One-Sample z-test and confidence interval estimate for a population mean:

Purpose: Study mean of one population.

Assumptions:
1) a random sample from a normal population with known variance, or
2) a relatively large (>30) random sample (if variance is unknown, it can be approximated by sample variance).

z-test:

Step 1: State Hypotheses (choose one of the three hypotheses below)

i) \( H_0: \mu = \mu_0 \) v.s. \( H_a: \mu \neq \mu_0 \) (Two-sided test)

ii) \( H_0: \mu = \mu_0 \) v.s. \( H_a: \mu > \mu_0 \) (Right-sided test)

iii) \( H_0: \mu = \mu_0 \) v.s. \( H_a: \mu < \mu_0 \) (Left-sided test)

Step 2: Compute z test statistic:

\[
z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}
\]

or

\[
z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad \text{(for large sample and } \sigma \text{ is unknown)}
\]

Step 3. Decision Rule:

\text{p-value approach: } \text{Compute p-value,}

\[
p\text{-value} = \begin{cases} 2 \cdot \min\{P(Z \geq z), P(Z \leq z)\} & \text{if } H_a: \mu \neq \mu_0 \\ P(Z \geq z) & \text{if } H_a: \mu > \mu_0 \\ P(Z \leq z) & \text{if } H_a: \mu < \mu_0 \\ \text{reject } H_0 \text{ if } p\text{-value} < \alpha \end{cases}
\]

\text{Critical value approach: } \text{Determine critical value(s) using } \alpha ,

\[
\text{reject } H_0 \text{ against}
\]

i) \( H_a: \mu \neq \mu_0 \) if \( z > z_{\alpha/2} \) or \( z < -z_{\alpha/2} \)

ii) \( H_a: \mu > \mu_0 \) if \( z > z_{\alpha} \)

iii) \( H_a: \mu < \mu_0 \) if \( z < -z_{\alpha} \)

Step 4: Draw conclusion (with a fixed significance level \( \alpha \))

Confidence interval: The \((1-\alpha)\%\) confidence interval estimate for population mean is

\[
\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \quad \text{(if large sample and unknown variance)}
\]
One-Sample \( t \)-test and confidence interval estimate for a population mean:
Purpose: Study mean of one population.
Assumptions: a random sample from a normal population with unknown variance.

\( t \)-test:

Step 1: State Hypotheses (choose one of the three hypotheses below)
   i) \( H_0: \mu = \mu_0 \) v.s. \( H_a: \mu \neq \mu_0 \) (Two-sided test)
   ii) \( H_0: \mu = \mu_0 \) v.s. \( H_a: \mu > \mu_0 \) (Right-sided test)
   iii) \( H_0: \mu = \mu_0 \) v.s. \( H_a: \mu < \mu_0 \) (Left-sided test)

Step 2: Compute \( t \) test statistic:
   \[ t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \sim t \text{-distribution with d.f.} = n - 1 \]

Step 3. Decision Rule:

   - **p-value approach**: Compute \( p \)-value,
     - if \( H_a: \mu \neq \mu_0 \), \( p \)-value = \( 2 \cdot \min\{P(T \geq t), P(T \leq t)\} \)
     - if \( H_a: \mu > \mu_0 \), \( p \)-value = \( P(T \geq t) \)
     - if \( H_a: \mu < \mu_0 \), \( p \)-value = \( P(T \leq t) \)
     - reject \( H_0 \) if \( p \)-value < \( \alpha \)

   - **Critical value approach**: Determine critical value(s) using \( \alpha \), reject \( H_0 \) against
     - i) \( H_a: \mu \neq \mu_0 \) if \( t > t_{\alpha/2} \) or \( t < -t_{\alpha/2} \)
     - ii) \( H_a: \mu > \mu_0 \) if \( t > t_{\alpha} \)
     - iii) \( H_a: \mu < \mu_0 \) if \( t < -t_{\alpha} \)

Step 4: Draw conclusion (with a fixed significance level \( \alpha \))

* If sample is not from normal population and sample size is small, try **nonparametric tests** such as Binomial Test (Sign Test), or Wilcoxon Signed-Rank Test. These tests are for testing median. For testing mean, the population distribution needs to be symmetric about its mean and variable needs to continuous or at least ordinal (Signed Test) or interval scale (Signed-Rank Test).

