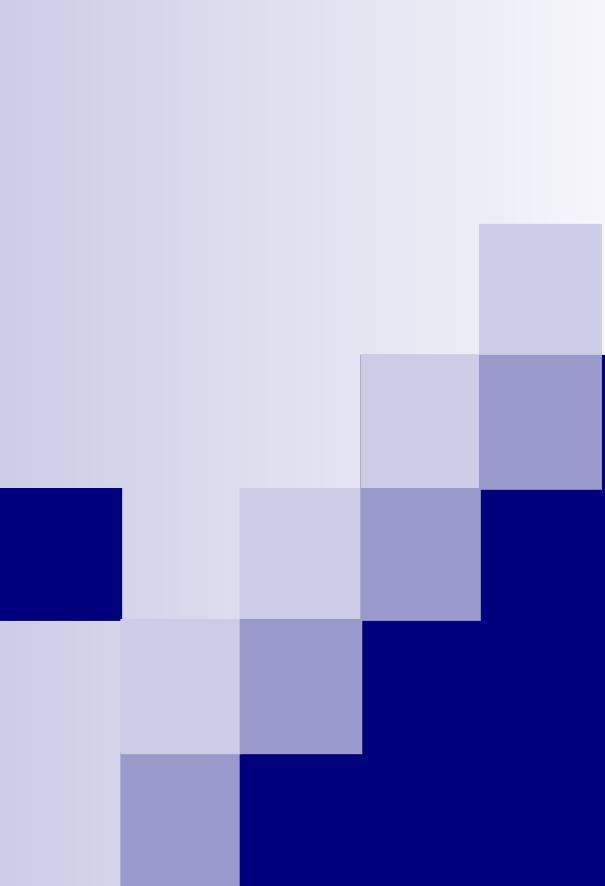




Sampling Distributi on

Ephrem Mannekulih (BSc, MSc)
**Biostatistics and Health
Informatics**

Sampling distribution

- Is the distribution of all *possible values* that can be assumed by some *sample statistic*, computed from samples of the same size *randomly* drawn from the same population.
- It is the bridge to link b/n descriptive (*sample statistic*) to inferential statistics (*Population parameter*).
- The concept of sampling distribution is key to the understanding of statistical inference.

.....cont'ed

- We consider **sample statistics** as **random variables** in sampling distribution

Example:

- **Probability distribution:** *Age* of individuals is a random variable
- **Sampling distribution:** *Mean age* is a random variable



Purpose of sampling distributions

- Sampling distributions serve two purposes
 - They allow us to answer probability questions about sample statistics
 - They provide the necessary theory for making statistical inference procedures valid.

Cont....

- Conclusions can not be drawn about values of population parameters based on **one** individual value.
- It should be based on sample statistic computed from adequate sample size.
- **Similarly**, we can not conclude about population parameters based on **single sample statistics**
- It should be based on sample statistic repeatedly computed from a random sample of the same size drawn from the same population



Why sampling distribution

- Conclusions can not be drawn about values of population parameters based on **one** sample statistic value.
- It should be based on sample statistic computed from adequate sample size.



Construction of sampling distributions

1. From a population of size **N**, randomly draw all possible samples of size **n**
2. Compute the statistic of interest for each sample
3. Create a frequency distribution of the statistic

Sampling with and without replacement

- **Sampling with replacement** is a method of sampling in which each member of the population selected for the sample is returned to the population before the next member is selected
- When sampling with replacement is employed, every member of the population is available at each draw.

Sampling with and without replacement

- **Sampling without replacement** is defined as a method of sampling in which the members of the sample are not returned to the population before subsequent members are selected.
- When sampling without replacement is employed, would lay selected sample aside until the entire sample is drawn.



