



# Sampling and Sampling Distributions

---



# Why Sample?

---

- Selecting a sample is less time-consuming than selecting every item in the population (**census**).
- Selecting a sample is less costly than selecting every item in the population.
- An analysis of a sample is less cumbersome and more practical than an analysis of the entire population.

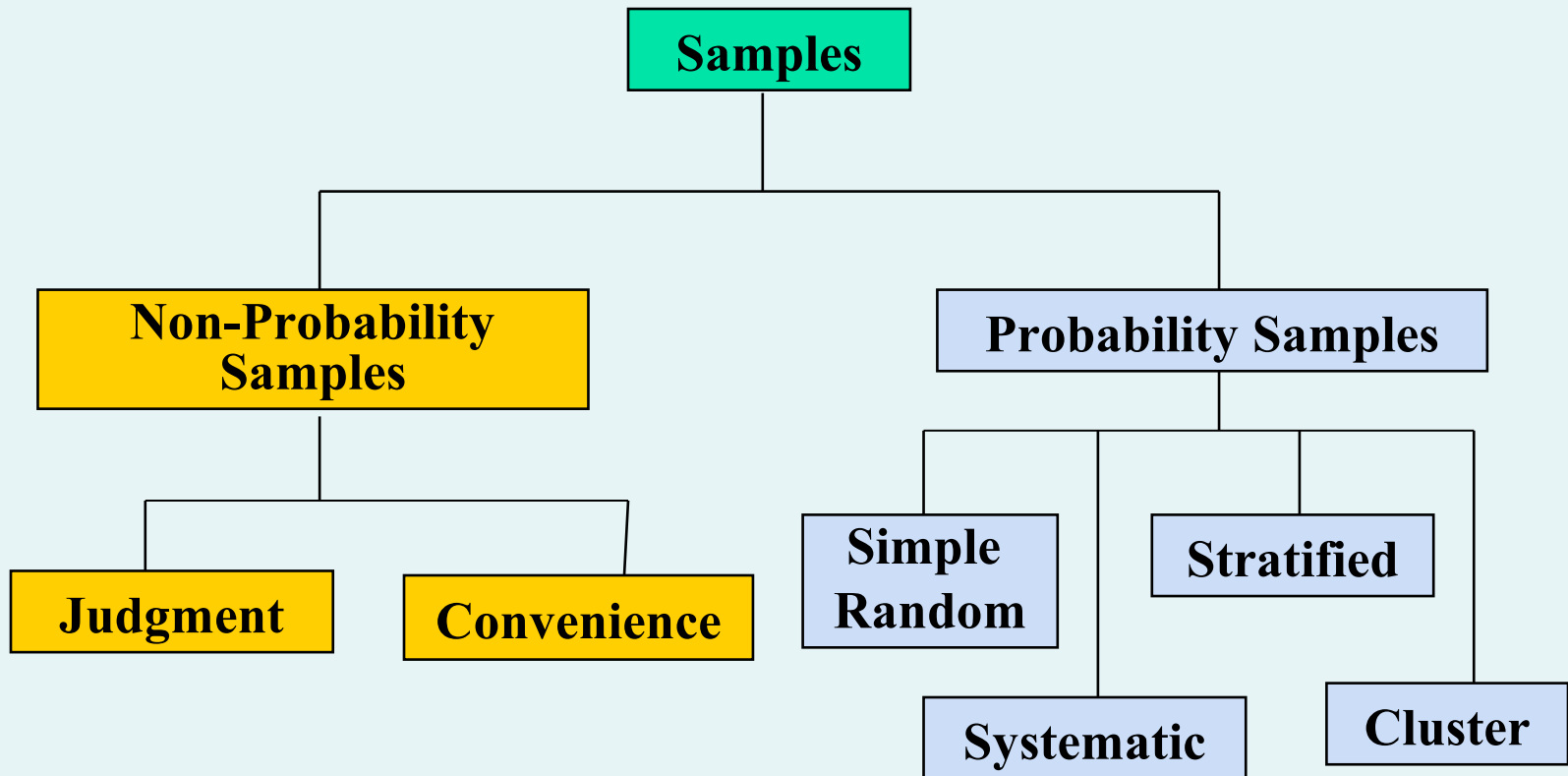


# A Sampling Process Begins With A Sampling Frame

---

- The sampling frame is a listing of items that make up the population
- Frames are data sources such as population lists, directories, or maps
- Inaccurate or biased results can result if a frame excludes certain portions of the population
- Using different frames to generate data can lead to dissimilar conclusions

# Types of Samples





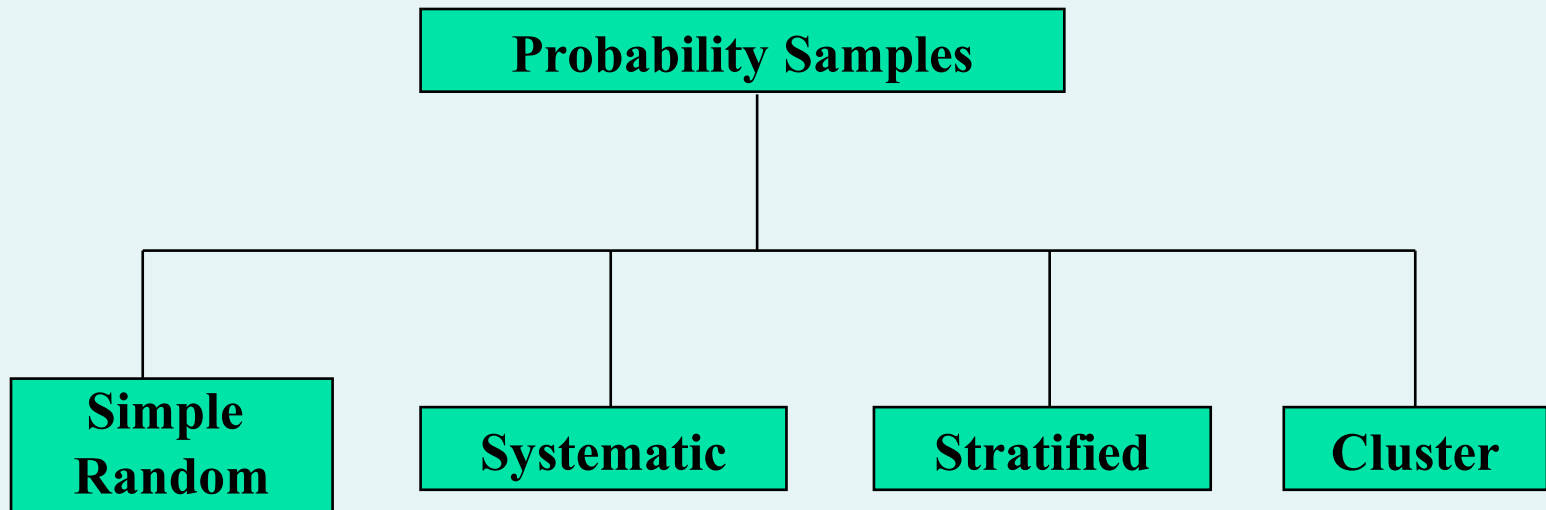
# Types of Samples: Nonprobability Sample

---

- In a nonprobability sample, items included are chosen without regard to their probability of occurrence.
  - In **convenience sampling**, items are selected based only on the fact that they are easy, inexpensive, or convenient to sample.
  - In a **judgment sample**, you get the opinions of pre-selected experts in the subject matter.

