Chapter 5:

Probabilistic Features of the Distributions of Certain Sample Statistics

5.1 Introduction

5.2 sampling Distribution

5.3 Distribution of the Sample Mean

5.4 Distribution of the Difference Between Two Sample Means

5.5 Distribution of the Sample Proportion

5.6 Distributions of the difference between two sample proportions

Introduction

In this Chapter we will discuss the probability distributions of some statistics.

As we mention earlier, a statistic is a measure computed from the random sample. As the sample values vary from sample to sample, the value of the statistic varies accordingly.

A statistic is a random variable; it has a probability distribution, a mean and a variance.

5.2 sampling Distribution :

The probability distribution of a statistic is called the sampling distribution of that statistic.

The sampling distribution of the statistic is used to make statistical inference about the unknown parameter.

5.3 Distribution of the Sample Mean ($\bar{X}$) :

Suppose that we have a population with mean μ and variance σ^2.
Suppose that X_1 , X_2 ,…, X_n is a random sample of size (n) selected randomly from this population.
We know that the sample mean is:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_n}{n}$$

Suppose that we select several random samples of size n =5 :

	1st sample	2nd sample	3rd sample	...	Last sample
Sample value	28	31	14	...	17
	30	20	31	...	32
	34	31	25	...	29
	34	40	27	...	31
	17	28	32	...	30
Sample Mean $\bar{x}$	28.4	29.9	25.8	...	27.8

- The value of the sample mean $\bar{X}$ varies from random sample to another.
- The value of $\bar{X}$ is random and it depends on the random sample.
- The sample mean $\bar{X}$ is a random variable.
- The probability distribution of $\bar{X}$ is called the sampling distribution of the sample mean $\bar{X}$.

- Questions :
  - What is the sampling distribution of the sample mean $\bar{X}$?
  - What is the mean of the sample mean $\bar{X}$?
  - What is the variance of the sample mean $\bar{X}$?

Result (1) : (Mean & variance of $\bar{X}$)

If X_1 , X_2 ,…, X_n is a random sample of size (n) from any distribution with mean μ and variance σ^2 , then:

1. The mean of $\bar{X}$ is : $\mu_{\bar{X}} = \mu$
2. The variance of $\bar{X}$ is : $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$
3. The standard deviation of $\bar{X}$ is called the standard error and is defined by : $\sigma_{\bar{X}} = \sqrt{\sigma_{\bar{X}}^2} = \frac{\sigma}{\sqrt{n}}$

Result (2): (Sampling from normal population)

If $X_1, X_2, \ldots, X_n$ is a random sample of size (n) from a normal population with mean μ and variance σ^2, that is Normal(μ, σ^2), then the sample mean $\bar{X}$ has a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$, that is :

1. $\bar{X} \sim$ Normal $(\mu, \frac{\sigma^2}{n})$
2. $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim$ Normal (0,1)

We use this result when sampling from normal distribution with known variance σ^2.

Result (3): (Central Limit Theorem: Sampling from Non-normal population)

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size (n) from a non-normal population with mean μ and variance σ^2. if the sample size n is large (n ³30), then the sample mean $\bar{X}$ has approximately a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$, that is :

1. $\bar{X} \approx$ Normal $(\mu, \frac{\sigma^2}{n})$ (approximately)
2. $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \approx$ Normal (0,1) (approximately)

Notes :

- "$\approx$" means "approximately distributed".
- We use this result when sampling from non-normal distribution with known variance σ^2 and with large sample size.

Result (4): (used when σ^2 is unknown + normal distribution)

If $X_1, X_2, \ldots, X_n$ is a random sample of size (n) from a normal distribution with mean μ and variance σ^2, that is Normal(μ, σ^2), then the statistic :

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

Notation : degrees of freedom = df = v

Application:

Example: (Sampling distribution of the sample mean)

Suppose that the time duration of a minor surgery is approximately normally distributed with mean equal to 800 seconds and a standard deviation of 40 seconds. Find the probability that a random sample of 16 surgeries will have average time duration of less than 775 seconds.