**How do we know whether our
sample represents the
population???**



Main types of sampling distributions

- A. Distribution of the sample mean
- B. Distribution of the difference between two means
- C. Distribution of the sample proportion
- D. Distribution of the difference between two proportions

A. Sampling distribution of sample mean

- Suppose we have a population of size **N=4**, constituting the ages of four outpatients.

x, Age (years): 18, 20, 22, 24

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

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

Now consider all possible samples of size (n=2)

1 st	2 nd Observation			
Obs	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



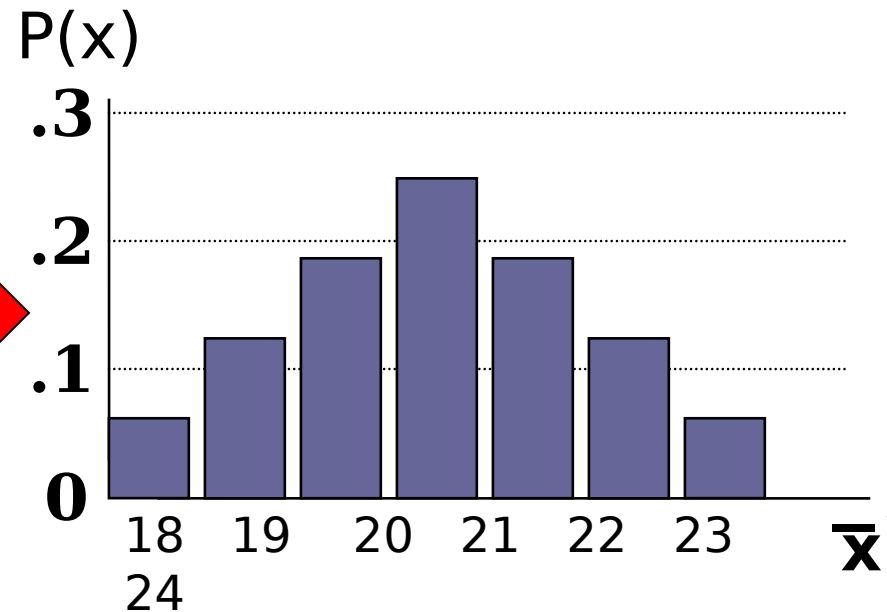
1 st	2 nd Observation			
Obs	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

- 16 Sample Means
- 16 possible samples (with replacement)

Sample Means Distribution

<i>Sample means</i>	<i>Freq</i>	<i>$P(\bar{x})$</i>
---------------------	-------------	--------------------------------

18	1	0.0625
19	2	0.1250
20	3	0.1875
21	4	0.2500
22	3	0.1875
23	2	0.1250
24	1	0.0625



Summary measures of mean distribution:

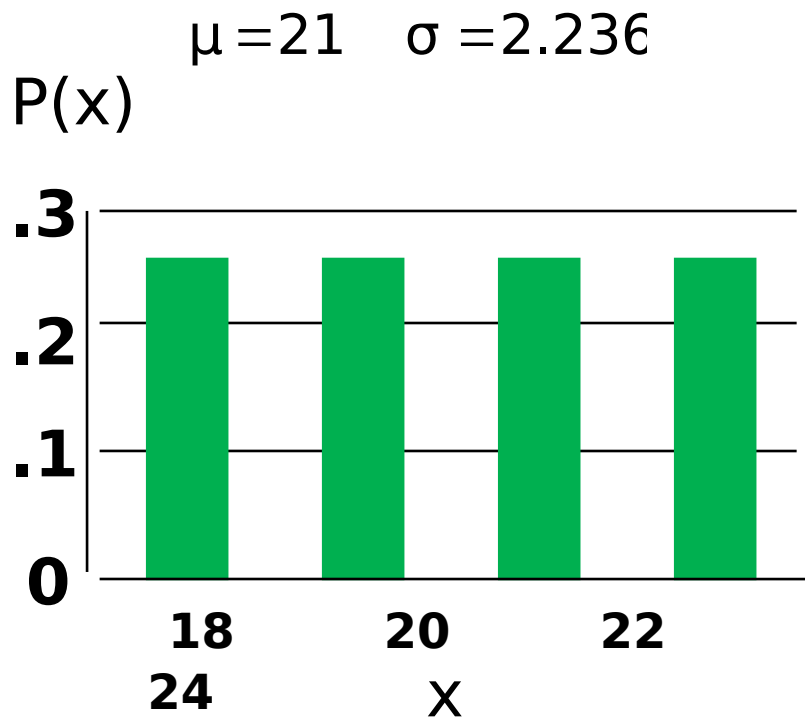
$$\mu_{\bar{x}} = \frac{\sum \bar{x}_i}{N} = \frac{18+19+21+\dots+24}{16} = 21$$