# Types of Samples: Probability Sample

- In a **probability sample**, items in the sample are chosen on the basis of known probabilities.





# Probability Sample: Simple Random Sample

---

- Every individual or item from the frame has an equal chance of being selected
- Selection may be with replacement (selected individual is returned to frame for possible reselection) or without replacement (selected individual isn't returned to the frame).
- Samples obtained from table of random numbers or computer random number generators.

# Probability Sample: Systematic Sample

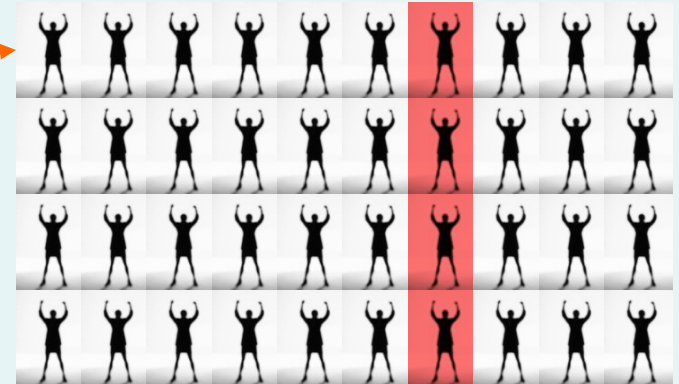
- Decide on sample size:  $n$
- Divide frame of  $N$  individuals into groups of  $k$  individuals:  $k=N/n$
- Randomly select one individual from the 1<sup>st</sup> group
- Select every  $k^{\text{th}}$  individual thereafter

$$N = 40$$

$$n = 4$$

$$k = 10$$

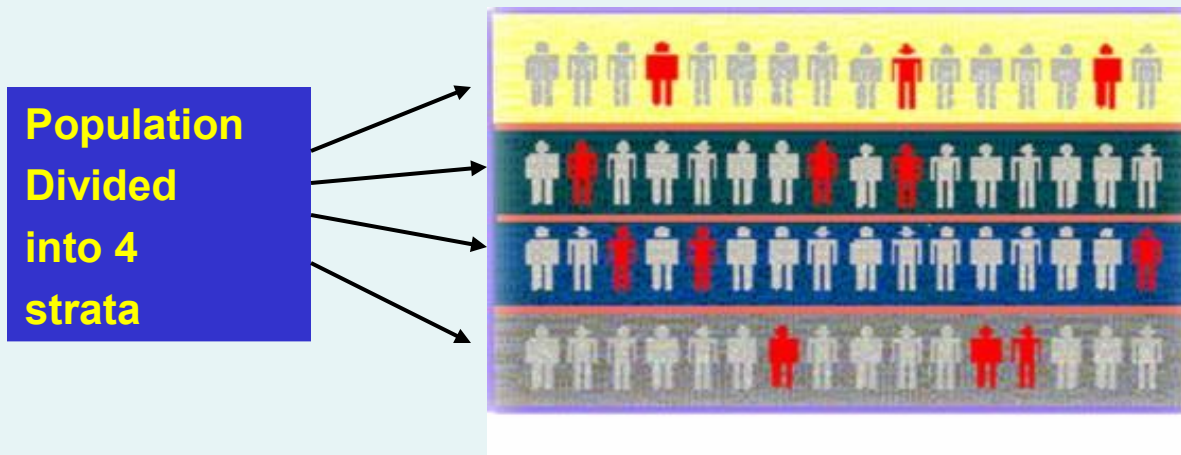
First Group





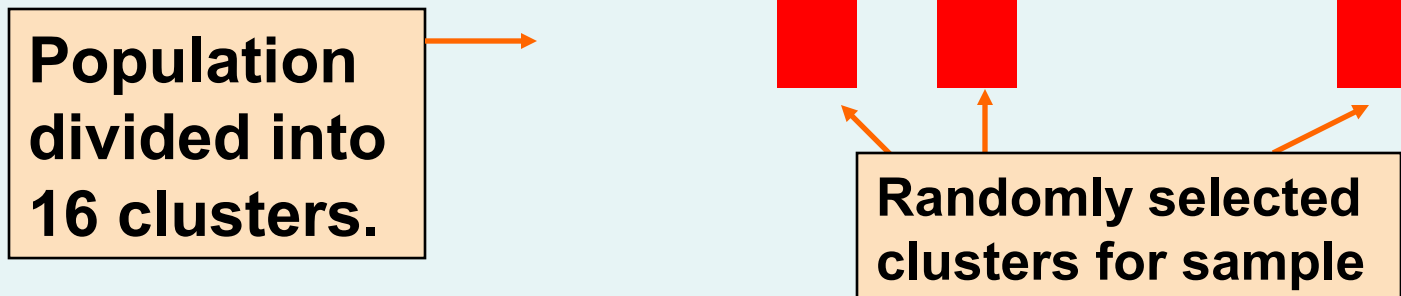
# Probability Sample: Stratified Sample

- Divide population into two or more subgroups (called *strata*) according to some common characteristic
- A simple random sample is selected from each subgroup, with sample sizes proportional to strata sizes
- Samples from subgroups are combined into one
- This is a common technique when sampling population of voters, stratifying across racial or socio-economic lines.



# Probability Sample Cluster Sample

- Population is divided into several “clusters,” each representative of the population
- A simple random sample of clusters is selected
- All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique
- A common application of cluster sampling involves election exit polls, where certain election districts are selected and sampled.





