CHAPTER 5

Probabilistic Features of the Distributions of Certain Sample Statistics

5.1. Introduction :

As we mention earlier:

1) a **statistic** is a measure computed form the random sample. As the sample values vary from sample to sample.

2) the value of the statistic varies accordingly.

Definition:

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.

5.3. Distribution of the Sample Mean:

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

	1st sample	2nd sample	3rd sample	...	Last sample
Sample values	28	31	14	.	17
	30	20	31	.	32
	34	31	25	.	29
	34	40	27	.	31
	17	28	32		30
Sample mean $\overline{X}$	28.4	29.9	25.8	...	27.8

– The value of the sample mean $\overline{X}$ varies from random sample to another.

– The value of $\overline{X}$ is random and it depends on the random sample.

– The sample mean $\overline{X}$ is a random variable.

– The probability distribution of $\overline{X}$ is called the sampling distribution of the sample mean $\overline{X}$.

Result (1): (mean & variance of $\overline{X}$)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from any distribution with mean μ and variance σ^2; then:

1. The mean of $\overline{X}$ is: $\qquad \mu_{\overline{X}} = \mu$.

2. The variance of $\overline{X}$ is: $\qquad \sigma_{\overline{X}}^2 = \dfrac{\sigma^2}{n}$.

3. The Standard deviation of $\overline{X}$ is call <u>the standard error</u> and is defined by: $\qquad \sigma_{\overline{X}} = \sqrt{\sigma_{\overline{X}}^2} = \dfrac{\sigma}{\sqrt{n}}$.

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 has a normal distribution with mean μ and variance σ^2 / n, that is:

1. $\overline{X} \sim$ Normal $\left(\mu, \dfrac{\sigma^2}{n} \right)$.

2. $Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim$ Normal $(0,1)$.

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 non-normal population with mean μ and variance σ^2. If the sample size n is large $(n \geq 30)$, then the sample mean has approximately a normal distribution with mean μ and variance σ^2/n, that is

1. $\overline{X} \approx \text{Normal}\left(\mu, \dfrac{\sigma^2}{n}\right)$ (approximately)

2. $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx \text{Normal}(0,1)$ (approximately)

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

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

$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

has a t- distribution with $(n-1)$ degrees of freedom, where S is the sample standard deviation given by:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

We write:

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

Notation: degrees of freedom = df = ν

The t- distribution (Student's t distribution):

- Recall that, if X_1, X_2, …, X_n is a random sample of size n from a normal distribution with mean μ and variance σ^2, i.e. $N(\mu, \sigma^2)$, then

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

We can apply this result only when σ^2 is known!

- If σ^2 is unknown, we replace the population variance σ^2 with the sample variance $S^2 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ to have the following statistic

$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

then the statistic:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a t-distribution with $(n-1)$ degrees of freedom ($df = \nu = n-1$), and we write T~ t(ν) or T~ t($n-1$).

Note:

- t-distribution is a continuous distribution.
- The value of t random variable range from $-\infty$ to $+\infty$ (that is, $-\infty < t < \infty$).
- The mean of t distribution is 0.
- It is symmetric about the mean 0.
- The shape of t-distribution is similar to the shape of the standard normal distribution.
- t-distribution $\rightarrow$ Standard normal distribution as n $\rightarrow \infty$.

Notation: (t_α)