Confidence interval: The (1- \( \alpha \))% confidence interval estimate for population mean is

\[ \bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \]
Two Independent Samples $z$-test and confidence interval estimate

Purpose: Study and compare means of two populations

Assumptions:
1) Two independent random samples from two normal populations with known variances, or
2) Two relatively large (>30) independent random samples (if variances are unknown, it can be approximated by sample variances).

$z$-test:

Step 1: State Hypotheses (choose one of the three pairs hypotheses below)

i) $H_0: \mu_1 = \mu_2$ v.s. $H_a: \mu_1 \neq \mu_2$ (Two-sided test)
ii) $H_0: \mu_1 = \mu_2$ v.s. $H_a: \mu_1 > \mu_2$ (Right-sided test)
iii) $H_0: \mu_1 = \mu_2$ v.s. $H_a: \mu_1 < \mu_2$ (Left-sided test)

Step 2: Compute $z$ test statistic:

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{or} \quad \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{for large } n_1, n_2$$

($\mu_1 - \mu_2 = 0$ when $H_0: \mu_1 = \mu_2$)

Step 3: Decision Rule:

$p$-value approach: Compute $p$-value,
- if $H_a: \mu_1 \neq \mu_2$, $p$-value = $2 \cdot \min\{P( Z \geq z ), P( Z \leq z ) \}$
- if $H_a: \mu_1 > \mu_2$, $p$-value = $P( Z \geq z )$
- if $H_a: \mu_1 < \mu_2$, $p$-value = $P( Z \leq z )$

reject $H_0$ if $p$-value < $\alpha$

Critical value approach: Determine critical value(s) using $\alpha$ ,
reject $H_0$ against

i) $H_a: \mu_1 \neq \mu_2$ if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$
ii) $H_a: \mu_1 > \mu_2$ if $z > z_{\alpha}$
iii) $H_a: \mu_1 < \mu_2$ if $z < -z_{\alpha}$

Step 4: Draw conclusion (with a fixed significance level $\alpha$)

Confidence interval: The (1– $\alpha$)% confidence interval estimate for population mean is

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{or} \quad \frac{\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{for large } n_1, n_2$$
Two Independent Samples \textit{t}-test and confidence interval estimate

Purpose: Study and compare means of two populations
Assumptions: Two independent random samples from two normal populations with unknown variances.

\textit{t}-test:

Step 1: State Hypotheses (choose one of the three hypotheses below)

i) \( H_0 : \mu_1 = \mu_2 \) v.s. \( H_a : \mu_1 \neq \mu_2 \) (Two-sided test)

ii) \( H_0 : \mu_1 = \mu_2 \) v.s. \( H_a : \mu_1 > \mu_2 \) (Right-sided test)

iii) \( H_0 : \mu_1 = \mu_2 \) v.s. \( H_a : \mu_1 < \mu_2 \) (Left-sided test)

A set of more general hypotheses:

\( H_0 : \mu_1 - \mu_2 = D_0 \) v.s. \( H_a : \mu_1 - \mu_2 \neq D_0 \)

\( H_0 : \mu_1 - \mu_2 = D_0 \) v.s. \( H_a : \mu_1 - \mu_2 > D_0 \)

\( H_0 : \mu_1 - \mu_2 = D_0 \) v.s. \( H_a : \mu_1 - \mu_2 < D_0 \)

Step 2: Compute \( t \)-test statistic:

If \( \sigma_1^2 \neq \sigma_2^2 \),

\[
\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2) \sim t\text{-distribution with d.f.} = \min(n_1, n_2)
\]

If \( \sigma_1^2 = \sigma_2^2 \),

\[
\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2) \sim t\text{-distribution with d.f.} = n_1 + n_2 - 2,
\]

where

\[
s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.
\]

\( (\mu_1 - \mu_2 = 0 \text{ when } H_0 : \mu_1 = \mu_2 ) \)

Step 3. Decision Rule:

\textbf{p-value approach:} Compute \( p \)-value,

if \( H_0 : \mu_1 \neq \mu_2 \), \( p \)-value = \( 2 \cdot \min\{ P( T \geq t ) , P( T \leq t ) \} \)

if \( H_0 : \mu_1 > \mu_2 \), \( p \)-value = \( P( T \geq t ) \)

if \( H_0 : \mu_1 < \mu_2 \), \( p \)-value = \( P( T \leq t ) \)

reject \( H_0 \) if \( p \)-value < \( \alpha \)