Solution:

X= the duration of the surgery
μ=800 , σ = 40 ,σ^2= 1600 , n=16

X~N(800, 1600)

Calculating mean, variance, and standard error (standard deviation) of the sample mean $\bar{X}$:

The mean of $\bar{X} = \mu_{\bar{X}} = \mu = 800$

The variance of $\bar{X} = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{1600}{16} = 100$

The standard deviation of $\bar{X} = \sqrt{\sigma^2_{\bar{X}}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{16}} = 10$

Using the central limit theorem $\bar{X}$ has a normal distribution with mean $\mu_{\bar{X}}$ =800 and variance $\sigma^2_{\bar{X}}$ =100

That is :

$$\bar{X} \sim \text{Normal } (800 , 100)$$

$$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X}-800}{10} \sim \text{Normal } (0,1)$$

The probability that a random sample of 16 surgeries will have average time duration of less than 775 seconds equals to :

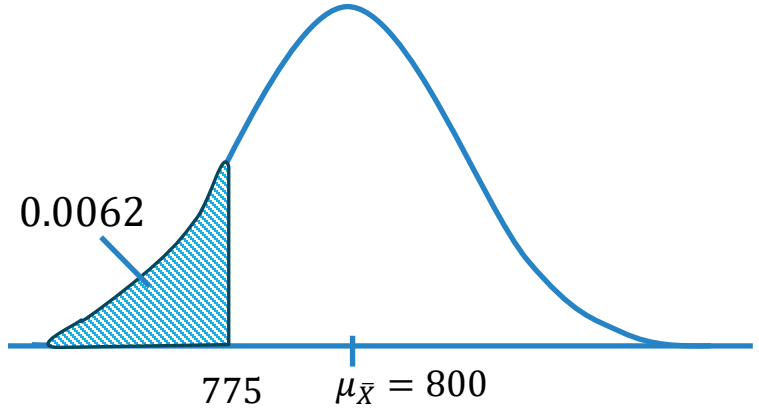
$$P(\bar{X} < 775) = P\left(\frac{\bar{X}-800}{10} < \frac{775-800}{10}\right) = P(Z < -2.5) = 0.00621$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N(800,100)$$

$$\mu_{\bar{X}} = 800$$

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = 100$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 800$$



Example :

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

Solution :

X= the serum iron value

μμ=120 , σ = 15 ,σ^2= 225 , n=50 (large)

X ≈ N(120, 225)

Calculating mean, variance, and standard error (standard deviation) of the sample mean $\bar{X}$:

The mean of $\bar{X} = \mu_{\bar{X}} = \mu = 120$

The variance of $\bar{X} = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{225}{50} = 4.5$

The standard deviation of $\bar{X} = \sqrt{\sigma^2_{\bar{X}}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{50}} = 2.12$

Using the central limit theorem $\bar{X}$ has a normal distribution with mean $\mu_{\bar{X}} = 120$ and variance $\sigma^2_{\bar{X}} = 4.5$

That is :

$$\bar{X} \sim \text{Normal}(120, 4.5)$$

$$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X}-120}{2.12} \sim \text{Normal}(0,1)$$

The probability that a random sample of size 50 men will yield a mean between 115 and 125 microgram/100ml equals to :

$$P(115 < \bar{X} < 125) = P\left(\frac{115-120}{2.12} < \frac{\bar{X}-800}{10} < \frac{125-120}{2.12}\right)$$

