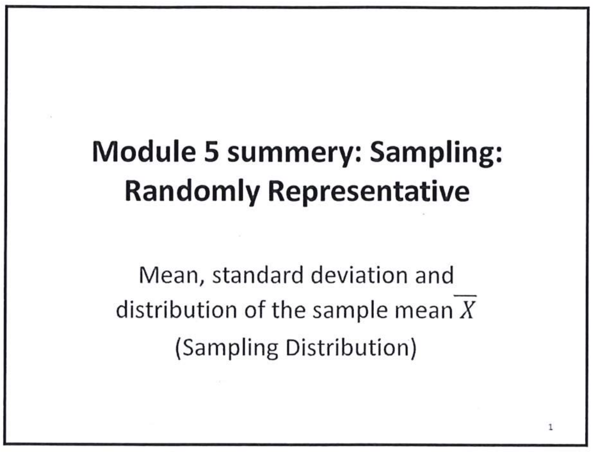
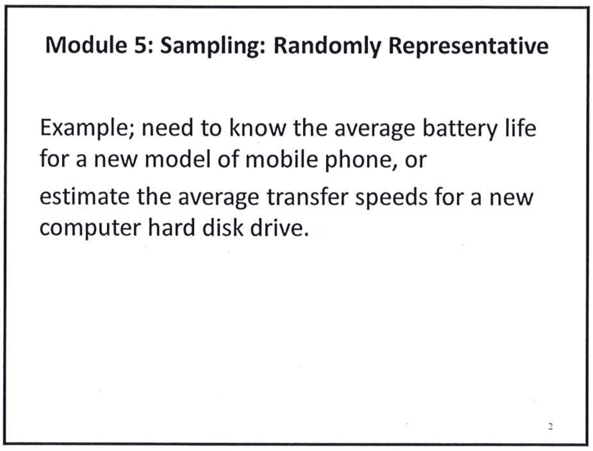


Module 5 summery: Sampling: Randomly Representative

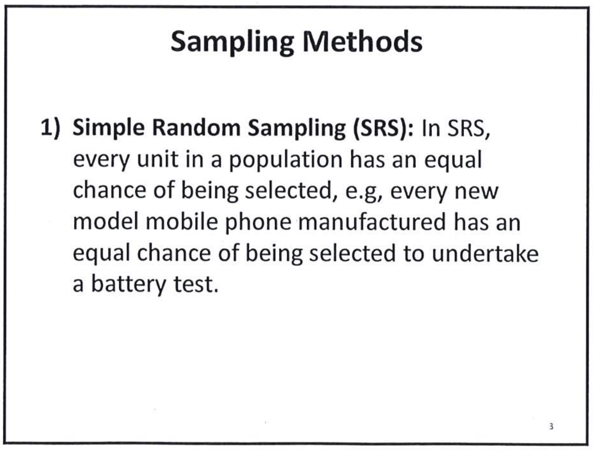
Mean, standard deviation and distribution of the sample mean $\overline{X}$ (Sampling Distribution)



Module 5: Sampling: Randomly Representative

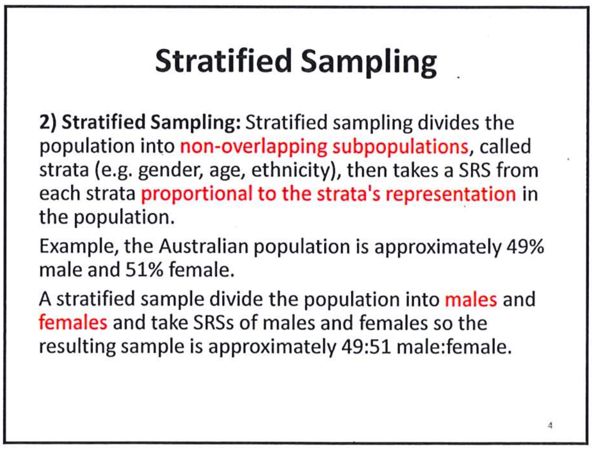
Example; need to know the average battery life for a new model of mobile phone, or

estimate the average transfer speeds for a new computer hard disk drive.



