Chapter 9: One- and Two-Sample Estimation Problems:

9.1 Introduction:
- Suppose we have a population with some unknown parameter(s).
  Example: Normal($\mu, \sigma$)
  $\mu$ and $\sigma$ are parameters.
- We need to draw conclusions (make inferences) about the unknown parameters.
- We select samples, compute some statistics, and make inferences about the unknown parameters based on the sampling distributions of the statistics.

* Statistical Inference
  (1) Estimation of the parameters (Chapter 9)
    $\rightarrow$ Point Estimation
    $\rightarrow$ Interval Estimation (Confidence Interval)
  (2) Tests of hypotheses about the parameters (Chapter 10)

9.3 Classical Methods of Estimation:

Point Estimation:
A point estimate of some population parameter $\theta$ is a single value $\hat{\theta}$ of a statistic $\hat{\Theta}$. For example, the value $\hat{x}$ of the statistic $X$ computed from a sample of size $n$ is a point estimate of the population mean $\mu$.

Interval Estimation (Confidence Interval = C.I.):
An interval estimate of some population parameter $\theta$ is an interval of the form $(\hat{\theta}_L, \hat{\theta}_U)$, i.e., $\hat{\theta}_L < \theta < \hat{\theta}_U$. This interval contains the true value of $\theta$ "with probability $1-\alpha$", that is $P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1-\alpha$.
- $(\hat{\theta}_L, \hat{\theta}_U) = (\hat{\theta}_L < \theta < \hat{\theta}_U)$ is called a $(1-\alpha)100\%$ confidence interval (C.I.) for $\theta$.
- $1-\alpha$ is called the confidence coefficient
- $\hat{\theta}_L =$ lower confidence limit
• $\hat{\theta}_U$ = upper confidence limit
• $\alpha=0.1, 0.05, 0.025, 0.01$ (0<$\alpha$<1)

9.4 Single Sample: Estimation of the Mean ($\mu$):
Recall:
• $E(\bar{X}) = \mu$  
  $\bar{X} = \sum_{i=1}^{n} X_i / n$ is the sample mean of a random sample of size $n$ from a population (distribution) with mean $\mu$ and known variance $\sigma^2$, then a $(1-\alpha)100\%$ confidence interval for $\mu$ is:
\[ \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim t(n-1) \]

Point Estimation of the Mean ($\mu$):
• The sample mean $\bar{X} = \sum_{i=1}^{n} X_i / n$ is a "good" point estimate for $\mu$.

Interval Estimation (Confidence Interval) of the Mean ($\mu$):
(i) First Case: $\sigma^2$ is known:
Result:
If $\bar{X} = \sum_{i=1}^{n} X_i / n$ is the sample mean of a random sample of size $n$ from a population (distribution) with mean $\mu$ and known variance $\sigma^2$, then a $(1-\alpha)100\%$ confidence interval for $\mu$ is:
\[ \bar{X} - Z_{a} \sigma / \sqrt{n} < \mu < \bar{X} + Z_{a} \sigma / \sqrt{n} \]

Notation:
$Z_a$ is the Z-value leaving an area of $a$ to the right; i.e., $P(Z>\bar{Z}_a)=a$ or equivalently, $P(Z<\bar{Z}_a)=1-a$
\[
(\overline{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})
\]
⇔ \(\overline{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\)
⇔ \(\overline{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\)
where \(Z_{\frac{\alpha}{2}}\) is the Z-value leaving an area of \(\frac{\alpha}{2}\) to the right; i.e., \(P(Z > Z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}\), or equivalently, \(P(Z < Z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}\).

Note:
We are \((1-\alpha)100\%\) confident that \(\mu \in (\overline{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})\).

**Example 9.2:**
The average zinc concentration recorded from a sample of zinc measurements in 36 different locations is found to be 2.6 gram/milliliter. Find a 95% and 99% confidence interval (C.I.) for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3.

**Solution:**
\(\mu = \) the mean zinc concentration in the river.
(unknown parameter)

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mu = ??)</td>
<td>(n = 36)</td>
</tr>
<tr>
<td>(\sigma = 0.3)</td>
<td>(\overline{X} = 2.6)</td>
</tr>
</tbody>
</table>

First, a point estimate for \(\mu\) is \(\overline{X} = 2.6\).