$$= P(-2.36 < Z < 2.36)$$

$$= P(Z < 2.36) - P(Z < -2.36)$$

$$= 0.99086 - 0.00914$$

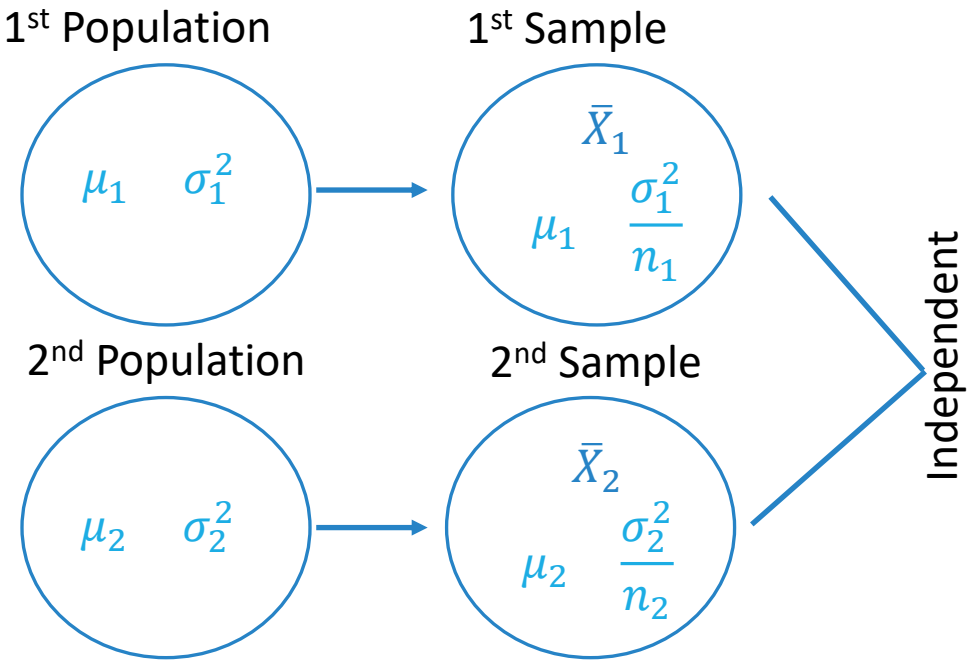
$$= 0.98172$$

5.4 Distribution of the Difference Between Two Sample Means ($\bar{X}_1 - \bar{X}_2$) :

Suppose that we have two populations:

- 1-st population with mean μ_1 and variance σ_1^2 .
- 2-nd population with mean μ_2 and variance σ_2^2 .
- We are interested in comparing μ_1 and μ_2, or equivalently, making inferences about the difference between the means $(\mu_1 - \mu_2)$.
- We independently select a random sample of size n_1 fromthe 1-st population and another random sample of size n_2 from the 2-nd population .
- Let $\bar{X}_1$ and S_1^2 be the sample mean and the sample variance of the 1-st sample .
- Let $\bar{X}_2$ and S_2^2 be the sample mean and the sample variance of the 2-nd sample .

•The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is used to make inferences about $\mu_1 - \mu_2$.



The sampling distribution of $\bar{X}_1 - \bar{X}_2$:

Result :

The mean , variance and the standard deviation of $\bar{X}_1 - \bar{X}_2$ are :

1. The mean of $\bar{X}_1 - \bar{X}_2$ is : $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$
2. The variance of $\bar{X}_1 - \bar{X}_2$ is : $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
3. The standard deviation of $\bar{X}$ is called the standard error and is defined by : $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1 - \bar{X}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

- Note: Square roots distribute over multiplication or division, but not addition or subtraction.

$$\sqrt{a+b} \neq \sqrt{a} + \sqrt{b}$$

- In general: Z= (value - Mean)/ Standard deviation

Result :

If the two random samples were selected from normal distributions (or non-normal distributions with large sample sizes) with known variances σ_1^2 and σ_2^2 , then the difference between the sample means $\bar{X}_1 - \bar{X}_2$ has a normal distribution with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$, that is :

- $\bar{X}_1 - \bar{X}_2$ ~ Normal $(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$
- $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ ~ Normal (0,1)

Example:

Suppose it has been established that for a certain type of client (type A) the average length of a home visit by a public health nurse is 45 minutes with standard deviation of 15 minutes, and that for second type (type B) of client the average home visit is 30 minutes long with standard deviation of 20 minutes. If a nurse randomly visits 35 clients from the first type and 40 clients from the second type, what is the probability that the average length of home visit of first type will be greater than the average length of home visit of second type by 20 or more minutes ?

Solution :

For the first type : μ_1=45 , σ_1=15 , σ_1^2=225 , n_1=35 (large)

For the second type : μ_2=30 , σ_2=20 , σ_2^2=400 , n_2=40 (large)

The mean, the variance and the standard deviation of $\bar{X}_1 - \bar{X}_2$ are:

1. The mean of $\bar{X}_1 - \bar{X}_2$ is : $\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2$= 45-30=15
2. The variance of $\bar{X}_1 - \bar{X}_2$ is : $\sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{225}{35} + \frac{400}{40}$ = 16.4286
3. The standard deviation of $\bar{X}$ is called the standard error and is defined by : $\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ $= \sqrt{16.4286}$ = 4.0532

The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is :

$\bar{X}_1 - \bar{X}_2$ ~ Normal (15 , 16.4286)