Sampling Methods

1) **Simple Random Sampling (SRS):** In SRS, every unit in a population has an equal chance of being selected, e.g, every new model mobile phone manufactured has an equal chance of being selected to undertake a battery test.

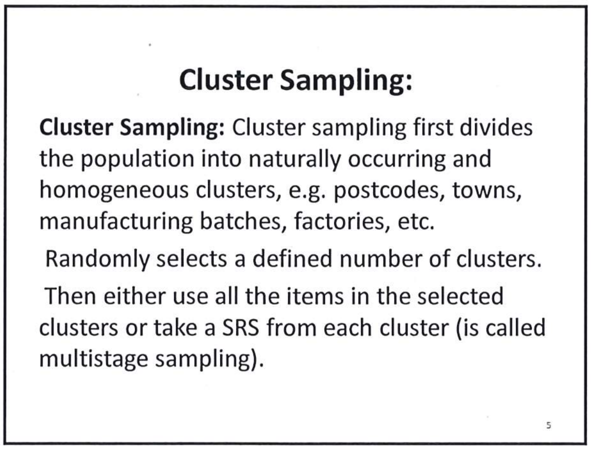


Stratified Sampling

2) Stratified Sampling: Stratified sampling divides the population into non-overlapping subpopulations, called strata (e.g. gender, age, ethnicity), then takes a SRS from each strata proportional to the strata's representation in the population.

Example, the Australian population is approximately 49% male and 51% female.

A stratified sample divide the population into males and females and take SRSs of males and females so the resulting sample is approximately 49:51 male:female.

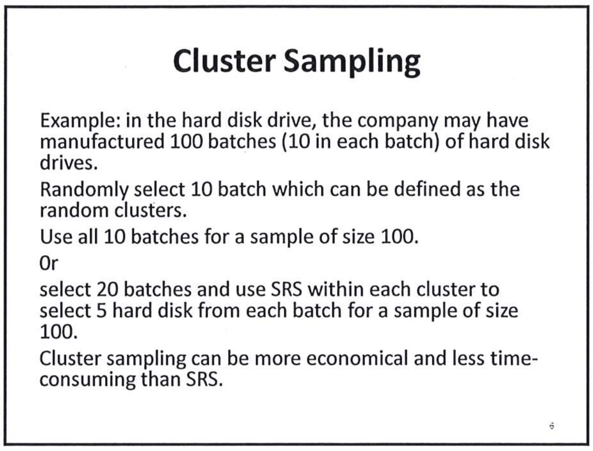


Cluster Sampling:

Cluster Sampling: Cluster sampling first divides the population into naturally occurring and homogeneous clusters, e.g. postcodes, towns, manufacturing batches, factories, etc.

Randomly selects a defined number of clusters.

Then either use all the items in the selected clusters or take a SRS from each cluster (is called multistage sampling).



Cluster Sampling

Example: in the hard disk drive, the company may have manufactured 100 batches (10 in each batch) of hard disk drives.

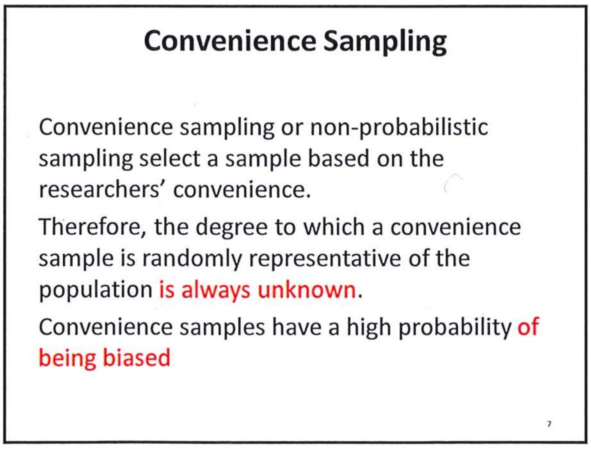
Randomly select 10 batch which can be defined as the random clusters.