(a) We want to find 95% C.I. for \(\mu\).
\(\alpha = ??\)
95% = \((1-\alpha)100\%
⇔ 0.95 = (1-\alpha)
⇔ \alpha = 0.05 \iff \alpha/2 = 0.025

\(Z \sim N(0,1)\)
\[\begin{align*}
Z_{0.025} &= -1.96 \\
Z_{0.975} &= 1.96
\end{align*}\]
\[ Z_{\alpha/2} = Z_{0.025} = 1.96 \]

A 95% C.I. for \( \mu \) is

\[ \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \]

\( \iff \bar{X} - Z_{\alpha} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha} \frac{\sigma}{\sqrt{n}} \)

\( \iff 2.6 - (1.96) \left( \frac{0.3}{\sqrt{36}} \right) < \mu < 2.6 + (1.96) \left( \frac{0.3}{\sqrt{36}} \right) \)

\( \iff 2.6 - 0.098 < \mu < 2.6 + 0.098 \)

\( \iff 2.502 < \mu < 2.698 \)

\( \iff \mu \in (2.502, 2.698) \)

We are 95% confident that \( \mu \in (2.502, 2.698) \).

(b) Similarly, we can find that (Homework) A 99% C.I. for \( \mu \) is

\[ 2.471 < \mu < 2.729 \]

\( \iff \mu \in (2.471, 2.729) \)

We are 99% confident that \( \mu \in (2.471, 2.729) \)

Notice that a 99% C.I. is wider than a 95% C.I..

Note:

Error

\[ \frac{Z_{\alpha}}{2} \frac{\sigma}{\sqrt{n}} \]

\[ \bar{X} - Z_{\alpha} \frac{\sigma}{\sqrt{n}} \]

\[ \bar{X} + Z_{\alpha} \frac{\sigma}{\sqrt{n}} \]

\[ Z_{\alpha} \frac{\sigma}{\sqrt{n}} \]

\[ Z_{\alpha} \frac{\sigma}{\sqrt{n}} \]

Theorem 9.1:
If \( \bar{X} \) is used as an estimate of \( \mu \), we can then be \( (1-\alpha)100\% \)

confident that the error (in estimation) will not exceed \( Z_{\alpha} \frac{\sigma}{\sqrt{n}} \).
Note:
max error of estimation = \( Z_{\alpha} \frac{\sigma}{\sqrt{n}} \) with \((1-\alpha)100\%\) confidence.

Example:
In Example 9.2, we are 95\% confident that the sample mean \( \bar{X} = 2.6 \) differs from the true mean \( \mu \) by an amount less than 
\[
Z_{\alpha} \frac{\sigma}{\sqrt{n}} = (1.96) \left( \frac{0.3}{\sqrt{36}} \right) = 0.098.
\]

Note:
Let \( e \) be the maximum amount of the error, that is 
\[
e = Z_{\alpha} \frac{\sigma}{\sqrt{n}},
\]
then:
\[
e = Z_{\alpha} \frac{\sigma}{\sqrt{n}} \iff \sqrt{n} = Z_{\alpha} \frac{\sigma}{e} \iff n = \left( Z_{\alpha} \frac{\sigma}{e} \right)^2
\]

Theorem 9.2:
If \( \bar{X} \) is used as an estimate of \( \mu \), we can then be \((1-\alpha)100\%\) confident that the error (in estimation) will not exceed a specified amount \( e \) when the sample size is 
\[
n = \left( Z_{\alpha} \frac{\sigma}{e} \right)^2.
\]

Note:
1. All fractional values of \( n = (Z_{\alpha} \sigma/e)^2 \) are rounded up to the next whole number.
2. If \( \sigma \) is unknown, we could take a preliminary sample of size \( n \geq 30 \) to provide an estimate of \( \sigma \). Then using 
\[
S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}
\]
we could determine approximately how many observations are needed to provide the desired degree of accuracy.