Z = $\frac{(\bar{X}_1 - \bar{X}_2) - 15}{\sqrt{16.4286}}$ ~ Normal (0,1)

the probability that the average length of home visit of first type will be greater than the average length of home visit of second type by 20 or more minutes equals to :

$P(\bar{X}_1 > \bar{X}_2 + 20)$ =

$= P(\bar{X}_1 - \bar{X}_2 > 20)$

$= P(Z > \frac{20-15}{4.0532})$

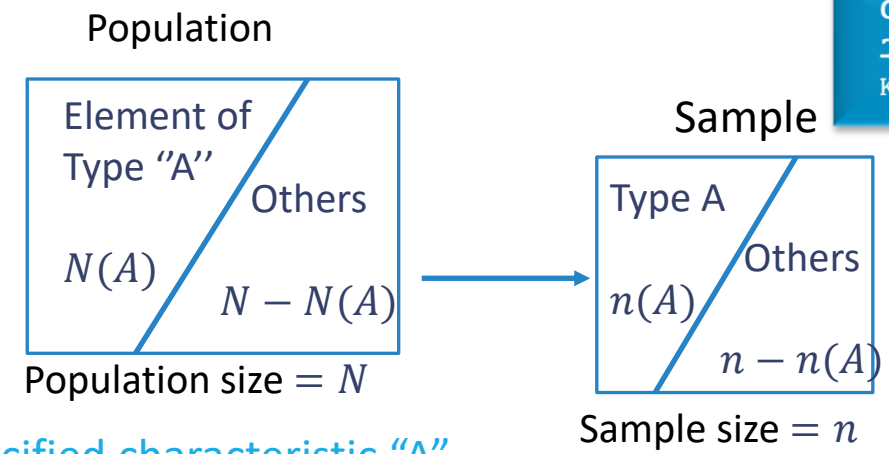
$= P(Z > 1.23)$

$= 1 - P(Z < 1.23)$

$= 1 - 0.89065$

$= 0.10935$

5.5 Distribution of the Sample Proportion ($\hat{P}$) :



- For the population:

$N(A)$ = number of elements in the population with a specified characteristic "A"

N = total number of elements in the population (population size)

The population proportion is :

$$P = \frac{N(A)}{N}$$ (P is a parameter)

- For the sample:

$n(A)$ = number of elements in the sample with the same characteristic "A"

n = sample size

The sample proportion is :

$$\hat{P} = \frac{N(A)}{N}$$ ($\hat{P}$ is a statistic)

- The sampling distribution of $\hat{P}$ is used to make inferences about p.

Result:

- The mean of the sample proportion $\hat{P}$ is the population proportion (P). ; that is:

$\mu_{\hat{P}} = P$

- The variance of the sample proportion $\hat{P}$ is :

$$\sigma^2_{\hat{P}} = \frac{P(1-P)}{n} = \frac{pq}{n} \qquad \text{(where q=1 –p)}$$

- The standard error (standard deviation) of the sample proportion $\hat{P}$ is :

$$\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}$$

- For large sample size ($n \geq 30, np > 5, nq > 5$), the sample proportion $\hat{P}$ has approximately a normal distribution with $\mu_{\hat{P}} = P$ and $\sigma^2_{\hat{P}} = \frac{pq}{n}$, that is:

$$\hat{P} \approx \text{Normal}\left(P, \frac{pq}{n}\right) \qquad \text{(approximately)}$$

$$Z = \frac{\hat{P}-P}{\sqrt{\frac{pq}{n}}} \sim \text{Normal (0,1)}$$

Example:

Suppose that 45% of the patients visiting a certain clinic are females. If a sample of 35 patients was selected at random, find the probability that:

1. The proportion of females in the sample will be greaterthan 0.4.
2. The proportion of females in the sample will be between 0.4 and 0.5.