Use all 10 batches for a sample of size 100.

Or

select 20 batches and use SRS within each cluster to select 5 hard disk from each batch for a sample of size 100.

Cluster sampling can be more economical and less time-consuming than SRS.

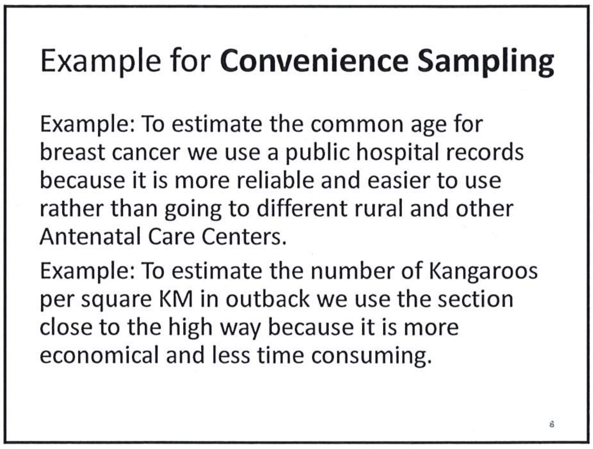


Convenience Sampling

Convenience sampling or non-probabilistic sampling select a sample based on the researchers' convenience.

Therefore, the degree to which a convenience sample is randomly representative of the population is always unknown.

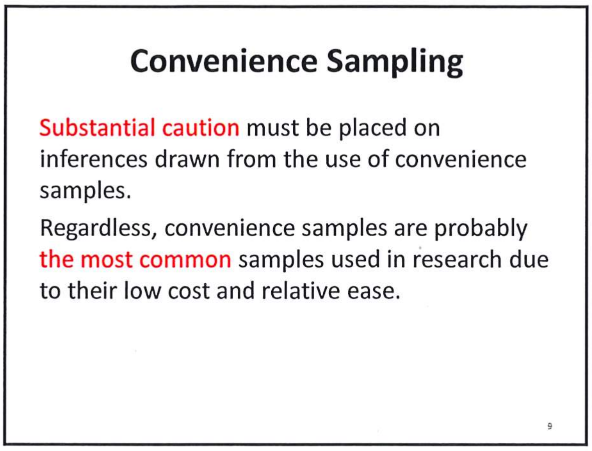
Convenience samples have a high probability of being biased



Example for **Convenience Sampling**

Example: To estimate the common age for breast cancer we use a public hospital records because it is more reliable and easier to use rather than going to different rural and other Antenatal Care Centers.

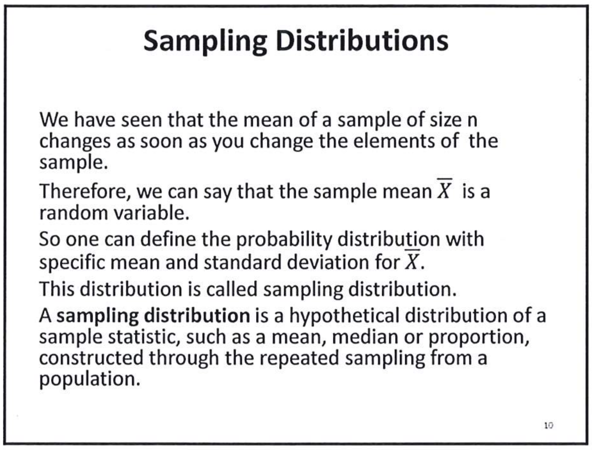
Example: To estimate the number of Kangaroos per square KM in outback we use the section close to the high way because it is more economical and less time consuming.



Convenience Sampling

Substantial caution must be placed on inferences drawn from the use of convenience samples.

