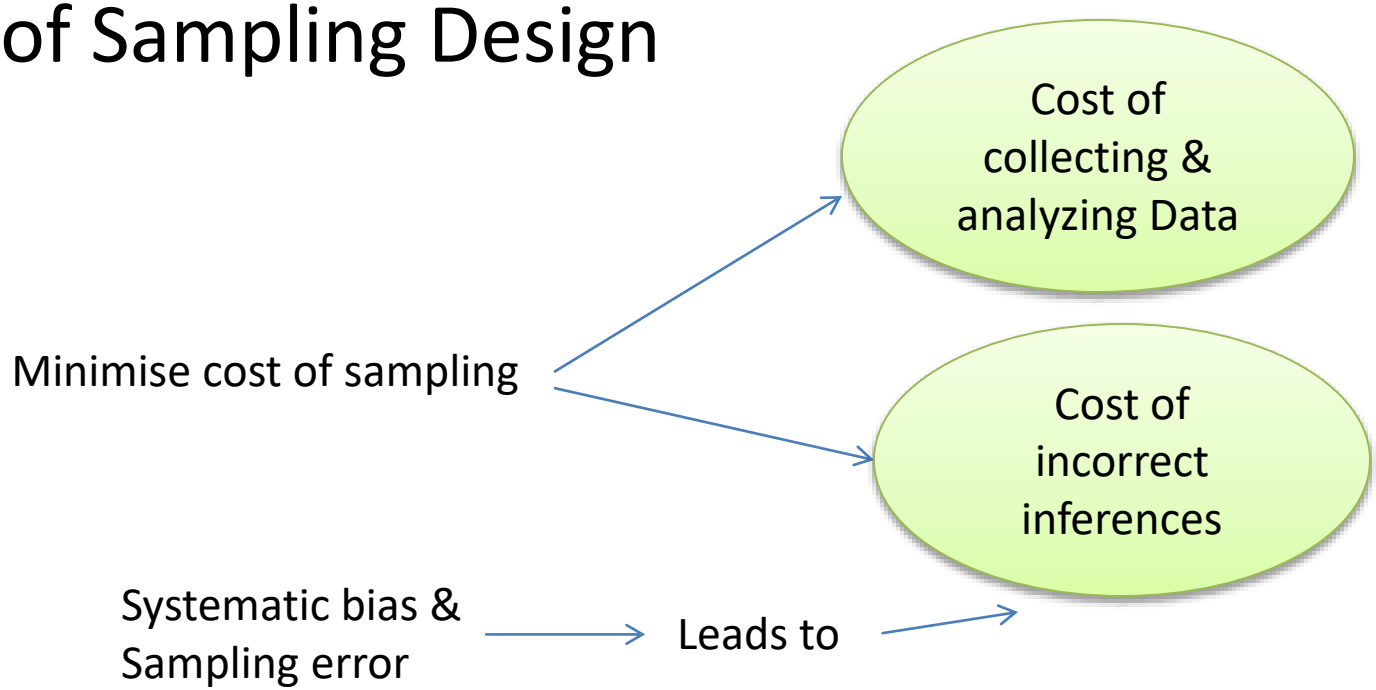
SAMPLING & ESTIMATION

Main Issues

- Universe/Population
- Sampling Frame
- Sampling Unit
- Sample Size
- Budgetary Constraints
- Sampling Procedure

- Universe/Population
  - CENSUS STUDY
- Sample
  - Sampling Unit
  - Sampling Frame: representation of the elements of the target population. Examples of a sampling frame include the telephone book, an association directory listing the firms in an industry, a customer database, a mailing list on a database purchased from a commercial organisation, a city directory, or a map. If a list cannot be compiled, then at least some directions for identifying the target population should be specified, such as random-digit dialling procedures in telephone surveys.
  - Sample Size
- Budgetary Constraints
- Sampling Procedure

Criteria of Sampling Design



Systematic bias – Inherent in the System
Design Errors: Selection error, Sampling frame error, Measurement scale error
Administering Error: Questioning error, Recording error
Response Error: Data error (intentional/ unintentional)
Non response Error: Failure to contact all members, Incomplete responses

Random/Sampling error – Random variation, controllable by sample size
difference between measure obtained from the sample and the true measure of the population

Sampling Methods

A. Non-random/Non-probability-based sampling: relies on the personal judgement of the researcher rather than on chance to select sample elements.