# Probability Sample: Comparing Sampling Methods

---

- Simple random sample and Systematic sample
  - Simple to use
  - May not be a good representation of the population's underlying characteristics
- Stratified sample
  - Ensures representation of individuals across the entire population
- Cluster sample
  - More cost effective
  - Less efficient (need larger sample to acquire the same level of precision)



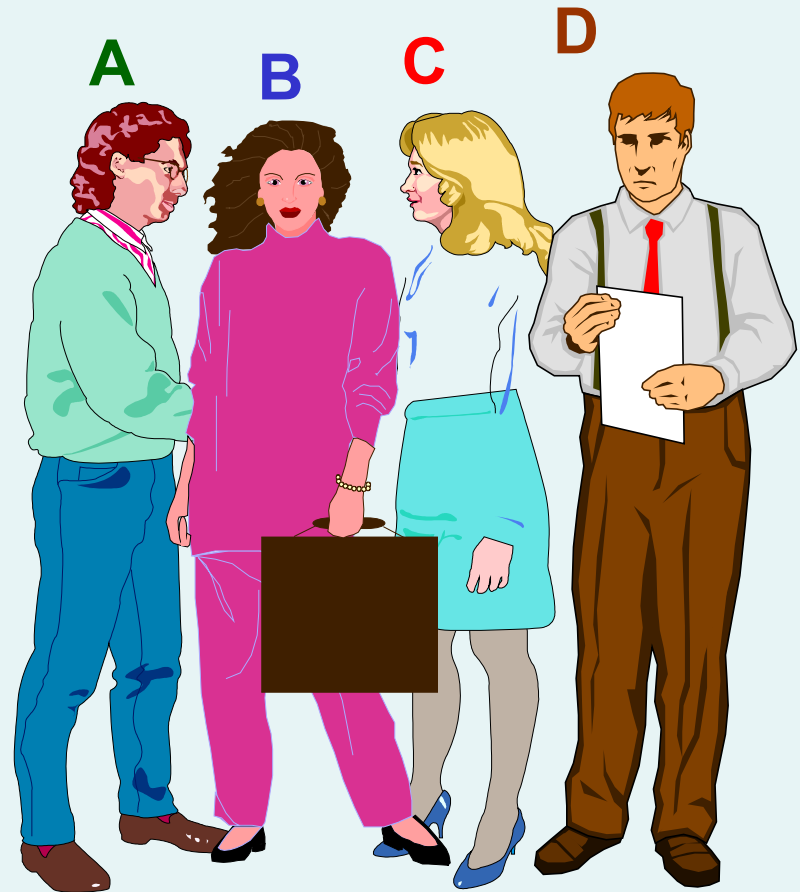
# Sampling Distributions

---

- A sampling distribution is a distribution of all of the possible values of a sample statistic for a given size sample selected from a population.
- For example, suppose you sample 50 students from your college regarding their mean CGPA. If you obtained many different samples of 50, you will compute a different mean for each sample. We are interested in **the distribution of all potential mean CGPA we might calculate for any given sample of 50 students.**

# Developing a Sampling Distribution

- Assume there is a population ...
- Population size  $N=4$
- Random variable,  $X$ , is **age** of individuals
- Values of  $X$ : 18, 20, 22, 24 (years)



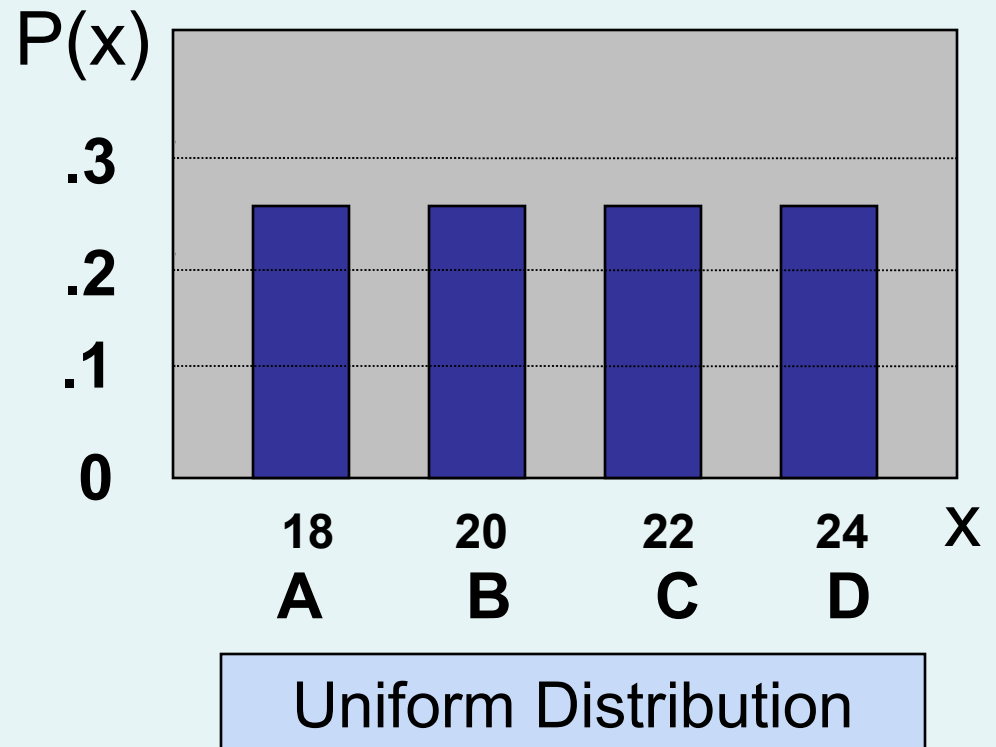
# Developing a Sampling Distribution

(continued)

Summary Measures for the Population Distribution:

$$\begin{aligned}\mu &= \frac{\sum X_i}{N} \\ &= \frac{18 + 20 + 22 + 24}{4} = 21\end{aligned}$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$$



# Developing a Sampling Distribution

(continued)

Now consider all possible samples of size  $n=2$

<b>1<sup>st</sup> Obs</b>	<b>2<sup>nd</sup> Observation</b>			
	18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 possible samples  
(sampling with  
replacement)



16 Sample  
Means

<b>1<sup>st</sup> Obs</b>	<b>2<sup>nd</sup> Observation</b>			
	18	20	22	24
<b>18</b>	18	19	20	21
<b>20</b>	19	20	21	22
<b>22</b>	20	21	22	23
<b>24</b>	21	22	23	24

# Developing a Sampling Distribution

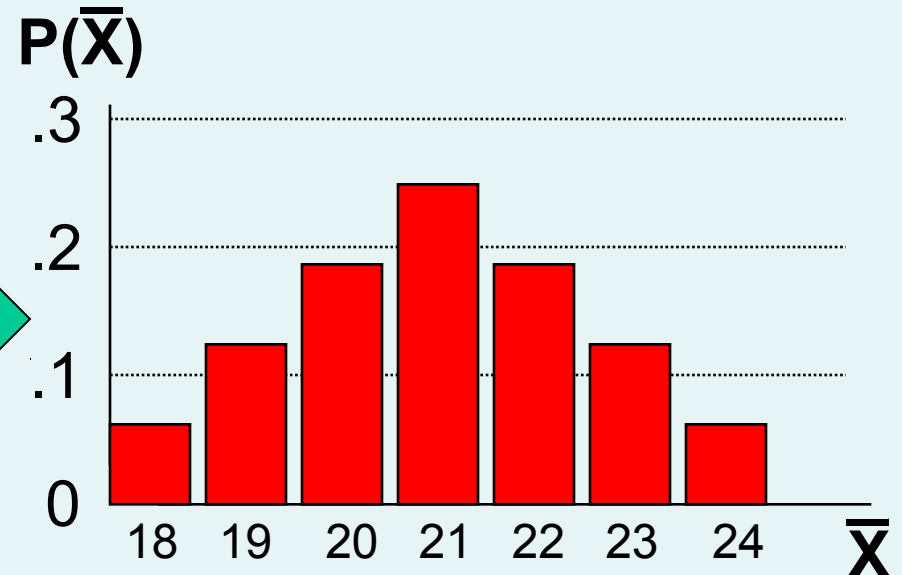
(continued)

