lim_{n\to\infty} P(Z \le z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-u^2/2}\,du$$

(Popl$\frac{n}{\cdot\cdot}$ is infinite / sampling is WR)

9. <u>Sampling Dist$\frac{n}{..}$ of Proportions</u>

Popl$\frac{n}{..}$ $\longrightarrow$ infinite & binomially distributed
p: prob. that any given member exhibits

Consider all possible samples of size 'n' drawn from this popl$\frac{n}{..}$ and for each sample, we determine the <u>statistic</u> that is <u>the proportion P of success</u>

P: proportion of heads turning up in n tosses

Sampling dist$\frac{n}{..}$ of proportions (whose mean μ_P and s.d. σ_P) as

$$\mu_P = p, \quad \sigma_P = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\left[\begin{array}{l} E(\bar{x}) = \mu, \ Var(\bar{x}) = \sigma^2/n \\ \mu_P \text{ \& } \sigma_P \text{ are obtained by putting} \\ \mu = p \text{ and } \sigma = \sqrt{pq} \end{array}\right]$$

For large n (≥ 30), the sampling dist$\frac{n}{..}$ is close to normal dist$\frac{n}{..}$.

For finite popl$\frac{n}{..}$ (sampling is WOR), $\sigma_P \leftarrow \sigma_{\bar{x}}$

9. Sampling Distⁿ of Differences & Sums

μ_1, σ_1 (P1) $\longrightarrow$ sample of size n_1 $\longrightarrow$ S_1 (compute statistic S_1 for each sample)

μ_2, σ_2 (P2) $\longrightarrow$ sample of size n_2 $\longrightarrow$ S_2 (compute for each sample of size n_2)

Let mean & s.d. of S_1 be μ_{S_1} and σ_{S_1} ; μ_{S_2} , σ_{S_2}

Distⁿ of $(S_1 - S_2)$ $\rightarrow$ sampling distⁿ of differences of the statistics

Mean : $\mu_{S_1 - S_2} = \mu_{S_1} - \mu_{S_2}$

S.D. : $\sigma_{S_1 - S_2} = \sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2}$ (S_1, S_2 are indep)

If in particular S_1 & S_2 be sample means of two populations, denoted by $\bar{X}_1$, $\bar{X}_2$ resp, then

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

$$\sigma_{\bar{X}_1} - \sigma_{\bar{X}_2} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

(sampling distⁿ of difference of means)

Results hold also for finite popⁿ, if sampling is WR.

The standardized variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\sim$ Normally distributed if n_1, n_2 be large ($n_1, n_2 \geq 30$)