\textbf{Critical value approach:} Determine critical value(s) using \( \alpha \),

reject \( H_0 \) against

i) \( H_a : \mu_1 \neq \mu_2 \) if \( t > t_{\alpha/2} \) or \( t < -t_{\alpha/2} \)

ii) \( H_a : \mu_1 > \mu_2 \) if \( t > t_\alpha \)

iii) \( H_a : \mu_1 < \mu_2 \) if \( t < -t_\alpha \)

Step 4: Draw conclusion (with a fixed significance level \( \alpha \) )

* If the normality assumption is not met for two independent samples \( t \)-test, try Median Test or Wilcoxon Rank Sum Test (Mann-Whitney U-Test) for testing location difference between two populations.

The \((1-\alpha)\% \) confidence interval estimate for the difference of the two population means is

If \( \sigma_1^2 \neq \sigma_2^2 \),

\[
\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \sqrt{ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} }.
\]

If \( \sigma_1^2 = \sigma_2^2 \),

\[
\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \sqrt{ \frac{s_p^2}{n_1} + \frac{s_p^2}{n_2} }.
\]
Example: (Two Independent Samples)
The number of milligrams of tar in filtered and nonfiltered cigarettes are both normally distributed. A research was done by measuring the number of milligrams of tar in two independent random samples of cigarettes, 9 filtered cigarettes and 10 non-filtered cigarettes. Data were recorded as the following:

<table>
<thead>
<tr>
<th>CIGTYPE</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filtered</td>
<td>9</td>
<td>1.0111</td>
<td>.2667</td>
<td>8.889E-02</td>
</tr>
<tr>
<td>Non-filtered</td>
<td>10</td>
<td>1.2900</td>
<td>.4067</td>
<td>1.286</td>
</tr>
</tbody>
</table>

\[ \bar{x}_1 = 1.0111, \quad s_1 = .2667, \quad \bar{x}_2 = 1.29, \quad s_2 = .4067 \]

At the 0.05 level of significance, test whether the average number of milligrams of tar in filtered is significantly different from non-filtered cigarettes or not.

<table>
<thead>
<tr>
<th>Levene's Test for Equality of Variances</th>
<th>t-test for Equality of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Sig.</td>
</tr>
<tr>
<td>TAR</td>
<td></td>
</tr>
<tr>
<td>Equal variances assumed</td>
<td>1.530</td>
</tr>
<tr>
<td>Equal variances not assumed</td>
<td></td>
</tr>
</tbody>
</table>

Hypothesis: \[ H_0: \mu_1 = \mu_2 \quad \text{v.s.} \quad H_a: \mu_1 \neq \mu_2 \quad \text{(Two-sided test)} \]

Test statistic:

\[ t = \frac{1.0111 - 1.29}{\sqrt{\frac{.12104}{9} + \frac{.12104}{10}}} = -1.745, \quad (s_p^2 = \frac{(9-1)(.2667)^2 + (10-1)(.4067)^2}{9 + 10 - 2} = .12104) \text{, d.f.=17} \]

Conclusion: Since the test of equal variance has an observed \( p \)-value of .233, the equal variance is assumed. The \( p \)-value for testing difference between means is .099, under equal variance assumption, the null hypothesis is not rejected. There is no sufficient evidence to support that there is any difference between the average amount of tar in these two types of cigarettes.

What if the goal is to test whether the filtered cigarettes have less amount of tar than the non-filtered cigarettes?

(To use SPSS, see lab handout.)
(See page 15 for Excel example.)
**Example: (Paired-Sample)**
The following data are weight changes of humans, tabulated after administration of a drug proposed to result in weight loss. Each weight change (in kg) is the weight after minus the weight before drug administration. A random sample of 12 subjects was selected to participate in this research.