## Sampling Distribution of All Sample Means

16 Sample Means

Sample Means Distribution

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24



(no longer uniform)





# Developing a Sampling Distribution

(continued)

Summary Measures of this Sampling Distribution:

$$\mu_{\bar{X}} = \frac{\sum \bar{X}_i}{N} = \frac{18 + 19 + 19 + \dots + 24}{16} = 21$$

$$\begin{aligned}\sigma_{\bar{X}} &= \sqrt{\frac{\sum (\bar{X}_i - \mu_{\bar{X}})^2}{N}} \\ &= \sqrt{\frac{(18 - 21)^2 + (19 - 21)^2 + \dots + (24 - 21)^2}{16}} = 1.58\end{aligned}$$

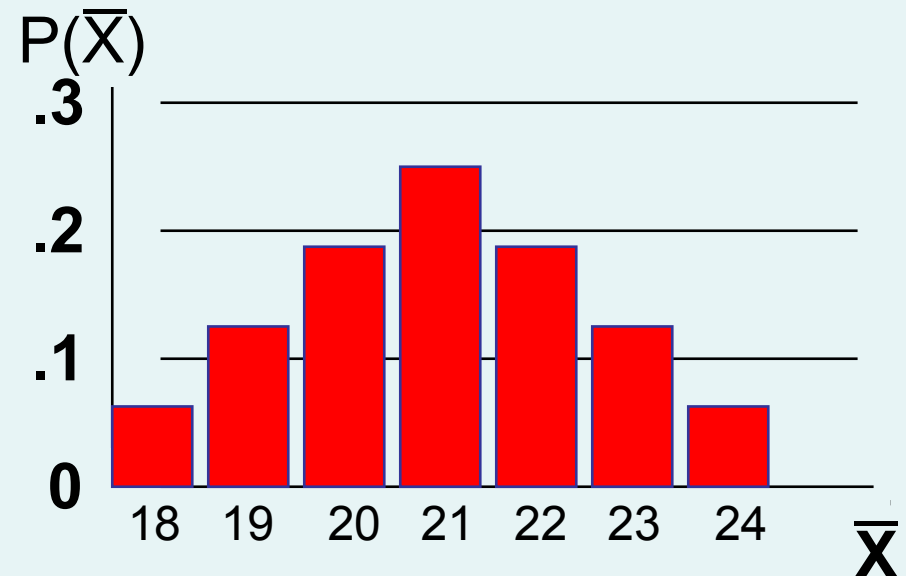
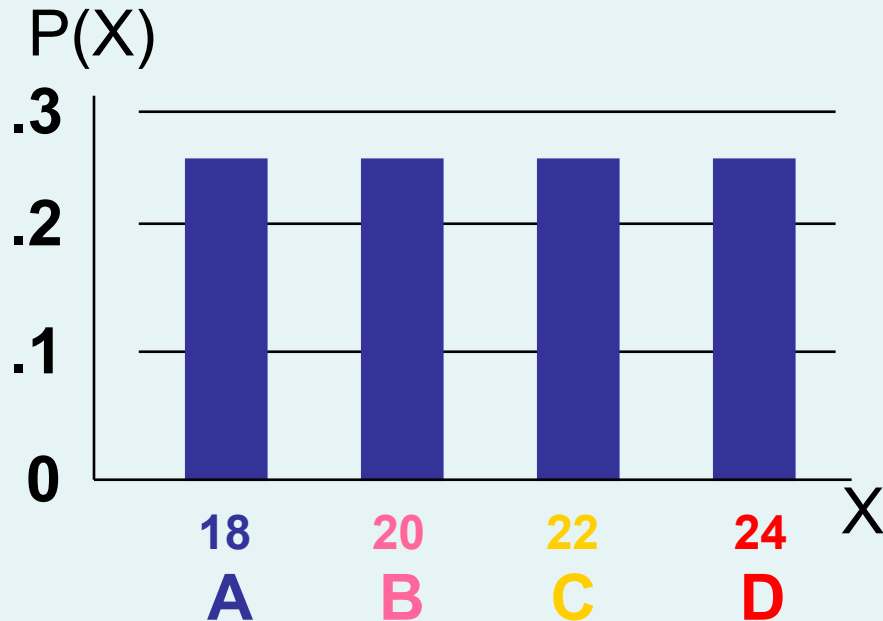
# Comparing the Population Distribution to the Sample Means Distribution

Population  
 $N = 4$

$$\mu = 21 \quad \sigma = 2.236$$

Sample Means Distribution  
 $n = 2$

$$\mu_{\bar{X}} = 21 \quad \sigma_{\bar{X}} = 1.58$$





# Sample Mean Sampling Distribution: Standard Error of the Mean

---

- Different samples of the same size from the same population will yield different sample means
- A measure of the variability in the mean from sample to sample is given by the **Standard Error of the Mean:**  
(This assumes that sampling is with replacement or sampling is without replacement from an infinite population)

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

- Note that the standard error of the mean decreases as the sample size increases



# Sample Mean Sampling Distribution: If the Population is Normal

---

- If a population is **normal** with mean  $\mu$  and standard deviation  $\sigma$ , the sampling distribution of  $\bar{X}$  is **also normally distributed** with

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$



# Z-value for Sampling Distribution of the Mean

---

- Z-value for the sampling distribution of  $\bar{X}$ :

$$Z = \frac{(\bar{X} - \mu_{\bar{X}})}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

where:

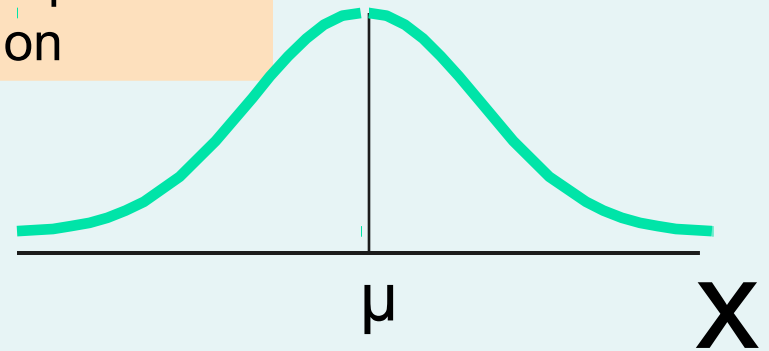
- $\bar{X}$  = sample mean
- $\mu$  = population mean
- $\sigma$  = population standard deviation
- $n$  = sample size