The mean and SD of the means

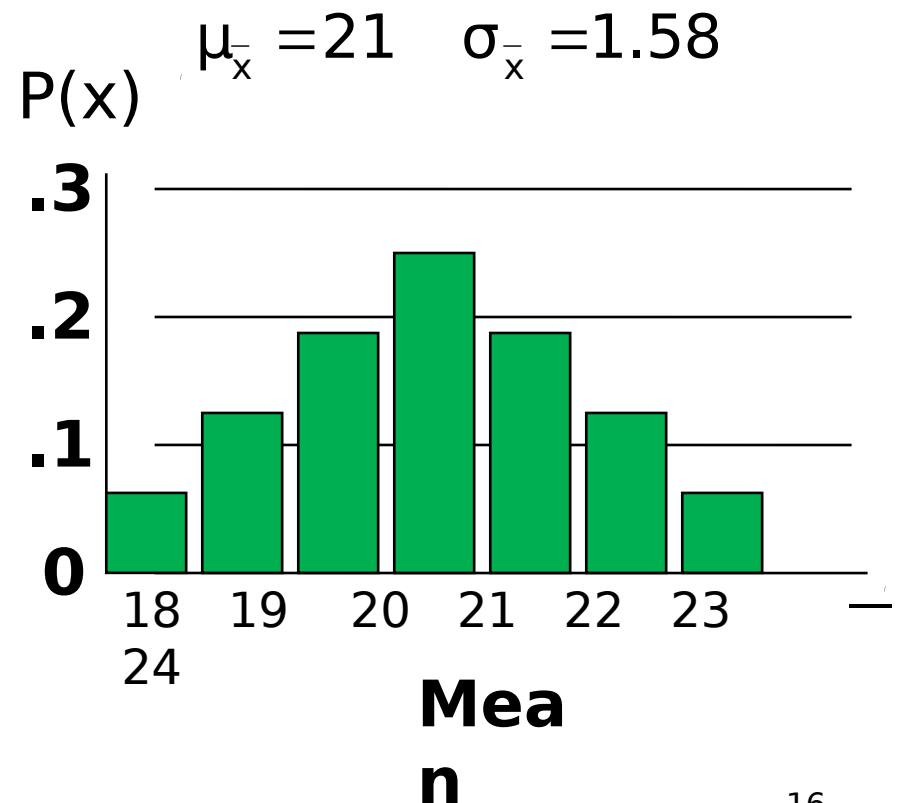
$$\begin{aligned}\sigma_{\bar{x}} &= \sqrt{\frac{\sum (x_i - \mu_{\bar{x}})^2}{N}} \\ &= \sqrt{\frac{(18-21)^2 + (19-21)^2 + \dots + (24-21)^2}{16}} = 1.58\end{aligned}$$

Comparing the population with its sampling distribution

Population ($N = 4$)



Sample means distribution ($n = 2$)



- The **mean** of the sampling distribution of $\bar{x}$ has the same value as the **mean** of the original population
- However, the **variance** is $\neq$ the original population variance; but is equal to the population variance **divided** by the sample size used to obtain sampling distribution.

- The square root of the sampling distribution variance is called **standard error of the mean** or, simply, **standard error**.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
- OR, the standard deviation of any **sample statistic** is called its **standard error**.

What is the difference between SE and SD?