<table>
<thead>
<tr>
<th>Subject</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>After</td>
<td>55.2</td>
<td>63.6</td>
<td>58.8</td>
<td>77.2</td>
<td>58.5</td>
<td>69.2</td>
<td>59.5</td>
<td>70.0</td>
<td>68.9</td>
<td>74.0</td>
<td>83.9</td>
<td>74.8</td>
</tr>
<tr>
<td>Before</td>
<td>55.4</td>
<td>63.9</td>
<td>60.1</td>
<td>78.8</td>
<td>59.2</td>
<td>68.7</td>
<td>59.9</td>
<td>70.0</td>
<td>69.2</td>
<td>73.7</td>
<td>84.9</td>
<td>75.3</td>
</tr>
<tr>
<td>Difference</td>
<td>-0.2</td>
<td>-0.3</td>
<td>-1.3</td>
<td>-1.6</td>
<td>-0.7</td>
<td>0.5</td>
<td>-0.4</td>
<td>0.0</td>
<td>-0.3</td>
<td>0.3</td>
<td>-1.0</td>
<td>-0.5</td>
</tr>
</tbody>
</table>

Do these data provide sufficient evidence to indicate that the drug will serve the purpose of weight loss? (at \(\alpha = 0.05\) level of significance)

**Paired-differences** are “After – Before”.
If these differences are mostly negative, and the average of the differences is negative, what does it imply? Weight loss is significant?

\[ H_0 : \bar{\delta} = 0 \text{ v.s. } H_a : \bar{\delta} \neq 0 \quad \text{or} \quad H_0 : \mu_d = 0 \text{ v.s. } H_a : \mu_d \neq 0 \]  
(Two-sided test)

**Test statistic (one-sample t-test):** (Assumption: the paired differences are normally distributed.)

\[
\bar{x} = -0.4583, \quad \text{(average of the differences)}
\]

\[
s_d = 0.6171, \quad \text{(standard deviation of the differences)}
\]

\[
t = \frac{-0.4583}{0.6171/\sqrt{12}} = -2.573, \quad \text{degrees of freedom} = 12 - 1 = 11.
\]

**Critical value** is \(-t_{0.05} = -1.796\), and **p-value** < .025 (approximated using \(t\)-distribution table, since \(t\) score of 2.201 has a \(p\)-value of .025)

**Decision rule**: If \(t \leq -t_{0.05} = -1.796\), or \(p\)-value < .05, then the null hypothesis is rejected.

Conclusion: ____________________________________________

If one wish to test if there is a significance difference in weight loss, then a two-tailed test would be used, and \(p\)-value of a two-tailed test would be less than .05 (i.e., twice of .025), and the critical values would be \(-t_{0.025} = -2.201\), and \(t_{0.025} = 2.201\).

(Paired-sample, Matched-sample, Repeated measures)

(To use SPSS, see lab handout.)
Inference on Proportions

Normal approximation to Binomial Distribution: When \( n \) (sample size) is large, the binomial distribution with probability of success for each Bernoulli trial, \( p \), can be approximated by a normal distribution with a mean \( np \) and a variance \( np(1-p) \).

Sampling distribution of sample proportion: A random sample of size \( n \) from a large population with proportion of successes (usually represented by a value 1) \( p \), and therefore proportion of failures (usually represented by a value 0) \( 1-p \), the sampling distribution of sample proportion, \( \hat{p} = \frac{x}{n} \), where \( x \) is the number of successes in the sample, is approximately normal with a mean \( p \) and standard deviation \( \sqrt{\frac{p(1-p)}{n}} \).

Confidence interval: The \((1-\alpha)\%\) confidence interval estimate for population proportion is

\[
\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

(or \( x \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \), if data are 1’s and 0’s)

Required Sample size to achieve a margin of error \( B \):

\[
n = \left( \frac{z_{\alpha/2}}{B} \right)^2 \hat{p}(1-\hat{p})
\]

where \( \hat{p} \) is from a pilot study or prior study.

Or

\[
n = \left( \frac{z_{\alpha/2}}{B} \right)^2 \cdot .25
\]

where .25 is from .5 \( \times \) .5 that would maximize \( n \)
Example: In a study, 1000 individuals from a community were randomly surveyed and checked whether they were covered by any health insurance during the 12 months. Among them, 122 didn’t have health insurance. Find a 95% confidence interval estimate to estimate the percentage of people in this community who didn’t have health insurance during the past 12 months.

\[ \hat{p} = \frac{122}{1000} = 0.122, \quad \alpha = 0.05, \quad z_{0.025} = 1.96 \]

\[ \hat{p} \pm z_{a/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow 0.122 \pm 1.96 \cdot \sqrt{\frac{0.122(1-0.122)}{1000}} \Rightarrow 0.122 \pm 0.020 \Rightarrow (0.102, 0.142) \]

How to interpret this interval?