- **Convenience sampling**: selection of sampling units is left primarily to the interviewer. Often, respondents are selected because they happen to be in the right place at the right time. Examples: (1) use of students and members of social organisations, (2) street interviews without qualifying the respondents, (3) some forms of email and Internet survey, (4) tear-out questionnaires included in a newspaper or magazine.
- **Judgmental sampling**: elements are selected based on the judgement of the researcher because he/she believes that they are representative of the population of interest or are otherwise appropriate. Examples: (1) test markets selected to determine the potential of a new product, (2) purchase engineers selected in industrial marketing research because they are considered to be representative of the company, (3) product testing with individuals who may be particularly fussy or who hold extremely high expectations, (4) expert witnesses used in court.

Quota sampling: two-stage restricted judgemental sampling that is used extensively in street interviewing.

- The first stage consists of developing control characteristics, or quotas, of population elements such as age or gender. To develop these quotas, the researcher lists relevant control characteristics and determines the distribution of these characteristics in the target population, such as Males 49%, Females 51% (resulting in 490 men and 510 women being selected in a sample of 1,000 respondents). Often, the quotas are assigned so that the proportion of the sample elements possessing the control characteristics is the same as the proportion of population elements with these characteristics. In other words, the quotas ensure that the composition of the sample is the same as the composition of the population with respect to the characteristics of interest.
- In the second stage, sample elements are selected based on convenience or judgement.

- **Snowball sampling:** an initial group of respondents is selected who possess the desired characteristics of the target population. After being interviewed, these respondents are asked to identify others who belong to the target population. Subsequent respondents are selected based on the referrals. By obtaining referrals from referrals, this process may be carried out in waves, thus leading to a snowballing effect. The main objective of snowball sampling is to estimate characteristics that are rare in the wider population.
- Examples: users of particular government or social services, such as parents who use nurseries or child minders, whose names cannot be revealed; special census groups, such as widowed males under 35; and members of a scattered minority ethnic group; Industrial buyer using some special equipment or technology;

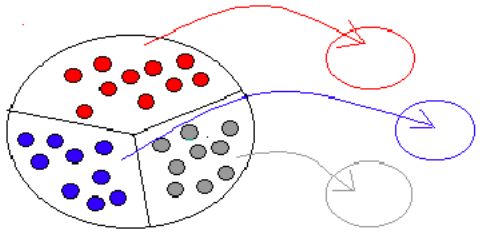
B. Random/Probability- based sampling

1. Simple random sampling

- Each element/item has equal chance of getting included in a sample. Randomness.
- Sampling with/without replacement
- Random number table, pseudo-random number generator.

2. Stratified Sampling

- Each stratum is a homogeneous group and different from other strata.
- Random selection from each stratum, proportionately.



3. Cluster sampling

- Least or no variation among clusters.
- Clusters are selected randomly for further analysis.
- Area sampling in geographical clusters.
- Multi-stage sampling as a special case.

4. Systematic sampling

- Elements selected at a uniform interval.
- Selection evenly spread, less cost & time, more convenient.
- the sample is chosen by selecting a random starting point and then picking every *i* th element in succession from the sampling frame.
- The sampling interval, *i*, is determined by dividing the population size *N* by the sample size *n* and rounding to the nearest whole number. For example, there are 100,000 elements in the population and a sample of 1,000 is desired. In this case, the sampling interval, *i*, is 100. A random number between 1 and 100 is selected. If, for example, this number is 23, the sample consists of elements 23, 123, 223, 323, 423, 523, and so on.

Sample Size Determination:

$$n = \frac{\sigma^2 z^2}{D^2} \quad \text{for mean}$$

$$n = \frac{p(1-p)z^2}{D^2} \quad \text{for proportion}$$

$D = \text{Level of precision}$

z is associated with Confidence Interval

SAMPLING DISTRIBUTION

- Sampling Distribution: Distribution of a sample statistics, usually mean.
- Standard error($\sigma_{\bar{x}}$): Standard deviation of the sampling distribution.
- Mean of sampling distribution($\mu_{\bar{x}}$) of means, taking all possible samples exhaustively, approaches to population mean (μ), particularly for normal population distribution.
- As sample size increases, standard error decreases.