# Sampling Distribution Properties

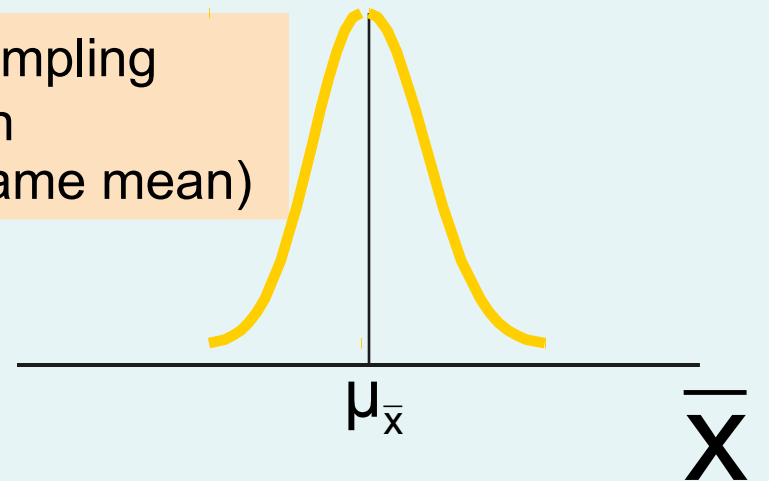
$$\mu_{\bar{X}} = \mu$$

(i.e.  $\bar{X}$  is unbiased)

Normal Population  
Distribution

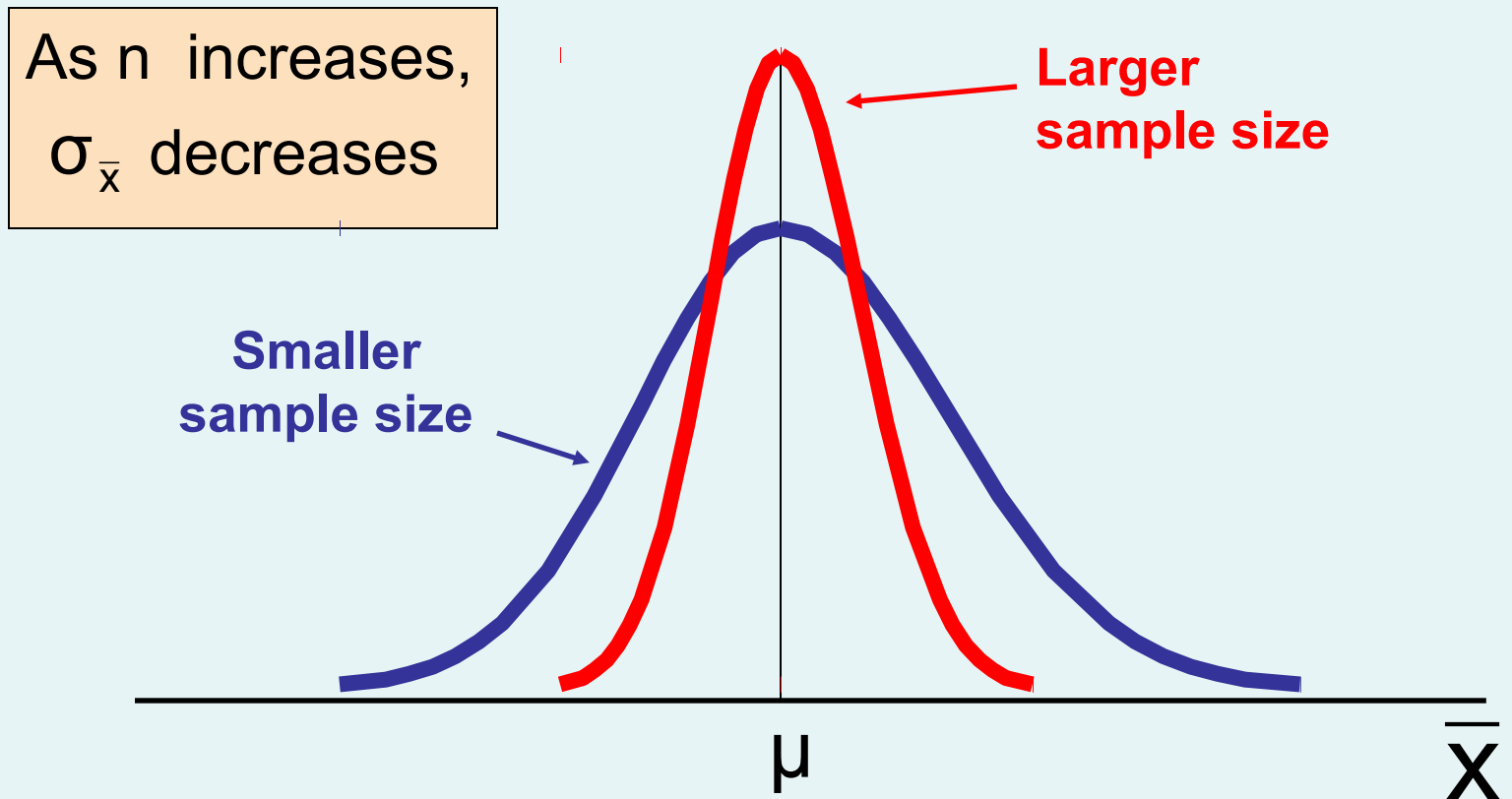


Normal Sampling  
Distribution  
(has the same mean)



# Sampling Distribution Properties

(continued)





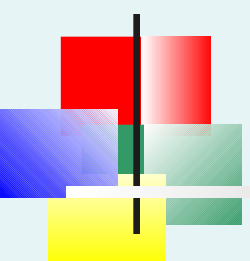
# Determining An Interval Including A Fixed Proportion of the Sample Means

---

Find a symmetrically distributed interval around  $\mu$  that will include 95% of the sample means when  $\mu = 368$ ,  $\sigma = 15$ , and  $n = 25$ .

- Since the interval contains 95% of the sample means 5% of the sample means will be outside the interval
- Since the interval is symmetric 2.5% will be above the upper limit and 2.5% will be below the lower limit.
- From the standardized normal table, the Z score with 2.5% (0.0250) below it is -1.96 and the Z score with 2.5% (0.0250) above it is 1.96.