- **Standard deviation** quantifies the amount of **variability among individuals** in a population or a sample
- **Standard error** quantifies the **variability among statistic of repeated samples** drawn from the population

What is the difference between SE and SD?

- **Standard deviation** quantifies the amount of **variability among individuals** in a population or a sample
- **Standard error** quantifies the **variability among statistic of repeated samples** drawn from the population
- ***When the SE is very small, it means that the sample best represents the population***



Sampling from Normally distributed population

If the sample is drawn from a normally distributed population,

- ❖ The distribution of mean is *normal*
- ❖ The population mean will be *equal* to the sample mean
- ❖ The variance of the population will be equal to the variance of the population divided by the sample size



Sampling distribution from non-normally distributed population:

- If the sample is drawn from non-normally distributed population, we depend on the mathematical theorem called “central limit theorem”

Central Limit Theorem

- Given a population of any **non-normal functional** form with a mean μ and finite variance σ^2 , the sampling distribution of the sample mean of size n from this population will have mean μ and variance σ^2/n and will be approximately normally distributed when the sample size is large.
- *Large sample size?????? Often taken as $n \geq 30$*
 - But the normality assumption needs to be always checked up regardless of the sample size since the *sampling technique is not as random as it is theorized*

If the population is not normal

Sampling
distribution
properties:

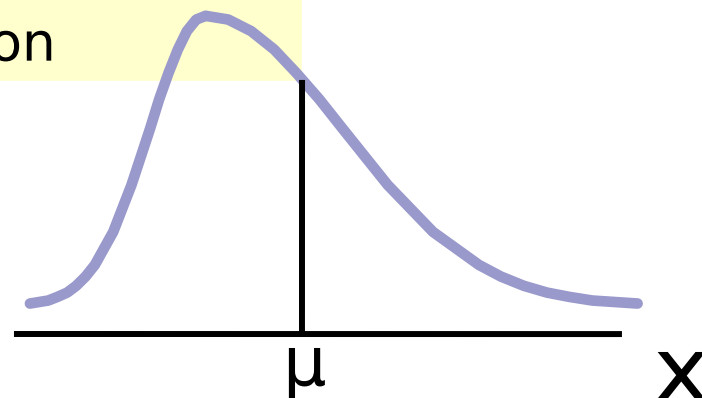
Central
Tendency

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

Variatio
n

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

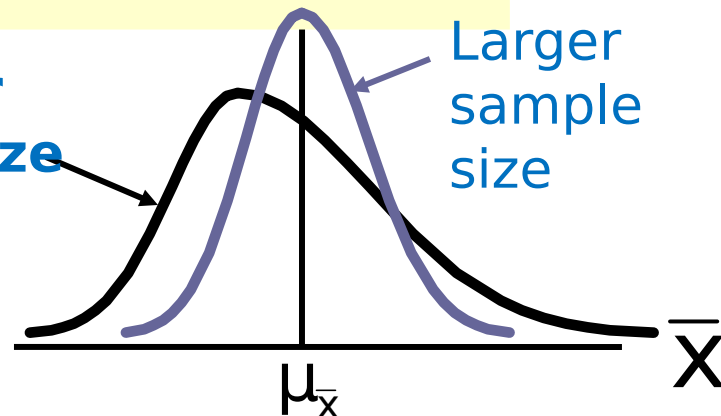
Population
Distribution



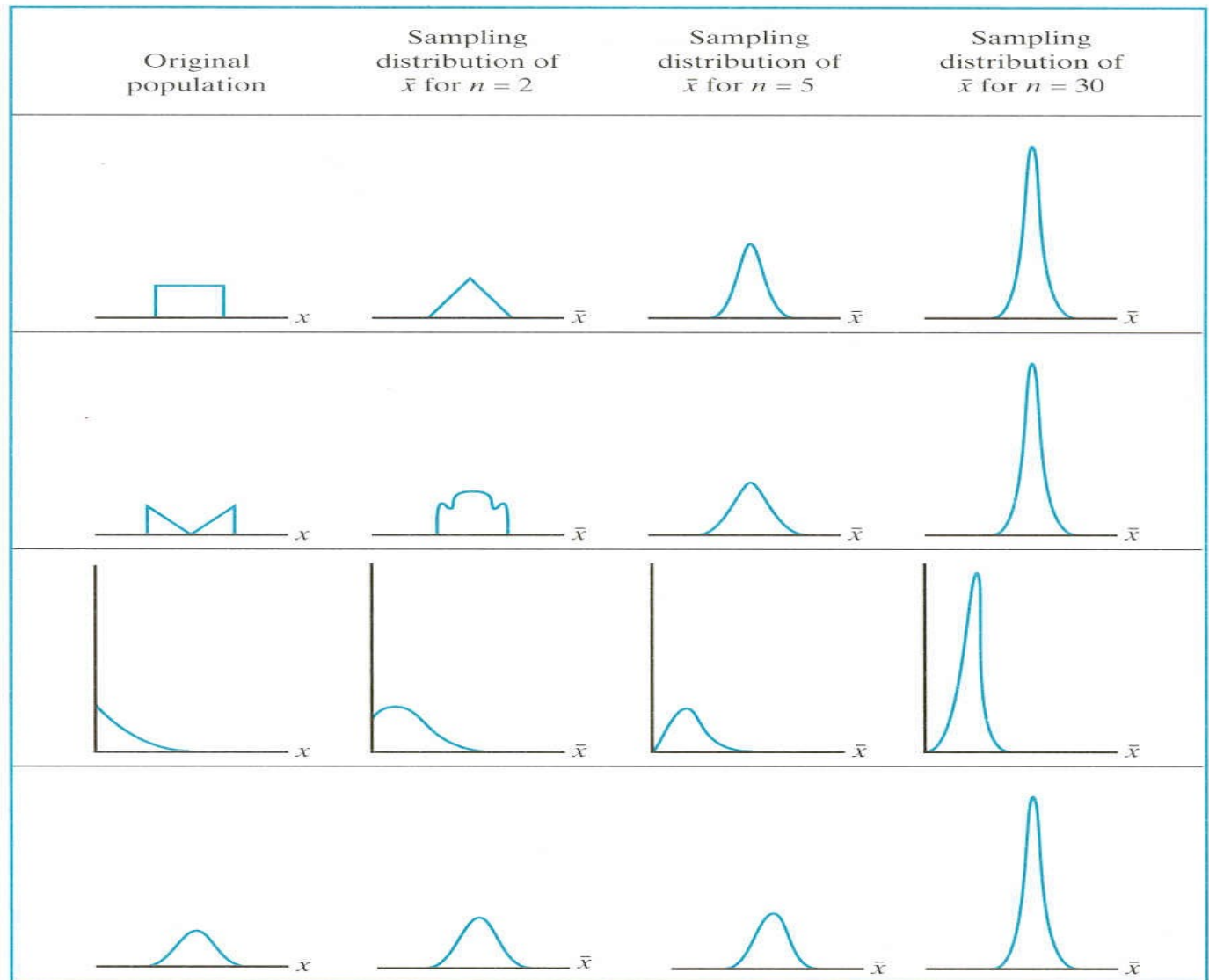
Sampling Distribution
(becomes normal as n
increases)

**Smaller
sample size**

**Larger
sample size**



Sampling distribution of $\bar{x}$ for different population and different sizes.



- 
- How to make probability statements about the population parameters?

