Sample and Sampling Distribution

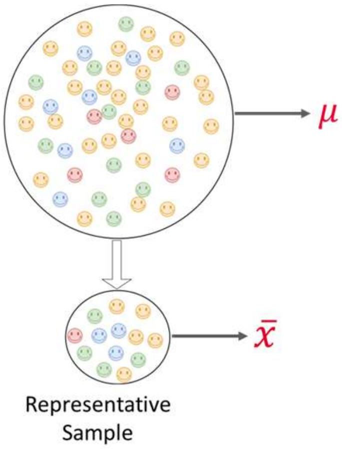
Center for Data Research and Analytics

Population and Parameter

- Research aims to know the population through the parameters

Sample and Estimate

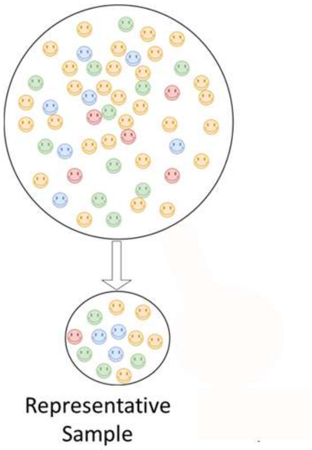
- The primary research interest is to find the parameter of the target population
- Population is not accessible in most of the cases because of
  - Time/budget constraint
  - The size of the population
  - Ethical constraint
- Research adopt a method of sampling from the population. A sample is a representative subset of the population that is assumed to have all properties of the population.
- We estimate the parameter of the population from the sample
- We are now concerned whether the $\bar{x}$ can really estimate the parameter μ



A sample is a representative part of the population. We get an estimate of the parameter from a sample

Sampling Methods

- We draw subjects from the population at random for a representative sample
- We use different randomization scheme depending on the characteristics of the population. Some of the popular schemes are as follows:
  - Simple Random Sampling
  - Stratified Random Sampling
  - Cluster Sampling
  - Multi-stage Sampling
- Biased selection of sample could bias the parameter estimate to any suitable level



Biased sample may provide the biased estimate of the parameter

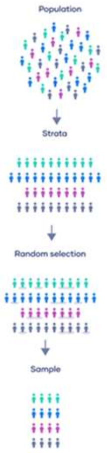
Simple Random Sampling

- The population is homogeneous, and subjects are independent and identical in terms of risk/probability of the selection we can used the SRS
- We select randomly so that every subjects have equal chance of being selected. Ideally, we need the complete list of subjects. We can adopt some method that works without the frame but that should be pre-defined.
- Since all subjects are independent and identical, we can apply all statistical methods and tests.

Selecting subjects by a researcher is not random.

Stratified Sampling

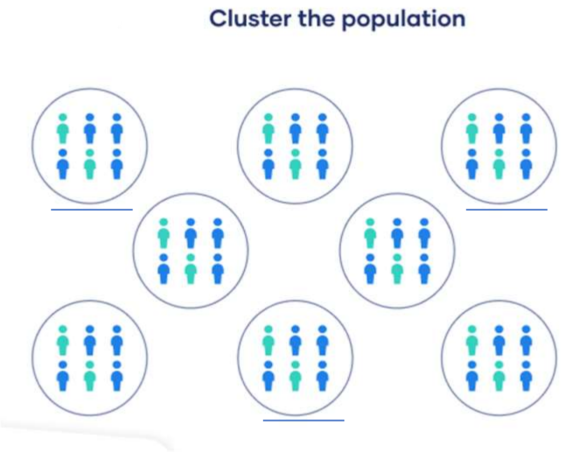
- The population is not homogeneous in terms of risk/probability of the outcome. E.g., risk of COVID-19 infection is higher in males than females
- We divide the populations in groups where the risk is appeared to be homogeneous.
- Perform simple random sample from each of the groups of the population.
- Sample may include the same proportion of subjects from each groups as in the population.
- If the stratum sizes varies too much, you might require to used the weighted estimates