Regardless, convenience samples are probably the most common samples used in research due to their low cost and relative ease.



Sampling Distributions

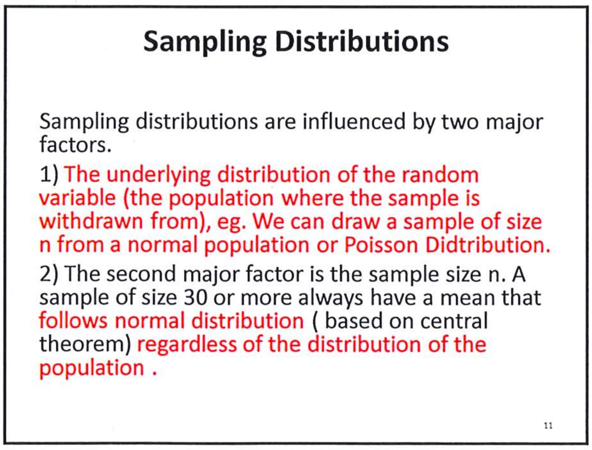
We have seen that the mean of a sample of size n changes as soon as you change the elements of the sample.

Therefore, we can say that the sample mean $\overline{X}$ is a random variable.

So one can define the probability distribution with specific mean and standard deviation for $\overline{X}$.

This distribution is called sampling distribution.

A **sampling distribution** is a hypothetical distribution of a sample statistic, such as a mean, median or proportion, constructed through the repeated sampling from a population.

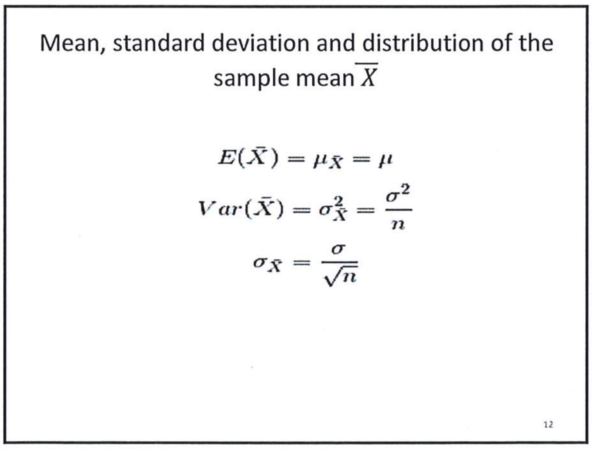


Sampling Distributions

Sampling distributions are influenced by two major factors.

1) The underlying distribution of the random variable (the population where the sample is withdrawn from), eg. We can draw a sample of size n from a normal population or Poisson Didtribution.

2) The second major factor is the sample size n. A sample of size 30 or more always have a mean that follows normal distribution (based on central theorem) regardless of the distribution of the population .



Mean, standard deviation and distribution of the sample mean $\overline{X}$

$$E(\bar{X}) = \mu_{\bar{X}} = \mu$$

$$Var(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Mean, standard deviation and distribution of the sample mean $\overline{X}$

Example: The YouTube video duration has μ = 193 Sec, σ^2= 193^2 , σ = 193. For a sample size n=10 we have:

$$\mu_{\bar{X}} = 193$$

$$\sigma^2_{\bar{X}} = \frac{193^2}{10} = 3724.9$$

$$\sigma_{\bar{X}} = \frac{193}{\sqrt{10}} = 61.03$$

Standard Error of $\overline{X}$ SE($\overline{X}$)

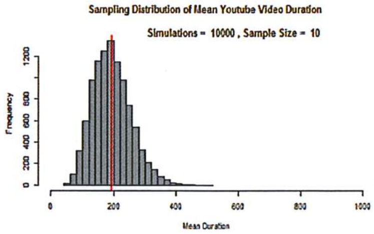
The SE($\overline{X}$) of mean gets smaller as the sample size n increases