Example 9.3:
How large a sample is required in Example 9.2 if we want to be 95\% confident that our estimate of \( \mu \) is off by less than 0.05?

Solution:
We have $\sigma = 0.3$, $Z_{\alpha/2} = 1.96$, $e = 0.05$. Then by Theorem 9.2,

$$n = \left( Z_{\alpha/2} \frac{\sigma}{e} \right)^2 = \left( 1.96 \times \frac{0.3}{0.05} \right)^2 = 138.3 \approx 139$$

Therefore, we can be 95% confident that a random sample of size $n=139$ will provide an estimate $\bar{X}$ differing from $\mu$ by an amount less than $e=0.05$.

**Interval Estimation (Confidence Interval) of the Mean ($\mu$):**

**(ii) Second Case: $\sigma^2$ is unknown:**

Recall:

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

**Result:**

If $\bar{X} = \frac{\sum X_i}{n}$ and $S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$ are the sample mean and the sample standard deviation of a random sample of size $n$ from a normal population (distribution) with unknown variance $\sigma^2$, then a $(1-\alpha)100\%$ confidence interval for $\mu$ is:

$$\left( \bar{X} - t_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2} \frac{S}{\sqrt{n}} \right)$$

$$\Leftrightarrow \bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

$$\Leftrightarrow \bar{X} - t_{\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2} \frac{S}{\sqrt{n}}$$

where $t_{\alpha/2}$ is the t-value with $v=n-1$ degrees of freedom leaving an area of $\alpha/2$ to the right; i.e., $P(T > t_{\alpha/2}) = \alpha/2$, or equivalently, $P(T < t_{\alpha/2}) = 1 - \alpha/2$.

**Example 9.4:**

The contents of 7 similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 liters. Find a 95% C.I. for the
mean of all such containers, assuming an approximate normal distribution.

Solution:

\[ n = 7 \quad \bar{X} = \frac{\sum X_i}{n} = 10.0 \quad S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}} = 0.283 \]

First, a point estimate for \( \mu \) is \( \bar{X} = \frac{\sum X_i}{n} = 10.0 \)

Now, we need to find a confidence interval for \( \mu \).

\( \alpha = ?? \)

\[
95\% = (1 - \alpha)100\% \iff 0.95 = (1 - \alpha) \iff \alpha = 0.05 \iff \alpha/2 = 0.025
\]

\[
t_{\alpha/2} = t_{0.025} = 2.447 \quad \text{(with } \nu = n - 1 = 6 \text{ degrees of freedom)}
\]

A 95% C.I. for \( \mu \) is

\[
\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}
\]

\( \iff \bar{X} - t_{\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2} \frac{S}{\sqrt{n}} \)

\( \iff 10.0 - (2.447)\left(\frac{0.283}{\sqrt{7}}\right) < \mu < 10.0 + (2.447)\left(\frac{0.283}{\sqrt{7}}\right) \)

\( \iff 10.0 - 0.262 < \mu < 10.0 + 0.262 \)

\( \iff 9.74 < \mu < 10.26 \)

\( \iff \mu \in (9.74, 10.26) \)

We are 95% confident that \( \mu \in (9.74, 10.26) \).

9.5 Standard Error of a Point Estimate:

- The standard error of an estimator is its standard deviation.
- We use \( \bar{X} = \frac{\sum X_i}{n} \) as a point estimator of \( \mu \), and we used the sampling distribution of \( \bar{X} \) to make a \((1 - \alpha)100\% \) C.I. for \( \mu \).
- The standard deviation of \( \bar{X} \), which is \( \sigma_{\bar{X}} = \sigma / \sqrt{n} \), is called the standard error of \( \bar{X} \). We write \( s.e.(\bar{X}) = \sigma / \sqrt{n} \).
\[ \bar{X} \pm Z_{\alpha} \frac{\sigma}{\sqrt{n}} = \bar{X} \pm Z_{\alpha} s.e(\bar{X}) \]

Note: a \((1-\alpha)100\%\) C.I. for \(\mu\), when \(\sigma^2\) is known, is

\[ \bar{X} \pm t_{\alpha} \frac{S}{\sqrt{n}} = \bar{X} \pm t_{\alpha} s.e(\bar{X}) \]  
(v=n−1 df)