Sample Size Example: If one wishes to do a survey to estimate the population proportion with 95% confidence and a margin of error of 3%, how large a sample is needed?

\[ n = \left( \frac{z_{0.025}}{B} \right)^2 \cdot 0.25 = \left( \frac{2.575}{0.3} \right)^2 \cdot 0.25 = 1067.11 \Rightarrow 1068 \]

One-Sample z-test for a population proportion:
Purpose: Study a population proportion.
Assumptions: a large random sample from a population with two categories.

**z-test:**

Step 1: **State Hypotheses** (choose one of the three hypotheses below)

i) \( H_0: p = \mu_0 \) v.s. \( H_a: p \neq \mu_0 \) (Two-sided test)

ii) \( H_0: p = \mu_0 \) v.s. \( H_a: p > \mu_0 \) (Right-sided test)

iii) \( H_0: p = \mu_0 \) v.s. \( H_a: p < \mu_0 \) (Left-sided test)

Step 2: **Compute z test statistic:**

\[ z = \frac{\hat{p} - \mu_0}{\sqrt{\frac{\mu_0(1-\mu_0)}{n}}} \]

Step 3. **Decision Rule:**

**p-value approach:** Compute p-value,

- if \( H_a: \mu \neq \mu_0 \), p-value = 2 \cdot \min\{P(Z \geq z), P(Z \leq -z)\}
- if \( H_a: \mu > \mu_0 \), p-value = P(Z \geq z)
- if \( H_a: \mu < \mu_0 \), p-value = P(Z \leq -z)

Reject \( H_0 \) if p-value < \( \alpha \)

**Critical value approach:** Determine critical value(s) using \( \alpha \),

- i) \( H_a: \mu \neq \mu_0 \) if \( z > z_{0.025} \) or \( z < -z_{0.025} \)
- ii) \( H_a: \mu > \mu_0 \) if \( z > z_{0.025} \)
- iii) \( H_a: \mu < \mu_0 \) if \( z < -z_{0.025} \)

Step 4: **Draw conclusion** (with a fixed significance level \( \alpha \) )
Example: A researcher hypothesized that the percentage of the people living in a community who has insurance coverage during the past 12 months is not 10%. In his study, 1000 individuals from the community were randomly surveyed and checked whether they were covered by any health insurance during the 12 months. Among them, 122 answered “no.” Test the researcher’s hypothesis at the level of significance of 0.05.

Hypothesis: \( H_0 : p = .10 \) v.s. \( H_a : p \neq .10 \) (Two-sided test)

Test Statistic: 
\[
z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.122 - .10}{\sqrt{\frac{.10(1-.10)}{1000}}} = 2.32
\]
\( p \)-value = 2 \times .01 = .02

Decision Rule: Reject null hypothesis if \( p \)-value < .05.

Conclusion: There is sufficient evidence to support the alternative hypothesis that the percentage is different from 10%.

Two Independent Samples z-test for Two Proportions
Purpose: Study and compare proportions of two populations

Step 1: Hypothesis:
1) \( H_0 : p_1 = p_2 \) v.s. \( H_a : p_1 \neq p_2 \)
2) \( H_0 : p_1 = p_2 \) v.s. \( H_a : p_1 > p_2 \)
3) \( H_0 : p_1 = p_2 \) v.s. \( H_a : p_1 < p_2 \)

Step 2: Test Statistic:
If a random sample of size \( n_1 \) from population 1 has \( x_1 \) successes, and a random sample of size \( n_2 \) from population 2 has \( x_2 \) successes, the sample proportions of these two samples are
\[
\hat{p}_1 = \frac{x_1}{n_1} \quad \text{and} \quad \hat{p}_2 = \frac{x_2}{n_2}, \quad \text{and overall sample proportion of successes} \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}
\]
\[
z = \frac{\hat{p}_1 - \hat{p}_2 -(p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad z \text{ has a standard normal distribution if } n_1 \text{ and } n_2 \text{ are large.}
\]