Solution:

n = 35 (large)

p = The population proportion of females = $\frac{45}{100}$ = 0.45

$\hat{P}$ = The sample proportion (proportion of females in the sample)

- The mean of the sample proportion $\hat{P}$ is : P=0.45
- The variance of the sample proportion $\hat{P}$ is : $\frac{(0.45)(0.55)}{35}$ = 0.0071
- The standard error (standard deviation) of the sample proportion $\hat{P}$ is : $\sqrt{0.0071}$ = 0.084
- $n \geq 30$, $np = 35(0.45) = 15.75 > 5$, $nq = 35(0.55) = 19.25 > 5$

- The probability that the proportion of females in the sample will be greaterthan 0.4 :

$P(\hat{P} > 0.4) =$

$= P(Z > \frac{0.4-0.45}{\sqrt{0.0071}})$

$= P(Z > -0.59) = 1 - P(Z < -0.59)$

$= 1 - 0.2776$

$= 0.7224$

- The probability that the proportion of females in the sample will be between 0.4 and 0.5 :

$P(0.4 < \hat{P} < 0.5) = P(\frac{0.4-0.45}{\sqrt{0.0071}} < Z < \frac{0.5-0.45}{\sqrt{0.0071}})$

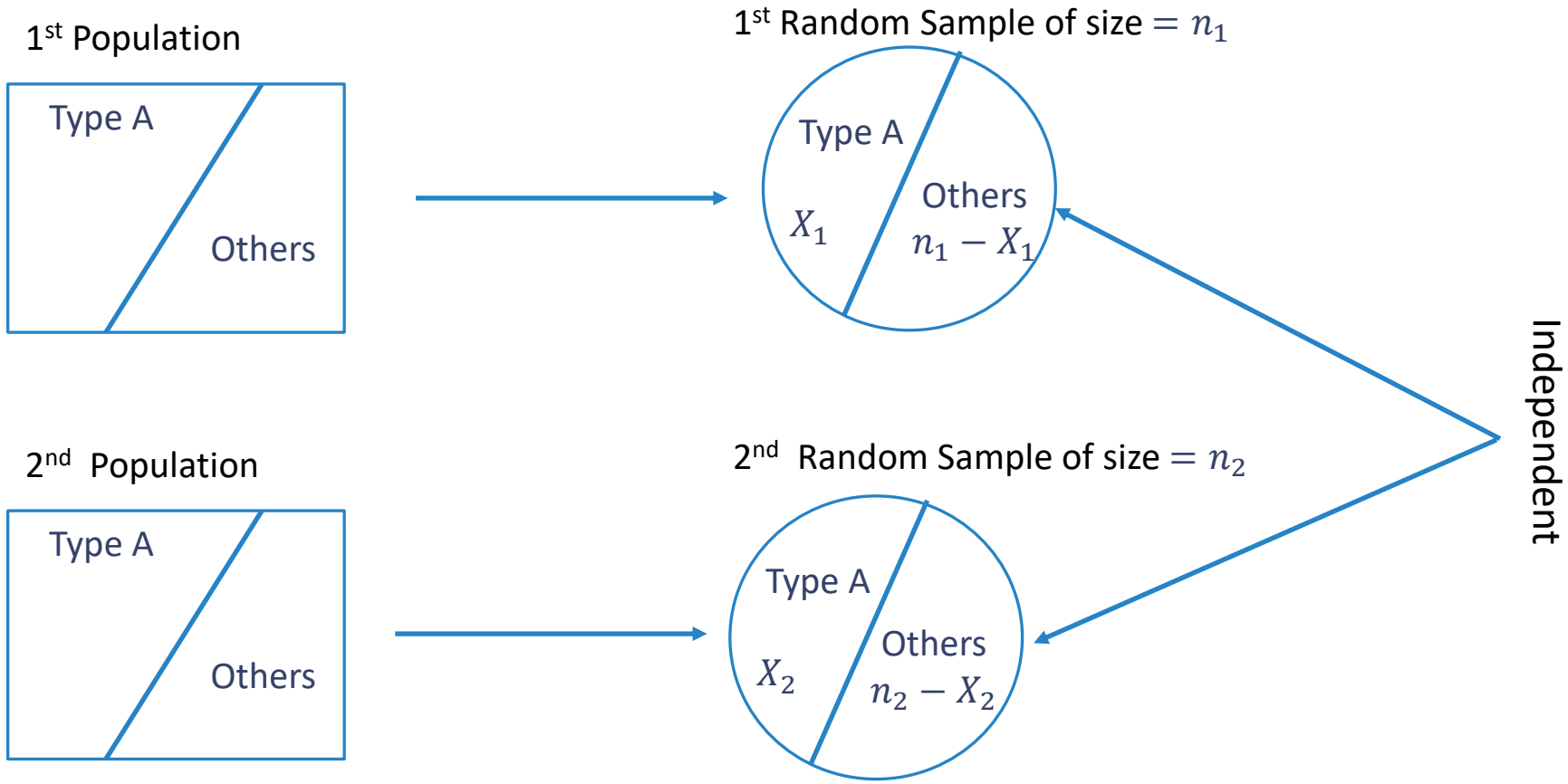
$= P(-0.59 < Z < 0.59)$

$= P(Z < 0.59) - P(Z < -0.59)$

$= 0.7224 - 0.2776$

$= 0.4448$

5.6 Distributions of the difference between two sample proportions ($\hat{P}_1 - \hat{P}_2$) :



Suppose that we have two populations:

- P_1= proportion of elements of type (A) in the 1-st population.
- P_2= proportion of elements of type (A) in the 2-nd population.
- We are interested in comparing P_1 and P_2 , or equivalently,making inferences about $P_1 - P_2$.
- We independently select a random sample of size n_1 fromthe 1-st population and another random sample of size n_2 from the 2-nd population:
- Let X_1 = no. of elements of type (A) in the 1-st sample.
- Let X_2 = no. of elements of type (A) in the 2-nd sample.
- $\hat{P}_1 = \frac{X_1}{n_1}$ = sample proportion of the 1-st sample .
- $\hat{P}_2 = \frac{X_2}{n_2}$ = sample proportion of the 2-nd sample .
- The sampling distribution of $\hat{P}_1 - \hat{P}_2$ is used to make inferences about $P_1 - P_2$.

The sampling distribution of $\hat{P}_1 - \hat{P}_2$:

Result :

The mean, the variance and the standard error (standard deviation) of $\hat{P}_1 - \hat{P}_2$ are :

- The mean of $\hat{P}_1 - \hat{P}_2$:

$$\mu_{\hat{P}_1 - \hat{P}_2} = P_1 - P_2$$