9.7 Two Samples: Estimating the Difference between Two Means \((\mu_1-\mu_2)\):

Recall: For two independent samples:

- \(\mu_{x_1-x_2} = \mu_1 - \mu_2 = E(\bar{X}_1 - \bar{X}_2)\)
- \(\sigma^2_{x_1-x_2} = \frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2} = Var(\bar{X}_1 - \bar{X}_2)\)
- \(\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1-\bar{X}_2}} = \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}\)
- \(Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}} \sim N(0,1)\)

Point Estimation of \(\mu_1-\mu_2\):
- \(\bar{X}_1 - \bar{X}_2\) is a "good" point estimate for \(\mu_1-\mu_2\).

Confidence Interval of \(\mu_1-\mu_2\):

\[(i)\ First\ Case: \ \sigma^2_1 \text{ and } \sigma^2_2 \text{ are known:}\]

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

- Result:

\[ (\bar{X}_1 - \bar{X}_2) - Z_{\alpha} \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + Z_{\alpha} \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}} \]
or \[(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha \over 2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\]
or \[\left(\bar{X}_1 - \bar{X}_2\right) - Z_{\alpha \over 2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{X}_1 - \bar{X}_2) + Z_{\alpha \over 2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\]

(ii) Second Case: \(\sigma_1^2 = \sigma_2^2 = \sigma^2\) is unknown:
- If \(\sigma_1^2\) and \(\sigma_2^2\) are unknown but \(\sigma_1^2 = \sigma_2^2 = \sigma^2\), then the pooled estimate of \(\sigma^2\) is
  \[S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\]
  where \(S_1^2\) is the variance of the 1-st sample and \(S_2^2\) is the variance of the 2-nd sample. The degrees of freedom of \(S_p^2\) is \(\nu = n_1 + n_2 - 2\).
- Result:
a \((1 - \alpha)100\%\) confidence interval for \(\mu_1 - \mu_2\) is:
  \[(\bar{X}_1 - \bar{X}_2) - t_{\alpha \over 2} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{\alpha \over 2} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}\]
or
  \[(\bar{X}_1 - \bar{X}_2) - t_{\alpha \over 2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{\alpha \over 2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\]
or
  \[(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha \over 2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\]
or
  \[\left(\bar{X}_1 - \bar{X}_2\right) - t_{\alpha \over 2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\bar{X}_1 - \bar{X}_2) + t_{\alpha \over 2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\]
where \(t_{\alpha \over 2}\) is the t-value with \(\nu = n_1 + n_2 - 2\) degrees of freedom.

Example 9.6: (1\textsuperscript{st} Case: \(\sigma_1^2\) and \(\sigma_2^2\) are known)
An experiment was conducted in which two types of engines, \(A\) and \(B\), were compared. Gas mileage in miles per gallon was measured. 50 experiments were conducted using engine type \(A\) and 75 experiments were done for engine type \(B\). The gasoline used and other conditions were held constant. The average gas
mileage for engine $A$ was 36 miles per gallon and the average for engine $B$ was 42 miles per gallon. Find 96% confidence interval for $\mu_B - \mu_A$, where $\mu_A$ and $\mu_B$ are population mean gas mileage for engines $A$ and $B$, respectively. Assume that the population standard deviations are 6 and 8 for engines $A$ and $B$, respectively.

**Solution:**

<table>
<thead>
<tr>
<th>Engine $A$</th>
<th>Engine $B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_A=50$</td>
<td>$n_B=75$</td>
</tr>
<tr>
<td>$\bar{X}_A=36$</td>
<td>$\bar{X}_B=42$</td>
</tr>
<tr>
<td>$\sigma_A=6$</td>
<td>$\sigma_B=8$</td>
</tr>
</tbody>
</table>

A point estimate for $\mu_B - \mu_A$ is $\bar{X}_B - \bar{X}_A=42-36=6$.