# Determining An Interval Including A Fixed Proportion of the Sample Means

*(continued)*

---

- Calculating the lower limit of the interval

$$\bar{X}_L = \mu + Z \frac{\sigma}{\sqrt{n}} = 368 + (-1.96) \frac{15}{\sqrt{25}} = 362.12$$

- Calculating the upper limit of the interval

$$\bar{X}_U = \mu + Z \frac{\sigma}{\sqrt{n}} = 368 + (1.96) \frac{15}{\sqrt{25}} = 373.88$$

- 95% of all sample means of sample size 25 are between 362.12 and 373.88

# Sample Mean Sampling Distribution: If the Population is **not** Normal

- We can apply the **Central Limit Theorem**:
  - Even if the population is **not normal**,
  - ...sample means from the population **will be approximately normal as long as the sample size is large enough...**

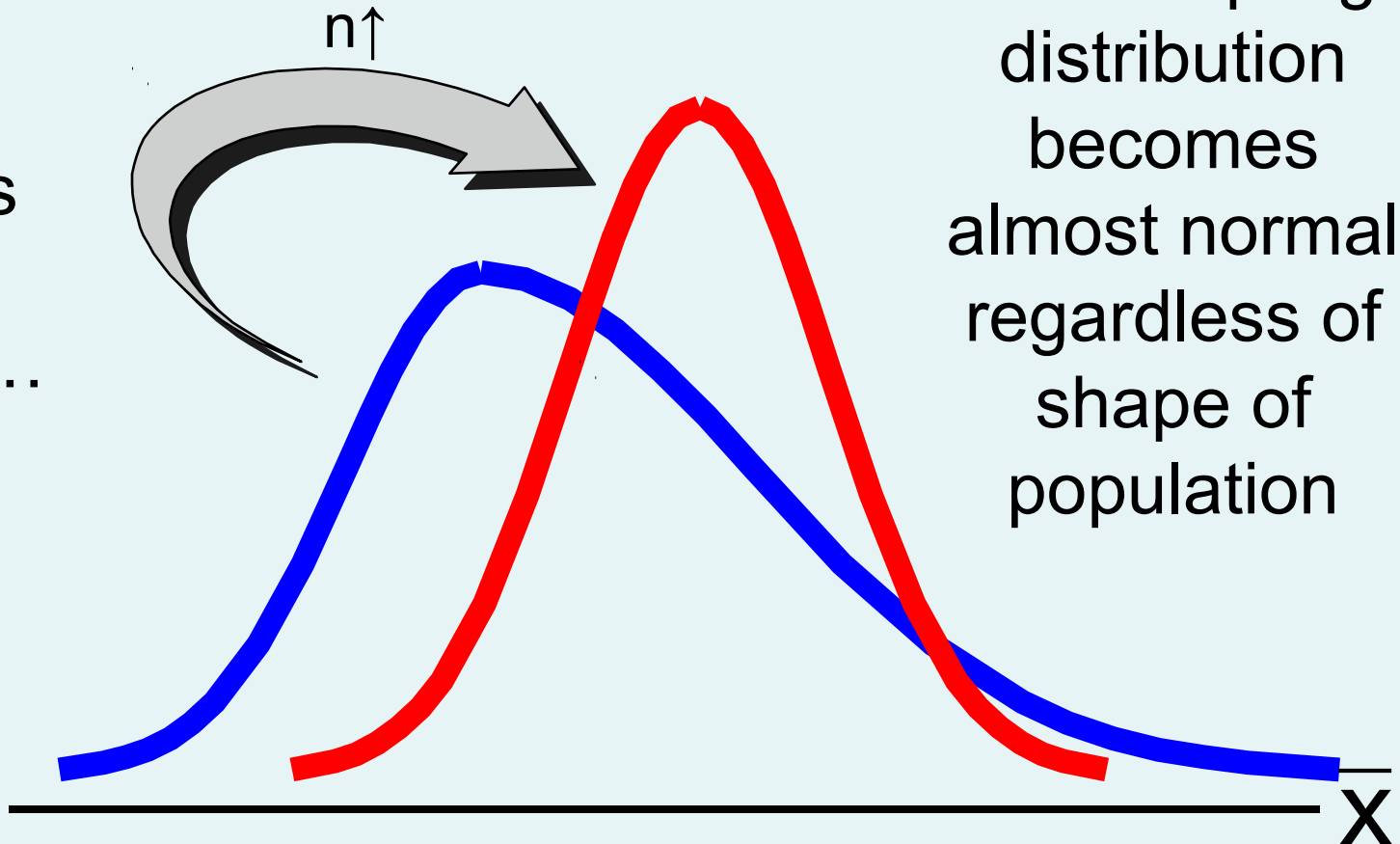
---

Properties of the sampling distribution:

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

# Central Limit Theorem

As the sample size gets large enough...



the sampling distribution becomes almost normal regardless of shape of population

# Sample Mean Sampling Distribution: If the Population is **not** Normal

(continued)

Sampling distribution  
properties:

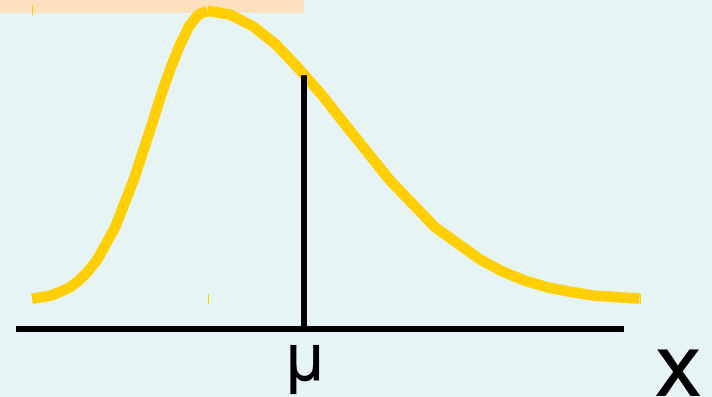
Central Tendency

$$\mu_{\bar{x}} = \mu$$

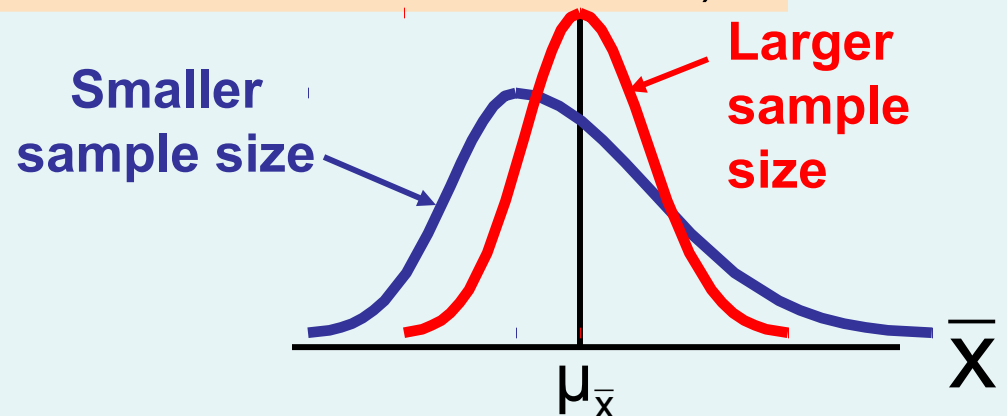
Variation

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

Population Distribution



Sampling Distribution  
(becomes normal as n increases)