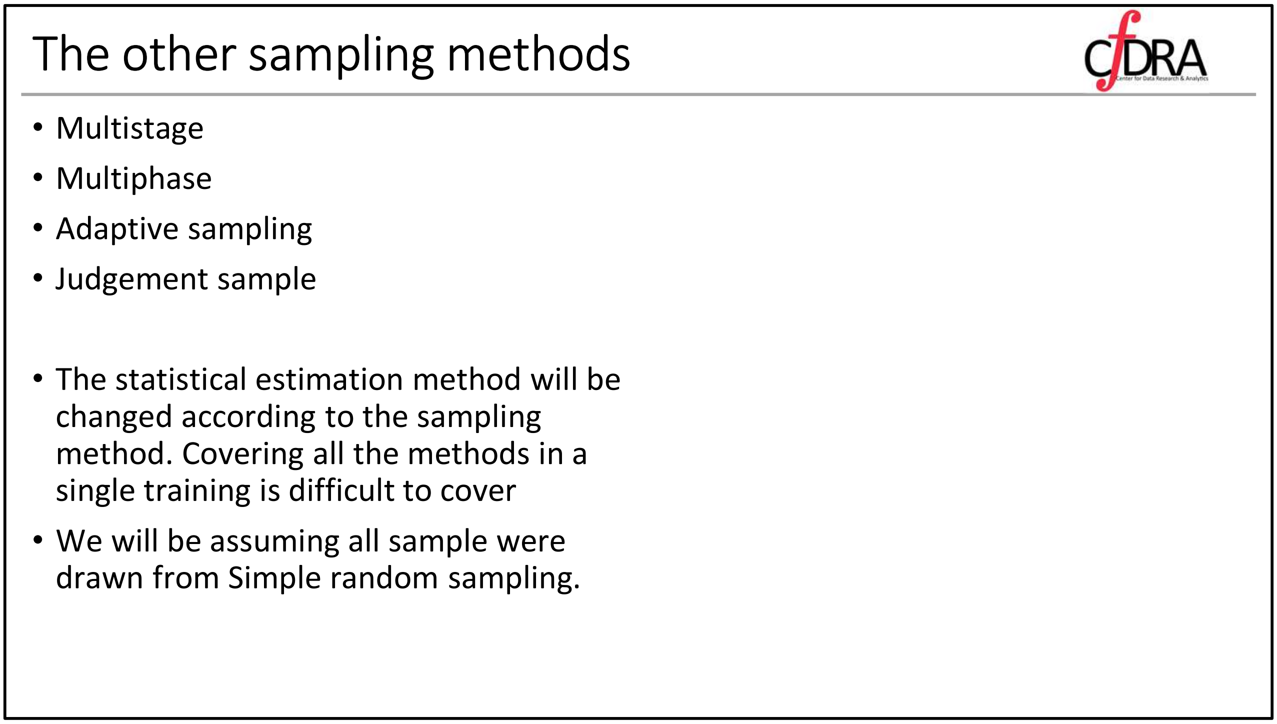


We perform simple random sample within relatively homogeneous stratas

Cluster Sampling

- If we can divide the whole population into a multiple homogeneous sub-groups where each of the subgroups are alike, and withing sub-groups they are heterogeneous.
- Each sub-groups are called cluster
- We can take a few subgroups randomly to get the representative sample of the population
- For example, we can select 10 random district to make a representative of whole Bangladesh people. Here Districts are clusters, and within districts people are heterogeneous.





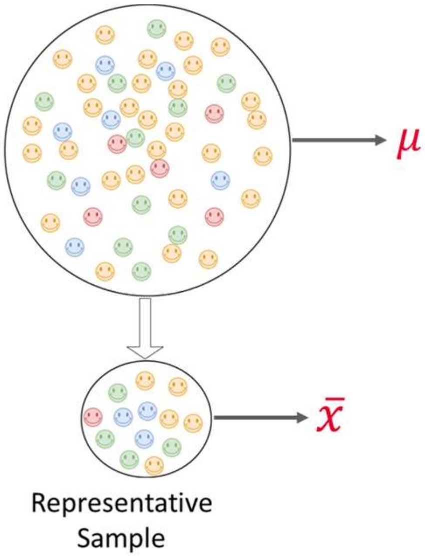
The other sampling methods

- Multistage
- Multiphase
- Adaptive sampling
- Judgement sample

- The statistical estimation method will be changed according to the sampling method. Covering all the methods in a single training is difficult to cover
- We will be assuming all sample were drawn from Simple random sampling.

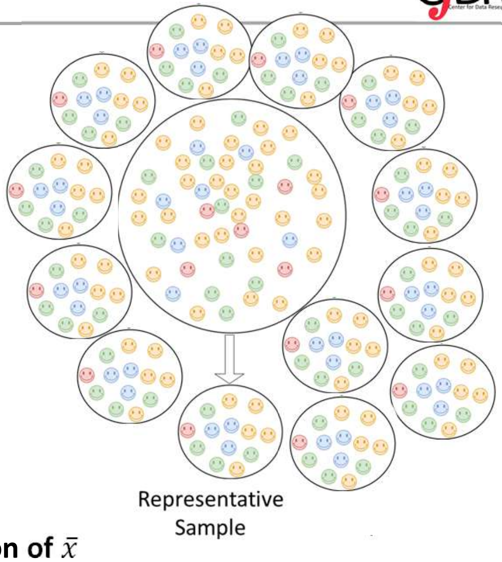
Estimator vs parameter

- Suppose we somehow ensure that the sample we received is a representative part of the population
- Subjects were randomly selected
- **Is it guaranteed** that an estimator will provide the right estimate of the true parameter
- Answer Yes, by the law of large number. That tell if you increase the sample size, eventually the sample estimator of means will converge to the true average.



Sampling distribution.