**Confidence interval:**

$\alpha = ??$

$96\% = (1-\alpha)100\% \leftrightarrow 0.96 = (1-\alpha) \leftrightarrow \alpha=0.04 \leftrightarrow \alpha/2 = 0.02$

$Z_{\alpha/2} = Z_{0.02} = 2.05$

A 96% C.I. for $\mu_B - \mu_A$ is

$$(\bar{X}_B - \bar{X}_A) - Z_{\alpha/2} \sqrt{\frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A}} < \mu_B - \mu_A < (\bar{X}_B - \bar{X}_A) + Z_{\alpha/2} \sqrt{\frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A}}$$

$$= (42 - 36) \pm Z_{0.02} \sqrt{\frac{8^2}{75} + \frac{6^2}{50}}$$

$$= 6 \pm (2.05) \sqrt{\frac{64}{75} + \frac{36}{50}}$$

$$= 6 \pm 2.571$$

$3.43 < \mu_B - \mu_A < 8.57$

We are 96% confident that $\mu_B - \mu_A \in (3.43, 8.57)$.

**Example 9.7:** (2nd Case: $\sigma_1^2=\sigma_2^2$ unknown) Reading Assignment
Example: (2\textsuperscript{nd} Case: $\sigma^2_1 = \sigma^2_2$ unknown)

To compare the resistance of wire $A$ with that of wire $B$, an experiment shows the following results based on two independent samples (original data multiplied by 1000):

Wire $A$: 140, 138, 143, 142, 144, 137
Wire $B$: 135, 140, 136, 142, 138, 140

Assuming equal variances, find 95% confidence interval for $\mu_A - \mu_B$, where $\mu_A$ ($\mu_B$) is the mean resistance of wire $A$ ($B$).

Solution:

<table>
<thead>
<tr>
<th>Wire $A$</th>
<th>Wire $B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_A = 6$</td>
<td>$n_B = 6$</td>
</tr>
<tr>
<td>$\bar{X}_A = 140.67$</td>
<td>$\bar{X}_B = 138.50$</td>
</tr>
<tr>
<td>$S^2_A = 7.86690$</td>
<td>$S^2_B = 7.10009$</td>
</tr>
</tbody>
</table>

A point estimate for $\mu_A - \mu_B$ is $\bar{X}_A - \bar{X}_B = 140.67 - 138.50 = 2.17$.

Confidence interval:

95% = (1$-$\(\alpha$))100% $\Leftrightarrow$ 0.95 = (1$-$\(\alpha$)) $\Leftrightarrow$ \(\alpha$ = 0.05 $\Leftrightarrow$ $\alpha$/2 = 0.025

$\nu = df = n_A + n_B - 2 = 10$

$t_{\frac{\alpha}{2}} = t_{0.025} = 2.228$

$S_p^2 = \frac{(n_A - 1)S^2_A + (n_B - 1)S^2_B}{n_A + n_B - 2} = \frac{(6 - 1)(7.86690) + (6 - 1)(7.10009)}{6 + 6 - 2} = 7.4835$

$S_p = \sqrt{S_p^2} = \sqrt{7.4835} = 2.7356$

A 95% C.I. for $\mu_A - \mu_B$ is

\[
(\bar{X}_A - \bar{X}_B) - t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} < \mu_A - \mu_B < (\bar{X}_A - \bar{X}_B) + t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}
\]

or

\[
(\bar{X}_A - \bar{X}_B) \pm t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}
\]

\[
(140.67 - 138.50) \pm (2.228)(2.7356)\sqrt{\frac{1}{6} + \frac{1}{6}}
\]

\[
2.17 \pm 3.51890
\]

\[-1.35 < \mu_A - \mu_B < 5.69
\]

We are 95% confident that $\mu_A - \mu_B \in (-1.35, 5.69)$. 