- The variance of $\hat{P}_1 - \hat{P}_2$ is :

$$\sigma^2_{\hat{P}_1 - \hat{P}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

- The standard error (standard deviation) of $\hat{P}_1 - \hat{P}_2$ is :

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

- $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

Result:

For large samples sizes $(n_1 \geq 30, n_2 \geq 30, n_1p_1 > 5, n_1q_1 > 5, n_2p_2 > 5\ , n_2q_2 > 5)$, we have :

$\hat{P}_1 - \hat{P}_2 \approx$ Normal $(P_1 - P_2\ , \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}\ )$ (approximately)

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} \sim \text{Normal } (0,1)$$

Example :

Suppose that 40% of Non-Saudi residents have medical insurance and 30% of Saudi residents have medical insurance in a certain city. We have randomly and independently selected a sample of 130 Non-Saudi residents and another sample of 120 Saudi residents. What is the probability that the difference between the sample proportions, will be between 0.05 and 0.2 ?

P_1 = population proportion of non-Saudi with medical insurance

P_2 = population proportion of Saudi with medical insurance

$\hat{P}_1$ = sample proportion of non-Saudi with medical insurance

$\hat{P}_2$ = sample proportion of Saudi with medical insurance

$P_1 = 0.4$, $n_1 = 130$

$P_2 = 0.3$, $n_2 = 120$

$\mu_{\hat{P}_1-\hat{P}_2} = P_1 - P_2 = 0.4 - 0.3 = 0.1$

$$\sigma^2_{\hat{P}_1-\hat{P}_2} = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2} = \frac{(0.4)(0.6)}{130} + \frac{(0.3)(0.7)}{120} = 0.0036$$

$$\sigma_{\hat{P}_1-\hat{P}_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} = \sqrt{0.0036} = 0.06$$

The probability that the difference between the sample proportions, will be between 0.05 and 0.2 is :

$$P(0.05 < \hat{P}_1 - \hat{P}_2 < 0.2) = P\left(\frac{0.05-(P_1-P_2)}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} < Z < \frac{0.2-(P_1-P_2)}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}}\right)$$

$$= P\left(\frac{0.05-0.1}{0.06} < Z < \frac{0.2-0.1}{0.06}\right)$$

$$= P(-0.83 < Z < 1.67)$$

$$= P(Z < 1.67) - P(Z < -0.83)$$

$$= 0.95254 - 0.20327 = 0.74927$$