- We couldn't deal with all types of measurements units and levels to integrate (sum) all points between a given interval
 - Tedious mathematics (calculus)
- Transformation of the sample statistic into a unit less measurement using **Z- transformation formula:**

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

Single mean distribution

Example

- Suppose it is known that in a certain large human population cranial length is approximately normally distributed with a mean of 185.6 mm and a standard deviation of 12.7 mm

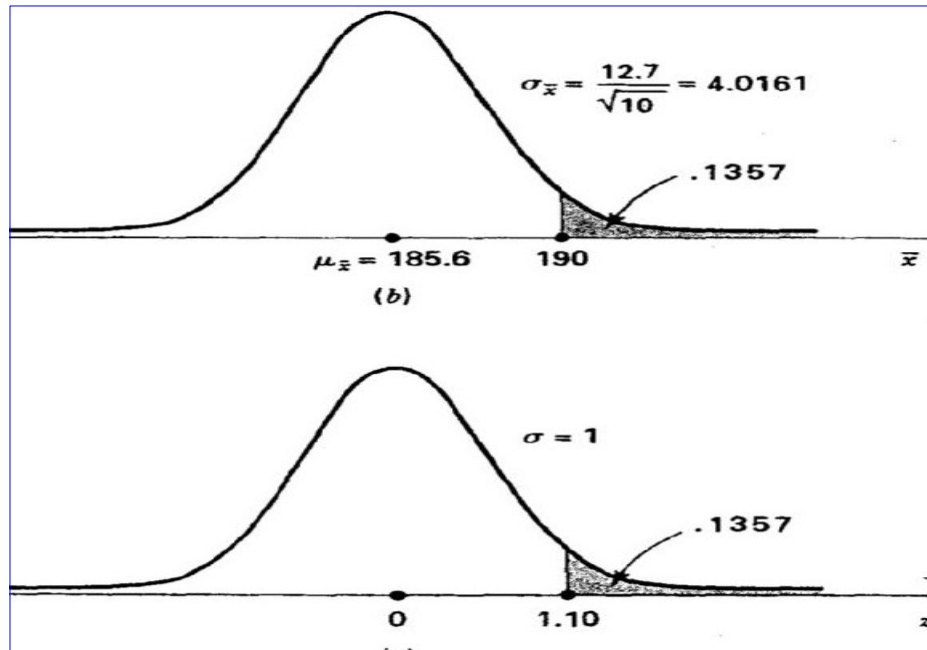
What is the probability that a random sample size of 10, from this population will have a mean greater than 190?

Given: $\mu = 185.6 \text{ mm}$

$\sigma = 12.7 \text{ mm}$

$n = 10$

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} = \frac{190 - 185.6}{12.7/\sqrt{10}} = 1.10$$



$$P(\bar{x} > 190\text{mm}) = P(Z > 1.10) = 1 - \text{pr}(Z < 1.10)$$

□ The probability of getting $(Z > 1.10) = 1 - (Z < 1.10) = 1 - 0.8643 = \mathbf{0.1357}$

□ The probability of getting cranial length > 190 mms out of 10 patients selected randomly from the population is 0.1357.

Example

- If the mean and standard deviation of serum iron for healthy men values in the population are 120 and 15 micrograms per 100 ml, respectively, what is the probability that a random sample of 50 normal men will yield a mean between 115 and 125 micrograms per 100 ml?

$$\begin{aligned}P(115 \leq \bar{x} \leq 125) &= P\left(\frac{115 - 120}{2.12} \leq z \leq \frac{125 - 120}{2.12}\right) \\&= P(-2.36 \leq z \leq 2.36) \\&= .9909 - .0091 \\&= .9818\end{aligned}$$

2) Distribution of the difference of two sample means

- ❑ Important to compare two population means (**comparative studies**)
- ❑ Are the two population means different?
- ❑ If yes by how much they differ?
- ❑ For example, mean serum cholesterol level for **sedentary office workers and daily laborers**

- It is generally assumed that the two populations are **normally distributed**
- For sampling from non-normal populations, **large samples** are recommended by the application of the **central limit theorem**
- Plotting sample differences $\bar{x}_1 - \bar{x}_2$ against its frequency gives a normal distribution with mean equal to $\mu_1 - \mu_2$

- The **variance** of the distribution of the sample differences is:

$$= (\sigma_1^2/n_1) + (\sigma_2^2/n_2)$$

- Thus, the **standard error** of the difference between sample means is:

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{\delta^2_1}{n_1} + \frac{\delta^2_2}{n_2}}$$

- The Z- transformation will be an extension for the distribution with single mean:

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

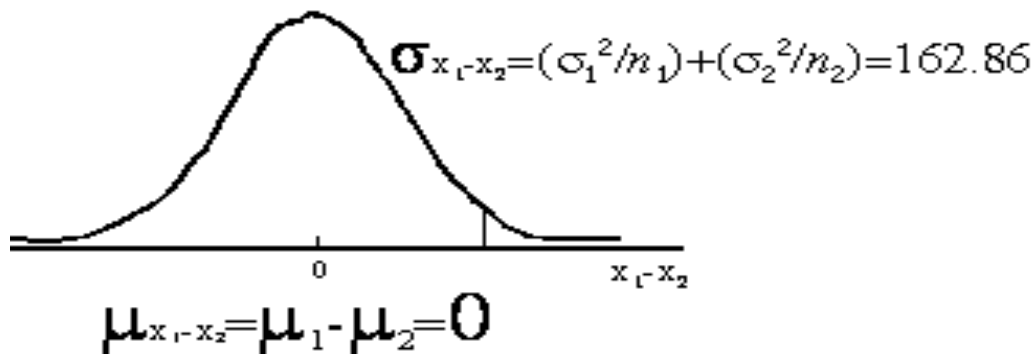
Example

- In a study of annual family expenditures for general health care, two populations were surveyed with the following results:

$$\begin{array}{lll} \text{Population 1: } n_1 = 40; & \bar{X}_1 & = \$346; \\ \sigma_1^2 = 2800 & \bar{X}_2 & \end{array}$$

$$\begin{array}{lll} \text{Population 2: } n_2 = 35; & & = \$300; \\ \sigma_2^2 = 3250 & \bar{X}_1 - \bar{X}_2 & \end{array}$$

- What is the probability of obtaining sample results as large as or larger than those shown if there is no difference in the means of the two populations?



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

$$z = \frac{(346 - 300) - (0)}{\sqrt{\frac{2800}{40} + \frac{3250}{35}}} = 3.60$$

- This gives the probability $P(Z > 3.60) = 1 - p(Z < 3.60) = 1 - 0.9998 = 0.0002$
- The probability that $\bar{x}_1 - \bar{x}_2$ is as large as 346 is **0.0002**

Example 2

- Suppose it has been established that for a certain type of client the average length of a home visit by a public health nurse is 45 minutes with a standard deviation of 15 minutes, and that for a second type of client the average home visit is 30 minutes long with a standard deviation of 20 minutes. If a nurse randomly visits 35 clients from the first and 40 from the second population, what is the probability that the average length of home visit is by 20 or

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{20 - 15}{\sqrt{16.4286}} = \frac{5}{4.0532} = 1.23$$

$$1 - .8907 = .1093.$$

3) Distribution of the sample proportion

- Population proportion (p) = the proportion of population having some characteristic
$$\bar{p} = \frac{x}{n} = \frac{\text{number of successes in the sample}}{\text{sample size}}$$

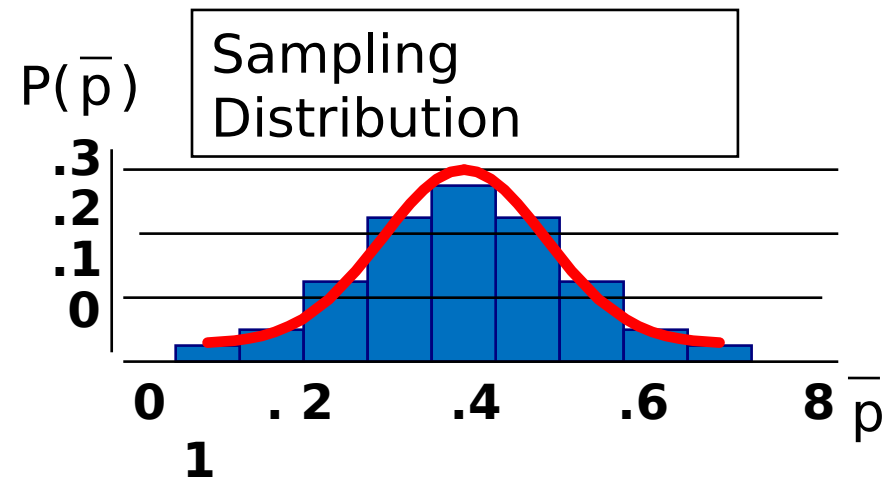
Characteristics:

- ✓ Mean of the distribution will be same to **population proportion**
- ✓ The shape of sampling distribution will be **approximately normal** if **large sample** is taken
$$np \geq 5$$
$$n(1-p) \geq 5$$
- ✓ SE of the sampling distribution $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$

- Approximation by a normal distribution if:

$$np \geq 5$$

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



Where $\mu_{\bar{p}} = p$ and $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$

(where p = population proportion)

- To establish probability statement about population proportion, the following Z-transformation formula can be used:

$$z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

TIP

- If sampling is **without replacement** and $n > 5\%$ of the population size, then we must use the finite population correction formula (

FPC fac

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$

Example

- A study on use of prenatal care among low income women, it was found that only 51% of these women had adequate prenatal care. If 200 women from this population were drawn at random,
- what is the probability that less than 45% will have received adequate prenatal care?


$$z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$z = \frac{0.45 - 0.51}{\sqrt{\frac{0.51 \times 0.49}{200}}} = -1.70$$

- The area to the left of -1.70 under the standard normal curve is 0.0446
- Therefore, $P(\bar{p} < 0.45) = P(Z < -1.70) = 0.0446$

Exercise

- It is known that 35 percent of the members of a certain population suffer from one or more chronic diseases. What is the probability that in a sample of 200 subjects drawn at random from this population 80 or more will have at least one chronic disease?

- Given: $p = 0.35$, $n = 200$, $\bar{p} = 80/200 = 0.4$

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

$$\bar{p} = \frac{x}{n} = \frac{\text{number of successes in the sample}}{\text{sample size}}$$



4) Distribution of the difference of two sample proportions

- If independent random samples of size n_1 and n_2 are drawn from two populations, where the proportions of observations with the characteristics of interest in the two populations are p_1 and p_2 respectively

$$\hat{p}_1 \quad \hat{p}_2$$

- The distribution of sample proportions $\hat{p}_1$ and $\hat{p}_2$ is approximately normally distributed with mean proportion difference:

$$SE(\bar{p}_1 - \bar{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- And SD:

- The Z- transformation will be an extension for the distribution with single proportion:

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}}$$

Example

- A national survey reported that 28% of white subjects reported they had experienced lower back pain three months prior to the survey. Among subjects of Hispanic origin, 21% reported lower back pain three months prior to the survey.

What is the probability that independent random samples $\hat{p}_1$ of $\hat{p}_2$ size 100 drawn from each of the populations will yield a value of as large as 0.10 or more?

$$n_1 = 100 \quad n_2 = 100$$

$$p_1 = 0.28 \quad p_2 = 0.21$$

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

$$z = \frac{0.10 - 0.07}{\sqrt{0.003675}} = 0.49$$

- It is found that the area under the standard normal curve lies to the right of $Z = 0.49$ is $1 - 0.6879 = 0.3121$
- Therefore, the probability of observing a difference as large as 0.10 or more is **0.3121**

Exercise

- Researchers found that among U.S. adults ages 75 or older, 34 % had lost all their natural teeth and for U.S. adults ages 65–74, 26 % had lost all their natural teeth. If a random sample of 200 adults ages 65–74 and an independent random sample of 250 adults ages 75 or older are drawn from these populations, find the probability that the difference in percent of total natural teeth loss is less than .2420.

$$z = \frac{.05 - (.08)}{\sqrt{.00186}} = -.70$$

.2420.