If we increase the n to 100 for the previous sample the

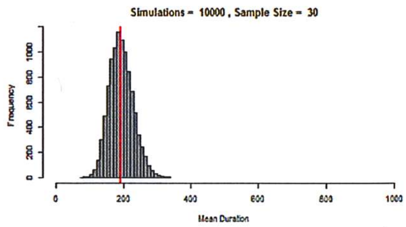
$$SE = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$ =193/sqrt(100)=

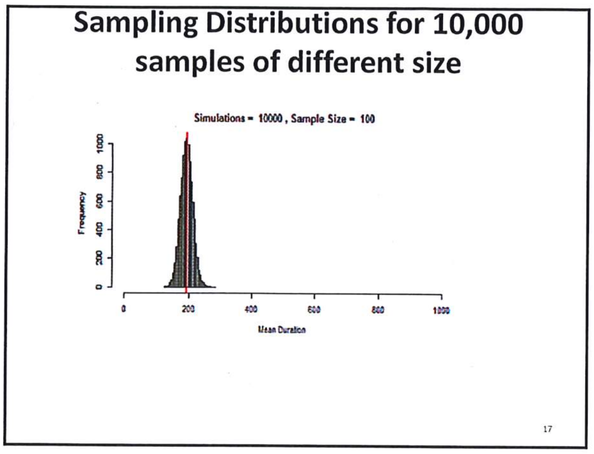
193/10= 19.3

Sampling Distributions for 10,000 samples of different size



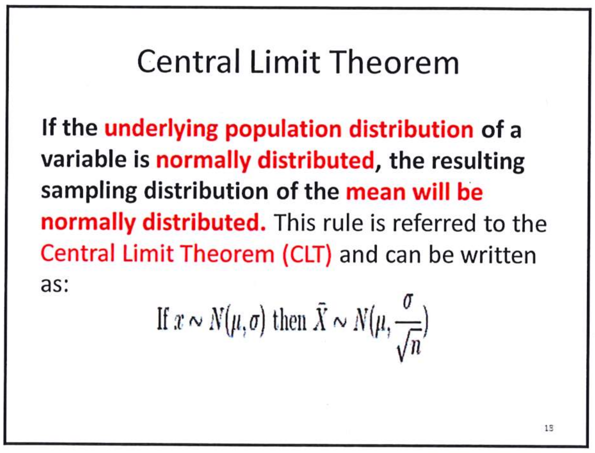
Sampling Distributions for 10,000 samples of different size





Sampling Distributions for 10,000 samples of different size

17

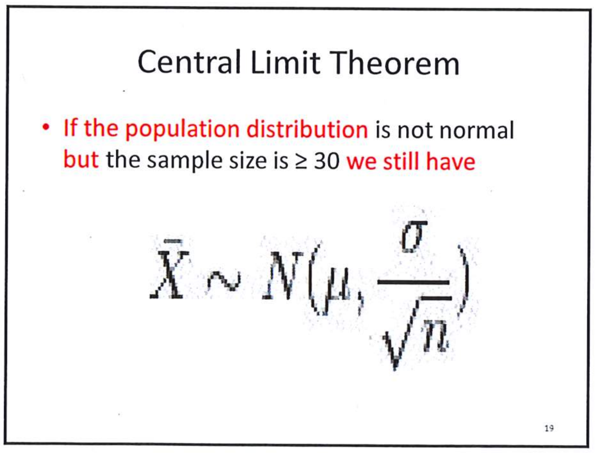


Central Limit Theorem

If the underlying population distribution of a variable is normally distributed, the resulting sampling distribution of the mean will be normally distributed. This rule is referred to the Central Limit Theorem (CLT) and can be written as:

$$\text{If } x \sim N(\mu, \sigma) \text{ then } \bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