Department of Statistics and O.R. – 93 – King Saud University
9.9 Single Sample: Estimating of a Proportion:

\[ \hat{p} = \text{Sample proportion of successes (elements of Type } A\text{) in the sample} = \frac{X}{n} \]

\[ p = \text{Population proportion of successes (elements of Type } A\text{) in the population} = \frac{A}{A+B} = \frac{\text{no. of elements of type } A\text{ in the population}}{\text{Total no. of elements}} \]

\[ n = \text{sample size} \]

\[ X = \text{no. of elements of type } A\text{ in the sample of size } n. \]

\[ \hat{p} = \text{Sample proportion of successes (elements of Type } A\text{) in the sample} = \frac{X}{n} \]

Recall that:

1. \( X \sim \text{Binomial}\ (n, p) \)
2. \( E(\hat{p}) = E\left(\frac{X}{n}\right) = p \)
3. \( \text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{pq}{n} ; q = 1 - p \)
4. For large \( n \), we have \( \hat{p} \sim N(p, \sqrt{\frac{pq}{n}}) \) (Approximately)

\[ Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1) \] (Approximately)

**Point Estimation for \( p \):**

A good point estimator for the population proportion \( p \) is given by the statistic (sample proportion):

\[ \hat{p} = \frac{X}{n} \]
Confidence Interval for $p$:

Result:

For large $n$, an approximate $(1-\alpha)100\%$ confidence interval for $p$ is:

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} ; \hat{q} = 1 - \hat{p}$$

or

$$\left(\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

or

$$\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Example 9.10:

In a random sample of $n=500$ families owing television sets in the city of Hamilton, Canada, it was found that $x=340$ subscribed to HBO. Find 95% confidence interval for the actual proportion of families in this city who subscribe to HBO.

Solution:

$p =$ proportion of families in this city who subscribe to HBO.

$n =$ sample size

$= 500$

$X =$ no. of families in the sample who subscribe to HBO.

$= 340$

$\hat{p} =$ proportion of families in the sample who subscribe to HBO.

$$= \frac{X}{n} = \frac{340}{500} = 0.68$$

$\hat{q} = 1 - \hat{p} = 1 - 0.68 = 0.32$

A point estimator for $p$ is

$$\hat{p} = \frac{X}{n} = \frac{340}{500} = 0.68$$

Now,

$95\% = (1-\alpha)100\% \iff 0.95 = (1-\alpha) \iff \alpha = 0.05 \iff \alpha/2 = 0.025$

$$Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$$
A 95% confidence interval for $p$ is:

$$
\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \quad ; \quad \hat{q} = 1 - \hat{p}
$$

$$
0.68 \pm 1.96 \sqrt{\frac{(0.68)(0.32)}{500}}
$$

$$
0.68 \pm 0.04
$$

$$
0.64 < p < 0.72
$$

We are 95% confident that $p \in (0.64, 0.72)$.

### 9.10 Two Samples: Estimating the Difference between Two Proportions:

Suppose that we have two populations:

- $p_1 = \text{proportion of the 1-st population}.$
- $p_2 = \text{proportion of the 2-nd population}.$
- 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 $n_1$ from the 1-st population and another random sample of size $n_2$ from the 2-nd population:
- Let $X_1 = \text{no. of elements of type } A \text{ in the 1-st sample}.$
  $X_1 \sim \text{Binomial}(n_1, p_1)$
E(X₁) = \( n₁ \cdot p₁ \)
Var(X₁) = \( n₁ \cdot p₁ \cdot q₁ \) \((q₁=1−p₁)\)
• Let \( X₂ \) = no. of elements of type \( A \) in the 2-nd sample.
\( X₂ \sim \text{Binomial}(n₂, p₂) \)
E(X₂) = \( n₂ \cdot p₂ \)
Var(X₂) = \( n₂ \cdot p₂ \cdot q₂ \) \((q₂=1−p₂)\)
• \( \hat{p}_1 = \frac{X₁}{n₁} = \) proportion of the 1-st sample
• \( \hat{p}_2 = \frac{X₂}{n₂} = \) proportion of the 2-nd sample
• The sampling distribution of \( \hat{p}_1 - \hat{p}_2 \) is used to make inferences about \( p₁ - p₂ \).