Assuming Normal Population Distribution

$$\textbf{Standard error}(\sigma_{\bar{x}}) = \frac{\textbf{Standard Deviation of Population}}{\sqrt{\mathbf{n}}}$$

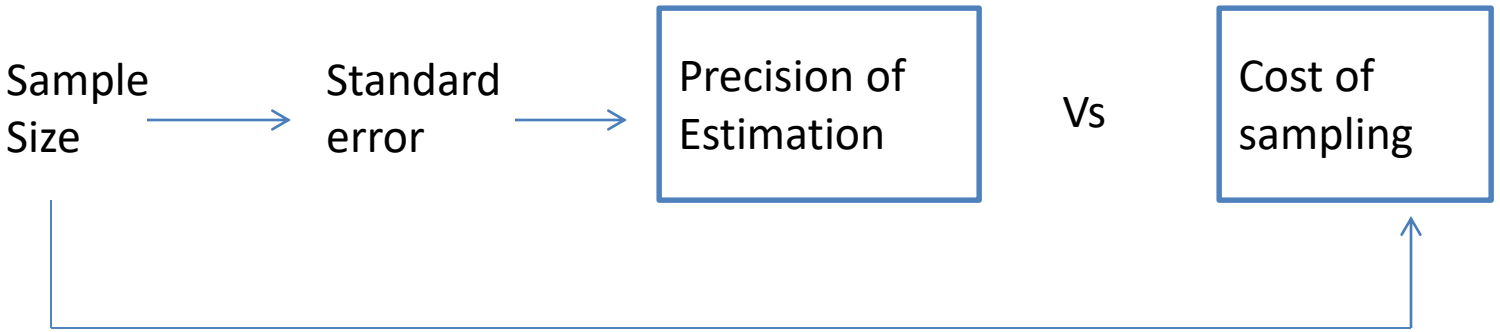
n = Sample size

$$\sigma_{\bar{x}} = \frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}} \textbf{ for larger or infinite population}$$

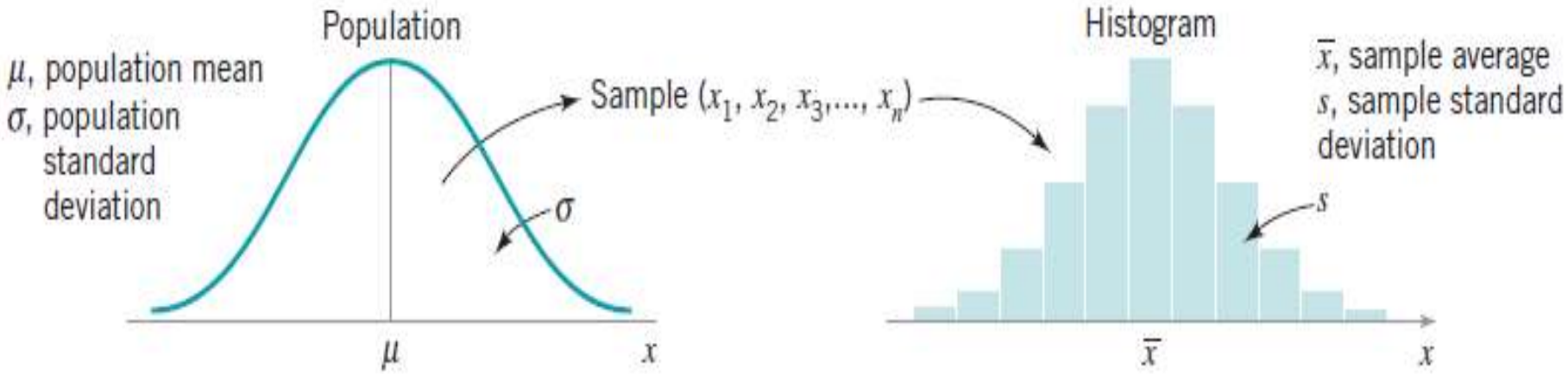
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \text{ for finite population of size N}$$

Central Limit Theorem:

- Irrespective of shape of population distribution, sampling distribution approaches to normal, as sample size increases.
- Mean of such sampling distribution is population mean.