18

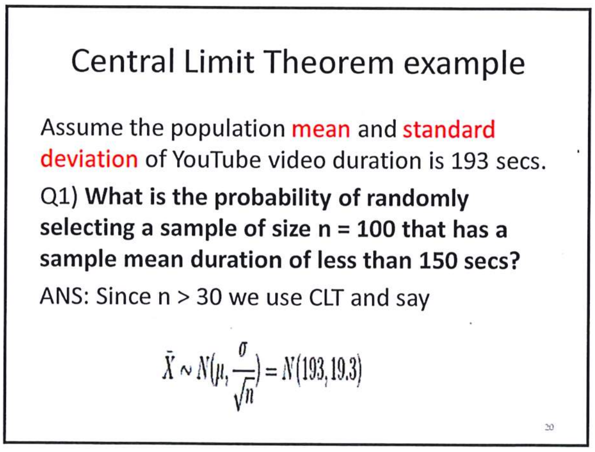


Central Limit Theorem

- If the population distribution is not normal but the sample size is ≥ 30 we still have

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

19



Central Limit Theorem example

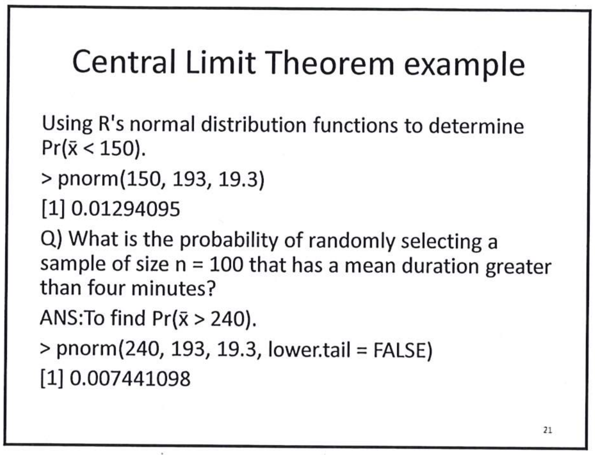
Assume the population mean and standard deviation of YouTube video duration is 193 secs.

Q1) **What is the probability of randomly selecting a sample of size n = 100 that has a sample mean duration of less than 150 secs?**

ANS: Since n > 30 we use CLT and say

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) = N(193, 19.3)$$

20



Central Limit Theorem example

Using R's normal distribution functions to determine Pr(x̄ < 150).

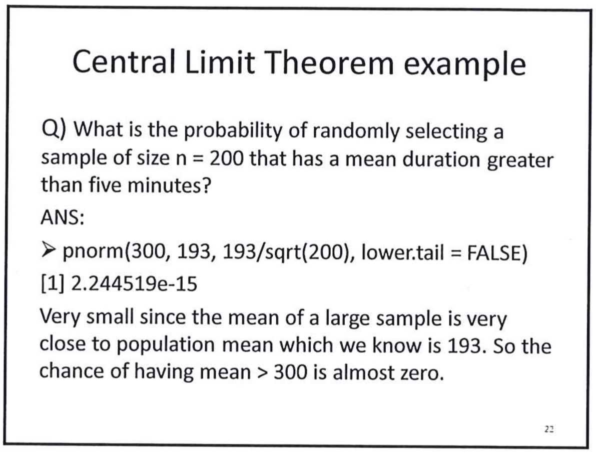
```
> pnorm(150, 193, 19.3)
[1] 0.01294095
```

Q) What is the probability of randomly selecting a sample of size n = 100 that has a mean duration greater than four minutes?

ANS:To find Pr(x̄ > 240).

```
> pnorm(240, 193, 19.3, lower.tail = FALSE)
[1] 0.007441098
```

21



Central Limit Theorem example

Q) What is the probability of randomly selecting a sample of size n = 200 that has a mean duration greater than five minutes?

ANS:

```
➢ pnorm(300, 193, 193/sqrt(200), lower.tail = FALSE)
[1] 2.244519e-15
```

Very small since the mean of a large sample is very close to population mean which we know is 193. So the chance of having mean > 300 is almost zero.

22