# How Large is Large Enough?

---

- For most distributions,  $n > 30$  will give a sampling distribution that is nearly normal
- For fairly symmetric distributions,  $n > 15$
- **For normal population distributions, the sampling distribution of the mean is always normally distributed**



# Example

---

- Suppose a population has mean  $\mu = 8$  and standard deviation  $\sigma = 3$ . Suppose a random sample of size  $n = 36$  is selected.
- What is the probability that the **sample mean** is between 7.8 and 8.2?



# Example

(continued)

## Solution:

- Even if the population is not normally distributed, the central limit theorem can be used ( $n > 30$ )
- ... so the sampling distribution of  $\bar{x}$  is approximately normal
- ... with mean  $\mu_{\bar{x}} = 8$
- ...and standard deviation  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$

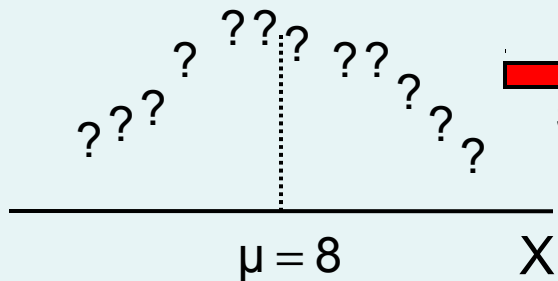
# Example

(continued)

Solution (continued):

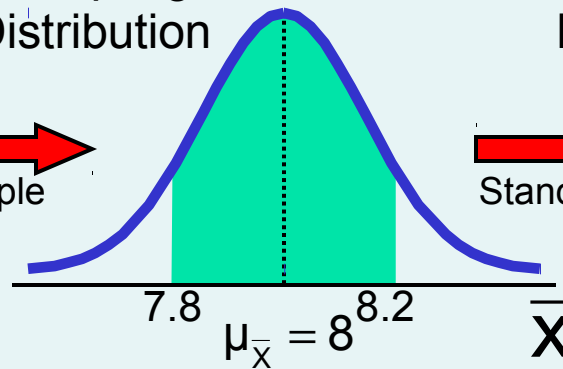
$$\begin{aligned} P(7.8 < \bar{X} < 8.2) &= P\left(\frac{7.8 - 8}{\frac{3}{\sqrt{36}}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{8.2 - 8}{\frac{3}{\sqrt{36}}}\right) \\ &= P(-0.4 < Z < 0.4) = \boxed{0.3108} \end{aligned}$$

Population  
Distribution



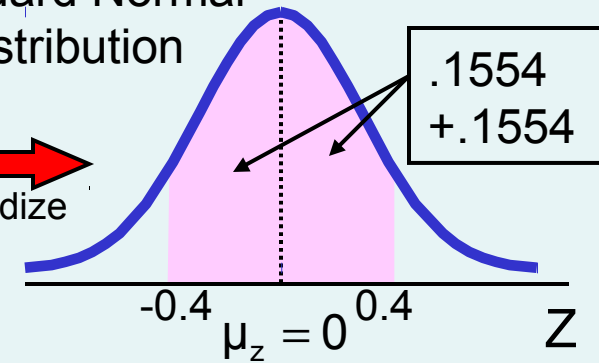
Sampling  
Distribution

Sample



Standard Normal  
Distribution

Standardize







# Population Proportions

---

$\pi$  = the proportion of the population having some characteristic

- Sample proportion ( $p$ ) provides an estimate of  $\pi$ :

$$p = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

- $0 \leq p \leq 1$
- $p$  is approximately distributed as a normal distribution when  $n$  is large

(assuming sampling with replacement from a finite population or without replacement from an infinite population)

# Sampling Distribution of $p$

- Approximated by a normal distribution if:

- $n\pi \geq 5$

and

$$n(1 - \pi) \geq 5$$

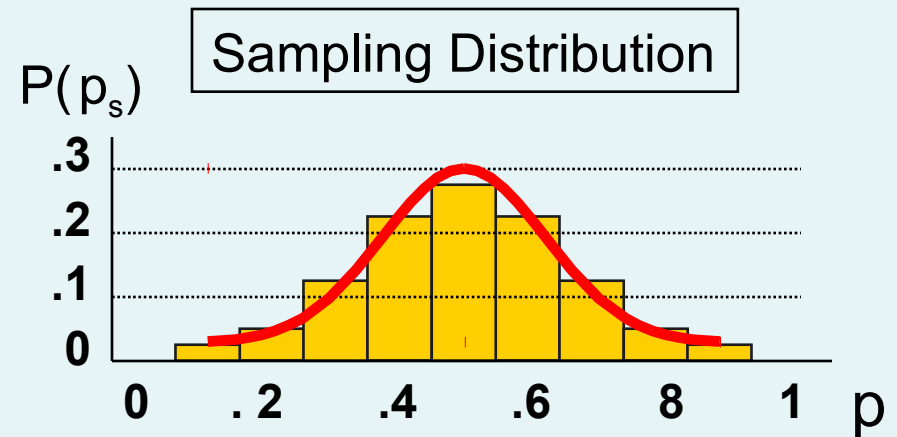
where

$$\mu_p = \pi$$

and

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

(where  $\pi$  = population proportion)





# Z-Value for Proportions

---

Standardize  $p$  to a  $Z$  value with the formula:

$$Z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}}$$



# Example

---

- If the true proportion of voters who support Proposition A is  $\pi = 0.4$ , what is the probability that a sample of size 200 yields a sample proportion between 0.40 and 0.45?
- i.e.: **if  $\pi = 0.4$  and  $n = 200$ , what is  $P(0.40 \leq p \leq 0.45)$  ?**



# Example

(continued)

- if  $\pi = 0.4$  and  $n = 200$ , what is  $P(0.40 \leq p \leq 0.45)$  ?