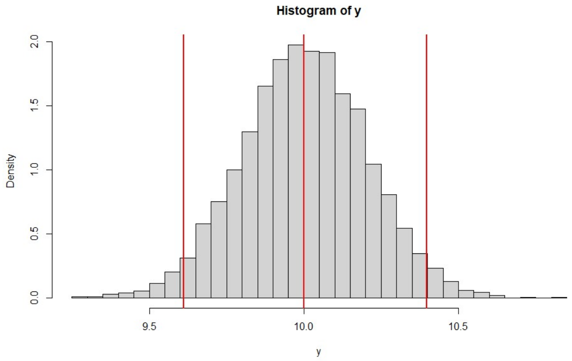
- The sampling distribution of an estimator provide the level of reliability which is termed as central limit theorem
- We can draw multiple samples from a population. So, each estimate is a random variable. We will get a different $\bar{x}$ with different samples
- The central limit theorem tells that the distribution of $\bar{x}$ will be normally distributed with mean of the true population parameter.



The central limit theorem gives us the distribution of $\bar{x}$

Sampling Distribution

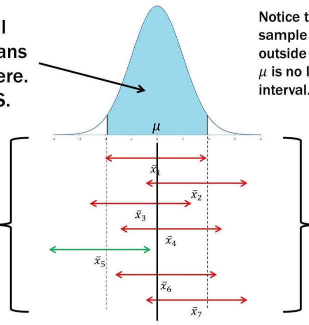
- Population mean was 10 and standard deviation was 2
- We sampled of size 100 each time and computed the mean
- 95% mean estimates are within the range of (9.6, 10.4)
- Mean of the Means = 10.0
- However, we never looked at the sampling distribution, rather we use this for inference. We will discuss in the next slides



We are now confident that if the true mean is 10, a mean from a sample will be around 10

95% Confidence Interval (CI)

- Suppose we draw sample from a population with mean = 10 and SD = 2
- The sample give us $\bar{x}$. We are interested in the distribution of $\bar{x}$.
- We use 95% CI for the $\bar{x}$.
- 95% CI: 9.6, 10.4 for a single sample
- The CI is interpreted as, if we draw samples multiple time and construct this kind of CI, then 95% of such CI will contain the true mean.



95% of all sample means $(\bar{x})$ are in here. THEN THIS.

Notice that as soon as a sample mean steps outside the dotted line, μ is no longer in its interval.

Many samples of the same size. THESE COME FIRST.

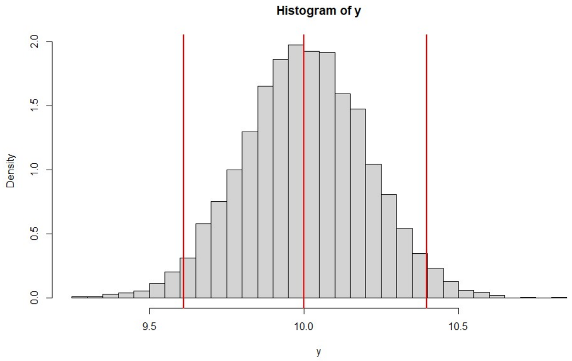
Samples of the same size have the same standard error $\sigma_{\bar{x}}$. So the 95% "width" is the same for all samples of that size.

Confidence interval is an interval estimate of the parameter. We try to grab the CI of the sampling distribution of $\bar{x}$.

Hypothesis testing

- Null hypothesis
  - The mean $\bar{x} = \mu = 10$

- How we can test the null (Using sampling distribution)
- We get our estimate $\bar{x} = 9.65$ from a sample
- Let us check how rate our $\bar{x}$ is in the sampling distribution under the null
- First, we see that our $\bar{x}$ is within the range of 95% samples
- The probability that the $\bar{x}$ is 96.21%
- P-value = 0.96



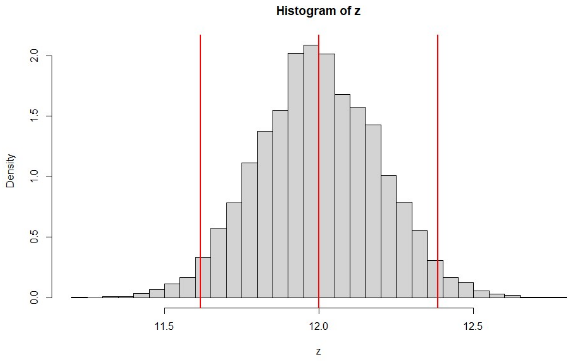
The true population mean was 10

The observed $\bar{x}$ is not a rare value under the null hypothesis

Hypothesis testing

- Null hypothesis
  - The mean $\bar{x} = \mu = 12$

- How we can test the null (Using sampling distribution)
- We get our estimate $\bar{x} = 9.65$ from a sample
- Let us check how rare our $\bar{x}$ is in the sampling distribution under the null
- First, we see that our $\bar{x}$ is **not** within the range of 95% samples
- The probability that $\bar{x}$ = 9.65 under the distribution is <0.25 for sure
- P-value < 0.001



The $\bar{x}$ is a rare value under the null