Point Estimate



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

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

Interval Estimate.

- **Confidence Level**:
  - Level of significance, α
  - Probability that is associated with an interval estimate (1- α), of any population parameter.
  - Higher confidence level => Wider confidence interval

Estimation of mean from large sample(usually n> 30):

As sample size is large, sampling distribution of mean is normal.

1. Compute $\sigma_{\bar{x}}$ from either known σ or estimated

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

2. Get Z value from standard normal distribution table corresponding to confidence level (1- α).
3. The confidence interval

$$\bar{x} \pm Z\sigma_{\bar{x}}$$

Estimation of means from small samples(n<30):

t-distribution:

- Applicable for smaller sample size.
- Unimodal and almost like a bell shape.
- Flatter than normal.
- Larger the sample size less flatter the distribution shape and closer to normal.
- Value of t varies with d.f.i.e.(n-1) as the distribution shape changes.

Step 1. Compute ($\sigma_{\bar{x}}$) as usual

Step 2. Get t value from t- distribution table corresponding to (n- 1) as d.f. and (1- confidence level) as the area under curve.

Step 3. $\bar{X} \pm t\ \sigma_{\bar{x}}$ is the confidence interval/limit.

Case		Two sided Confidence Interval (CI)
Population standard deviation, σ known		$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
Population standard deviation, σ unknown	Sample size n > 30	$\bar{x} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}}$
	Sample size n ≤ 30	$\bar{x} \pm t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}$

Example 1: A sample of size 20 was collected and the sample mean and standard deviation are estimated as 9.8525 and 0.0965. Find 95% two-sided CI for the mean.

- **Example 2**: The life in hours of a light bulb is known to be approximately normally distributed with standard deviation of 25 hours. A random sample of 40 bulbs has a mean life of 1014 hours.
  1. Construct a 95% two-sided CI on the mean life.
  2. Construct a 95% one-sided lower CI of the mean life.

One-sided confidence interval: Appropriate lower or upper confidence limit are found by replacing

$Z_{\alpha/2}$ by Z_{α} and $t_{\frac{\alpha}{2},n-1}$ by $t_{\alpha,n-1}$

- **Example 3**: The following result shows the investigation of the haemoglobin level of hockey players (in g/dl).

15.3	16.0	14.4	16.2	16.2
14.9	15.7	14.6	15.3	17.7
16.0	15.0	15.7	16.2	14.7
14.8	14.6	15.6	14.5	15.2

a) Find the 90% two-sided CI on the mean haemoglobin level.

15.43684211
0.83413996

b) Also construct 90% Upper CI on the mean haemoglobin level.

Confidence Interval on the Variance of a Normal Distribution

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$

Confidence Intervals on a Population Proportion

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + Z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example: An automatic filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of 0.0153. If the variance of fill volume is too large, an unacceptable proportion of bottles will be under- or overfilled. We will assume that the fill volume is approximately normally distributed. Calculate 95% upper-confidence interval for variance.

$$\sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{0.95,19}} \qquad \sigma^2 \leq \frac{(19)0.0153}{10.117} = 0.0287$$

$$\sigma \leq 0.17$$

Therefore, at the 95% level of confidence, the data indicate that the process standard deviation could be as large as 0.17

Example: In a random sample of 85 automobile engine crankshaft bearings, 10 have a surface finish that is rougher than the specifications allow. Therefore, a point estimate of the proportion of bearings in the population that exceeds the roughness specification is $\hat{p} = \frac{x}{n} = \frac{10}{85} = 0.12$. Compute 95% two-sided confidence interval for *p.*

$$\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.12 - 1.96\sqrt{\frac{0.12(0.88)}{85}} \leq p \leq 0.12 + 1.96\sqrt{\frac{0.12(0.88)}{85}}$$

$$0.05 \leq p \leq 0.19$$