---

Find  $\sigma_p$ : 
$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.4(1-0.4)}{200}} = 0.03464$$

Convert to  
standardized  
normal:

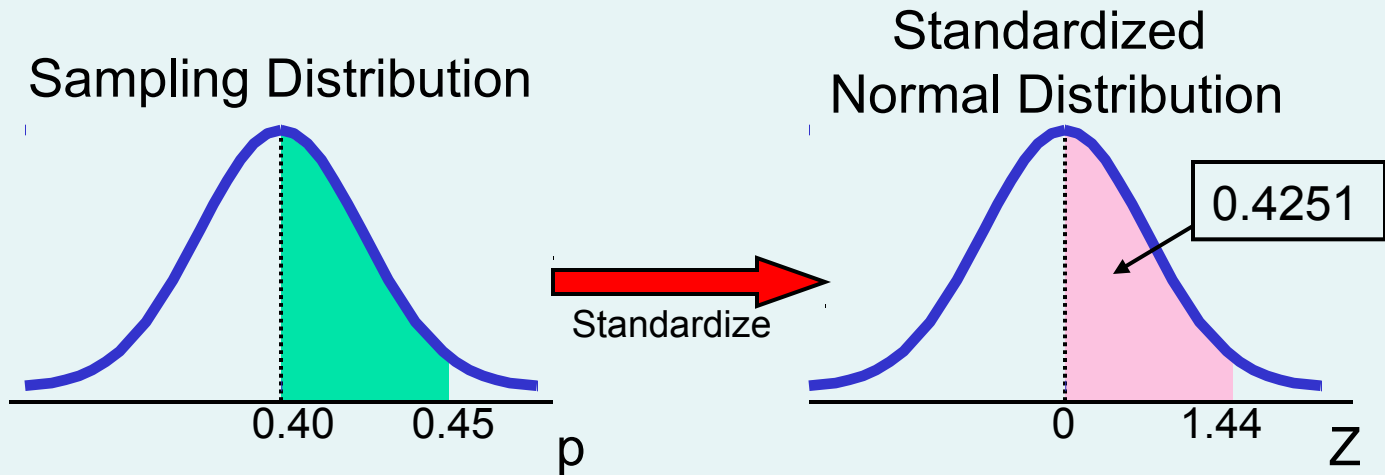
$$\begin{aligned} P(0.40 \leq p \leq 0.45) &= P\left(\frac{0.40 - 0.40}{0.03464} \leq Z \leq \frac{0.45 - 0.40}{0.03464}\right) \\ &= P(0 \leq Z \leq 1.44) \end{aligned}$$

# Example

(continued)

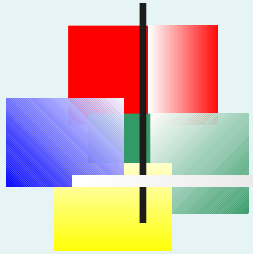
- if  $\pi = 0.4$  and  $n = 200$ , what is  $P(0.40 \leq p \leq 0.45)$  ?

Use standardized normal table:  $P(0 \leq Z \leq 1.44) = 0.4251$



# Types and Sources of Errors in Statistical Data

---





# Types of Errors

---

- In general, there are **two types of errors**:
  - a. non-sampling errors and
  - b. sampling errors.
- It is important to be aware of these errors, in particular non-sampling errors, so that they can be either minimised or eliminated from the data collected.





## Non-sampling errors

It is a general assumption in sampling theory that the true value of each unit in the population can be obtained and tabulated without any errors. In practice, this assumption may be violated due to several reasons and practical constraints. This results in errors in observations as well as in tabulation. Such errors which are due to factors other than sampling are called **non-sampling errors**.

The non-sampling errors are unavoidable in census and surveys. The data collected by complete enumeration in census is free from sampling error but would not remain free from non-sampling errors. The data collected through sample surveys can have both – sampling errors as well as non-sampling errors. Non-sampling errors arise because of the factors other than the inductive process of inferring about the population from a sample.

In general, the sampling errors decrease as the sample size increases whereas non-sampling error increases as the sample size increases.

In some situations, the non-sampling errors may be large and deserve greater attention than the sampling error.

In any survey, it is assumed that the value of the characteristic to be measured has been defined precisely for every population unit. Such a value exists and is unique. This is called the true value of the characteristic for the population value. In practical applications, data collected on the selected units are called **survey values** and differ from the true values. Such difference between the true and observed values is termed as **observational error** or **response error**. Such an error arises mainly from the lack of precision in measurement techniques and variability in the performance of the investigators.



# Sources of non-sampling errors

---

Non-sampling errors arise from:

- **defects in the sampling frame.**
- **failure to identify the target population.**
- **non response.**
- **responses given by respondents.**
- **data processing and**
- **reporting, among others.**



# Reducing non-sampling errors

---

- Can be minimised by adopting any of the following approaches:
  - using an up-to-date and accurate sampling frame.
  - careful selection of the time the survey is conducted.
  - planning for follow up of non-respondents.
  - careful questionnaire design.
  - providing thorough training and periodic retraining of interviewers and processing staff.



# Sampling Error

---

- Refer to the **difference between the estimate derived from a sample survey and the 'true' value that would result if a census of the whole population were taken under the same conditions.**
- These are errors that arise **because data has been collected from a part, rather than the whole** of the population.
- Because of the above, sampling errors are **restricted to sample surveys only** unlike non-sampling errors that can occur in both sample surveys and censuses data.



## Factors Affecting Sampling Error

---

It is affected by a number of factors including:

**a. sample size**

- In general, larger sample sizes decrease the sampling error, however this decrease is not directly proportional.
- As a rough rule of the thumb, **you need to increase the sample size fourfold to halve the sampling error but bear in mind that non sampling errors are likely to increase with large samples!**

**b. the sampling fraction**

- this is of lesser influence but as the sample size increases as a fraction of the population, the sampling error should decrease.



## Factors Affecting Sampling Error – cont'd

---

### c. the variability within the population.

- More variable populations give rise to larger errors as the samples or the estimates calculated from different samples are more likely to have greater variation.
- The effect of **variability within the population can be reduced by the use of stratification** that allows explaining some of the variability in the population.

### d. sample design.

- An efficient sampling design will help in reducing sampling error.



## Two types of Sampling Error

---

**Biased**  
Errors

- Selection of inappropriate method of sampling

**Unbiased**  
Errors

- Chance differences between members of the sample & the members of the remaining population



## Characteristics of the sampling error

---

- Generally decreases in magnitude as the sample size increases (but not proportionally).
- Depends on the variability of the characteristic of interest in the population.
- Can be accounted for and reduced by an appropriate sample plan.
- Can be measured and controlled in probability sample surveys.

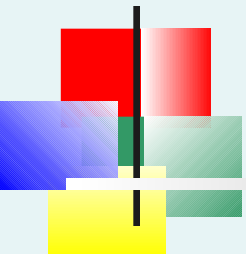




## Reducing sampling error

---

- If sampling principles are applied carefully within the constraints of available resources, sampling error can be kept to a minimum.



---

**Thank you!**