**Result:**
1. \( E(\hat{p}_1 - \hat{p}_2) = p₁ - p₂ \)
2. \( \text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p₁ \cdot q₁}{n₁} + \frac{p₂ \cdot q₂}{n₂} ; q₁ = 1−p₁, q₂ = 1−p₂ \)
3. For large \( n₁ \) and \( n₂ \), we have
\[
\hat{p}_1 - \hat{p}_2 \sim \text{N}(p₁ - p₂, \sqrt{\frac{p₁ \cdot q₁}{n₁} + \frac{p₂ \cdot q₂}{n₂}}) \text{ (Approximately)}
\]
\[
Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p₁ - p₂)}{\sqrt{\frac{p₁ \cdot q₁}{n₁} + \frac{p₂ \cdot q₂}{n₂}}} \sim \text{N}(0,1) \text{ (Approximately)}
\]

**Point Estimation for \( p₁ - p₂ \):**
A good point estimator for the difference between the two proportions, \( p₁ - p₂ \), is given by the statistic:
\[
\hat{p}_1 - \hat{p}_2 = \frac{X₁}{n₁} - \frac{X₂}{n₂}
\]

**Confidence Interval for \( p₁ - p₂ \):**
Result:
For large \( n₁ \) and \( n₂ \), an approximate \((1−α)100\%\) confidence interval for \( p₁ - p₂ \) is:
\[
(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{α}{2}} \sqrt{\frac{\hat{p}_1 \cdot \hat{q}_1}{n₁} + \frac{\hat{p}_2 \cdot \hat{q}_2}{n₂}}
\]
or
\[
(p_1 - p_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}{n_1} + \frac{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}{n_2}}, \quad (\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}{n_1} + \frac{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}{n_2}}
\]

or
\[
(\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}{n_1} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}{n_1} + \frac{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}{n_2}}
\]

**Example 9.13:**
A certain change in a process for manufacture of component parts is being considered. Samples are taken using both existing and the new procedure to determine if the new process results in an improvement. If 75 of 1500 items from the existing procedure were found to be defective and 80 of 2000 items from the new procedure were found to be defective, find 90% confidence interval for the true difference in the fraction of defectives between the existing and the new process.

**Solution:**
\[p_1 = \text{fraction (proportion) of defectives of the existing process}\]
\[p_2 = \text{fraction (proportion) of defectives of the new process}\]
\[\hat{p}_1 = \text{sample fraction of defectives of the existing process}\]
\[\hat{p}_2 = \text{sample fraction of defectives of the new process}\]

<table>
<thead>
<tr>
<th>Existing Process</th>
<th>New Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_1 = 1500 )</td>
<td>( n_2 = 2000 )</td>
</tr>
<tr>
<td>( X_1 = 75 )</td>
<td>( X_2 = 80 )</td>
</tr>
<tr>
<td>( \hat{p}_1 = \frac{X_1}{n_1} = \frac{75}{1500} = 0.05 )</td>
<td>( \hat{p}_2 = \frac{X_2}{n_2} = \frac{80}{2000} = 0.04 )</td>
</tr>
<tr>
<td>( \hat{q}_1 = 1 - 0.05 = 0.95 )</td>
<td>( \hat{q}_2 = 1 - 0.04 = 0.96 )</td>
</tr>
</tbody>
</table>

Point Estimation for \( p_1 - p_2 \):
A point estimator for the difference between the two proportions, \( p_1 - p_2 \), is:
\[\hat{p}_1 - \hat{p}_2 = 0.05 - 0.04 = 0.01\]

Confidence Interval for \( p_1 - p_2 \):
A 90% confidence interval for \( p_1 - p_2 \) is:
\[
(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}
\]

\[
(\hat{p}_1 - \hat{p}_2) \pm Z_{0.05} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}
\]

\[
0.01 \pm 1.645 \sqrt{\frac{(0.05)(0.95)}{1500} + \frac{(0.04)(0.96)}{2000}}
\]

\[
0.01 \pm 0.01173
\]

\[-0.0017 < p_1 - p_2 < 0.0217\]

We are 90% confident that \( p_1 - p_2 \in (-0.0017, 0.0217) \).

Note:
Since \( 0 \in 90\% \) confidence interval = \((-0.0017, 0.0217)\), there is no reason to believe that the new procedure produced a significant decrease in the proportion of defectives over the existing method \( (p_1 - p_2 \approx 0 \iff p_1 \approx p_2) \).