$T \sim t(v)$ $T \sim t(v)$

α $1-\alpha$ $1-\alpha$ α

t_α $t_{1-\alpha} = -t_\alpha$

$P(T < t_\alpha) = \alpha$
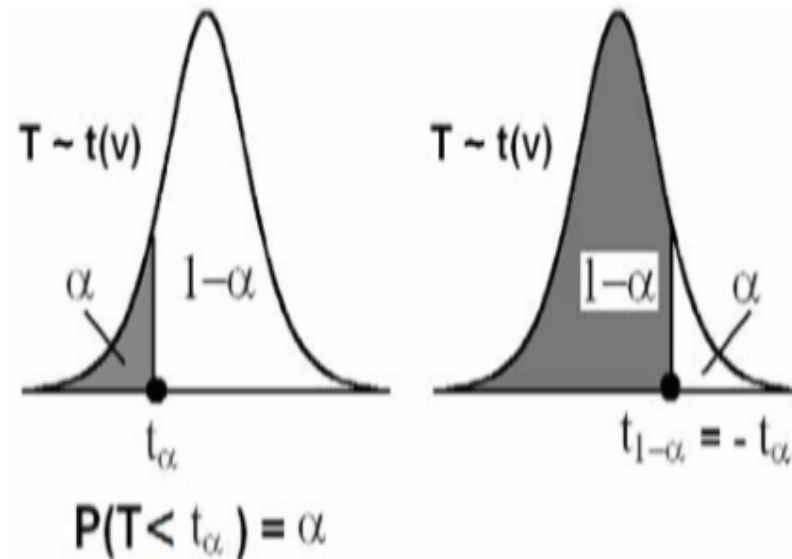
- t_α = The t-value under which we find an area equal to α
  = The t-value that leaves an area of α to the left.
- The value t_α satisfies: $P(T < t_\alpha) = \alpha$.
- Since the curve of the pdf of $T \sim t(v)$ is symmetric about 0, we have

$$t_{1-\alpha} = -t_\alpha$$

Critical Values of the t-distribution (t_α)

v=df	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
35	1.3062	1.6896	2.0301	2.4377	2.7238
40	1.3030	1.6840	2.0210	2.4230	2.7040
45	1.3006	1.6794	2.0141	2.4121	2.6896
50	1.2987	1.6759	2.0086	2.4033	2.6778
60	1.2958	1.6706	2.0003	2.3901	2.6603
70	1.2938	1.6669	1.9944	2.3808	2.6479
80	1.2922	1.6641	1.9901	2.3739	2.6387
90	1.2910	1.6620	1.9867	2.3685	2.6316
100	1.2901	1.6602	1.9840	2.3642	2.6259
120	1.2886	1.6577	1.9799	2.3578	2.6174
140	1.2876	1.6558	1.9771	2.3533	2.6114
160	1.2869	1.6544	1.9749	2.3499	2.6069
180	1.2863	1.6534	1.9732	2.3472	2.6034
200	1.2858	1.6525	1.9719	2.3451	2.6006
∞	1.282	1.645	1.960	2.326	2.576

Example:

Find the t-value with $v=14$ (df) that leaves an area of:
- (a)   0.95 to the left.
- (b)   0.95 to the right.

Solution:

$v = 14$ (df); $T \sim t(14)$

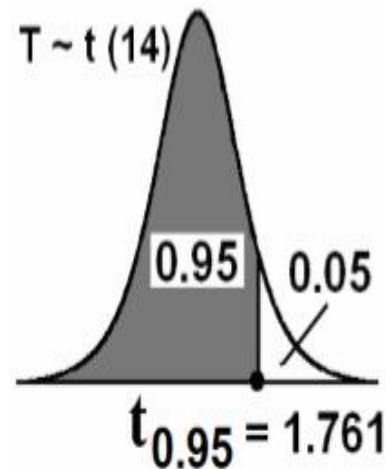(a) The t-value that leaves an area of 0.95 to the left is
$t_{0.95} = 1.761$.

(b) The t-value that leaves an area of 0.95 to the right is

$$t_{0.05} = -t_{1-0.05} = -t_{0.95} = -1.761$$

$T \sim t(14)$

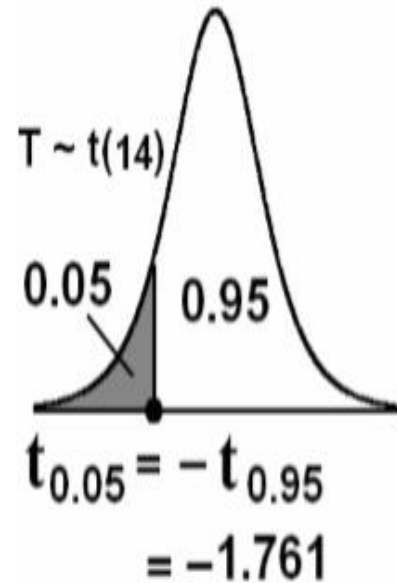0.05   0.95

$$t_{0.05} = -t_{0.95}$$
$$= -1.761$$

Table of t - Distribution

	0.05
14	−1.761

$$t_{0.05} = -1.761$$

Note: Some t-tables contain values of α that are greater than or equal to 0.90. When we search for small values of α in these tables, we may use the fact that:

$$t_{1-\alpha} = -t_{\alpha}$$

Example:

For $\nu = 10$ degrees of freedom (df), find $t_{0.93}$ and $t_{0.07}$.

Solution:

$t_{0.93} = (1.372 + 1.812)/2 = 1.592$ (from the table)

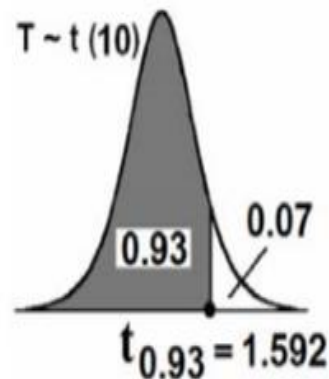$t_{0.07} = -t_{1-0.07} = -t_{0.93} = -1.592$ (using the rule: $t_{1-\alpha} = -t_{\alpha}$)

Example:

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:

Mean of $\bar{X}$: $\quad \mu_{\bar{X}} = \mu = 800$

Variance of $\bar{X}$: $\quad \sigma_{\bar{X}}^2 = \dfrac{\sigma^2}{n} = \dfrac{1600}{16} = 100$

Standard error (standard deviation) of $\bar{X}$: $\quad \sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{40}{\sqrt{16}} = 10$

Using the central limit theorem, $\bar{X}$ has a normal distribution with mean $\mu_{\bar{X}} = 800$ and variance $\sigma_{\bar{X}}^2 = 100$, that is:

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

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

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

$$P(\bar{X} < 775) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{775 - \mu}{\sigma/\sqrt{n}}\right) = P\left(\frac{\bar{X} - 800}{10} < \frac{775 - 800}{10}\right)$$

$$= P\left(Z < \frac{775 - 800}{10}\right) = P(Z < -2.50) = 0.0062$$

Example:

If the mean and standard deviation of serum iron values for healthy mean 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:

Mean of $\bar{X}$: $\qquad \mu_{\bar{X}} = \mu = 120$

Variance of $\bar{X}$: $\qquad \sigma_{\bar{X}}^2 = \dfrac{\sigma^2}{n} = \dfrac{225}{50} = 4.5$

Standard error (standard deviation) of $\bar{X}$: $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{15}{\sqrt{50}} = 2.12$

Using the central limit theorem, $\overline{X}$ has a normal distribution with mean $\mu_{\overline{X}} = 120$ and variance $\sigma_{\overline{X}}^2 = 4.5$, that is:

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(120, 4.5)$$

$$\Leftrightarrow \ Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - 120}{2.12} \sim N(0,1)$$

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

$$P(115 < \overline{X} < 125) = P\left(\frac{115 - \mu}{\sigma/\sqrt{n}} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{125 - \mu}{\sigma/\sqrt{n}} \right)$$

$$= P\left(\frac{115 - 120}{2.12} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{125 - 120}{2.12} \right) = P(-2.36 < Z < 2.36)$$

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

$$= 0.9909 - 0.0091$$

$$= 0.9818$$

5.4. Distribution of the difference between Two Sample Means $\overline{X}_1 - \overline{X}_2$:

Result:

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

Mean of $\overline{X}_1 - \overline{X}_2$ is:
$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2$$

Variance of $\overline{X}_1 - \overline{X}_2$ is:
$$\sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Standard error (standard) deviation of $\overline{X}_1 - \overline{X}_2$ is:
$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\sigma^2_{\overline{X}_1 - \overline{X}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

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 $((\sigma_1^2 / n_1) + (\sigma_2^2 / n_2))$, that is:

- $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2 , \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2} \right)$

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

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:

Mean of $\bar{X}_1 - \bar{X}_2$ is:

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 = 45 - 30 = 15$$

Variance of $\bar{X}_1 - \bar{X}_2$ is:

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{225}{35} + \frac{400}{40} = 16.4286$$

Standard error (standard) deviation of $\bar{X}_1 - \bar{X}_2$ is:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1 - \bar{X}_2}} = \sqrt{16.4286} = 4.0532$$

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

$$\bar{X}_1 - \bar{X}_2 \sim N(15, 16.4286)$$

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

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 is:

$$P(\bar{X}_1 - \bar{X}_2 > 20) = P\left(\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} > \frac{20 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}\right)$$

$$= P\left(Z > \frac{20 - 15}{4.0532}\right) = P(Z > 1.23) = 1 - P(Z < 1.23)$$

$$= 1 - 0.8907$$

$$= 0.1093$$

5.5. Distribution of the Sample Proportion $\hat{p}$:

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:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}}$$

Result:

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

$$\hat{p} \sim N\left(p, \frac{pq}{n} \right) \quad \text{(approximately)}$$

$$Z = \frac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}} \sim N(0,1) \quad \text{(approximately)}$$

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 greater than 0.4.
2. the proportion of females in the sample will be between 0.4 and 0.5.

Solution:

1. The probability that the sample proportion of females ($\hat{p}$) will be greater than 0.4 is:

$$P(\hat{p} > 0.4) = 1 - P(\hat{p} < 0.4) = 1 - P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{0.4 - p}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$$= 1 - P\left(Z < \frac{0.4 - 0.45}{\sqrt{\frac{0.45(1 - 0.45)}{35}}}\right) = 1 - P(Z < -0.59)$$

$$= 1 - 0.2776 = 0.7224$$

2. The probability that the sample proportion of females ($\hat{p}$) will be between 0.4 and 0.5 is:

$$P(0.4 < \hat{p} < 0.5) = P(\hat{p} < 0.5) - P(\hat{p} < 0.4)$$

$$= P\left(\frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} < \frac{0.5 - p}{\sqrt{\dfrac{p(1-p)}{n}}} \right) - 0.2776$$

$$= P\left(Z < \frac{0.5 - 0.45}{\sqrt{\dfrac{0.45(1-0.45)}{35}}} \right) - 0.2776$$

$$= P(Z < 0.59) - 0.2776$$
$$= 0.7224 - 0.2776$$
$$= 0.4448$$

5.6. Distribution of the Sample Proportion $\hat{p}_1 - \hat{p}_2$:

Result:

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

- Mean of $\hat{p}_1 - \hat{p}_2$ is:

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$$

- Variance of $\hat{p}_1 - \hat{p}_2$ is:

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

- 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

$(n_1 \geq 30, n_2 \geq 30, n_1 p_1 > 5, n_1 q_1 > 5, n_2 p_2 > 5, n_2 q_2 > 5)$, we have that $\hat{p}_1 - \hat{p}_2$ has approximately normal distribution with mean $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ and variance $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}$, that is:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \right) \quad \text{(Approximately)}$$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}} \sim N(0,1) \quad \text{(Approximately)}$$

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, $\hat{p}_1 - \hat{p}_2$, will be between 0.05 and 0.2?

Solution:

$$p_1 = 0.4 \qquad n_1 = 130$$
$$p_2 = 0.3 \qquad n_2 = 120$$

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 = 0.4 - 0.3 = 0.1$$

$$\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} = \frac{(0.4)(0.6)}{130} + \frac{(0.3)(0.7)}{120} = 0.0036$$

Solution:

$$P(0.05 < \hat{p}_1 - \hat{p}_2 < 0.2) = P(\hat{p}_1 - \hat{p}_2 < 0.2) \;-\; P(\hat{p}_1 - \hat{p}_2 < 0.05)$$

$$= P\left(\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}} < \frac{0.2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}} \right)$$

$$- P\left(\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}} < \frac{0.05 - (p_1 - p_2)}{\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}} \right)$$

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

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

$$= 0.9515 - 0.2033$$

$$= 0.7482$$