Step 3. Decision Rule:
\( p \)-value approach: Compute \( p \)-value,
- if \( H_a : p_1 \neq p_2 \), \( p \)-value = \( 2 \times \min\{P(Z \geq z), P(Z \leq z)\} \)
- if \( H_a : p_1 > p_2 \), \( p \)-value = \( P(Z \geq z) \)
- if \( H_a : p_1 < p_2 \), \( p \)-value = \( P(Z \leq z) \)
- reject \( H_0 \) if \( p \)-value < \( \alpha \)

Critical value approach: Determine critical value(s) using \( \alpha \), reject \( H_0 \) against
- i) \( H_a : p_1 \neq p_2 \) if \( z > z_{\alpha/2} \) or \( z < -z_{\alpha/2} \)
- ii) \( H_a : p_1 > p_2 \) if \( z > z_{\alpha} \)
- iii) \( H_a : p_1 < p_2 \) if \( z < -z_{\alpha} \)

Step 4: Conclusion: _______________________________________________
**Confidence interval:** The \((1 - \alpha)\%\) confidence interval estimate for the difference of two population proportions is

\[
\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}
\]

**Example:** Suppose a group of researchers wish to test whether the percentage of smokers decreased over a 10-year period. In 1989, among a random sample 1500 adults, 551 were smokers. In 1999, 2000 adults were randomly selected and 652 of them were smokers. Do the data indicate that the percentage of smokers decreased over this 10-year period?

\[
\hat{p}_1 = \frac{551}{1500} = .367 \text{ (from 1989)}
\]
\[
\hat{p}_2 = \frac{652}{2000} = .326 \text{ (from 1999)}
\]
\[
\hat{p} = \frac{551 + 652}{1500 + 2000} = .344 \text{ (overall percentage of smokers)}
\]

**Hypothesis:**

\(H_0: \ p_1 = p_2\) v.s. \(H_a: \ p_1 > p_2\)

**Test Statistic:**

\[
z = \frac{.367 - .326 - 0}{\sqrt{.344(1-.344)\left(\frac{1}{1500} + \frac{1}{2000}\right)}} = 2.53
\]

\(p\)-value = 0.006

**Decision Rule:** Using the level of significance at 0.05, the null hypothesis would be rejected if \(p\)-value is less than 0.05.

**Conclusion:** Since \(p\)-value = 0.006 < 0.05, the null hypothesis is rejected. There is sufficient evidence to support the alternative hypothesis that the percentage of smokers has decreased.

**The 95% confidence interval estimate for the difference of the two population proportions is:**

\[
.367 - .326 \pm 1.96 \cdot \sqrt{\frac{.367(1-.367)}{1500} + \frac{.326(1-.326)}{2000}}
\]

\[\Rightarrow .041 \pm .032\]

\[\Rightarrow (0.009, 0.073)\]
One-Way Analysis of Variance (One-Way ANOVA, study the effect of one factor)
Purpose: Study and compare means for more than two populations

Health index for three ethnic groups

<table>
<thead>
<tr>
<th>CASE I</th>
<th>CASE II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5.90</td>
<td>5.90</td>
</tr>
<tr>
<td>5.51</td>
<td>6.31</td>
</tr>
<tr>
<td>5.01</td>
<td>4.52</td>
</tr>
<tr>
<td>5.92</td>
<td>4.42</td>
</tr>
<tr>
<td>5.50</td>
<td>3.54</td>
</tr>
<tr>
<td>5.00</td>
<td>6.93</td>
</tr>
<tr>
<td>5.89</td>
<td>7.51</td>
</tr>
<tr>
<td>5.50</td>
<td>4.73</td>
</tr>
<tr>
<td>4.99</td>
<td>4.48</td>
</tr>
<tr>
<td>5.91</td>
<td>7.89</td>
</tr>
<tr>
<td>5.49</td>
<td>7.20</td>
</tr>
<tr>
<td>4.98</td>
<td>5.55</td>
</tr>
<tr>
<td>5.88</td>
<td>3.78</td>
</tr>
<tr>
<td>5.49</td>
<td>5.72</td>
</tr>
<tr>
<td>4.98</td>
<td>3.52</td>
</tr>
<tr>
<td>Average</td>
<td>Average</td>
</tr>
<tr>
<td>5.90</td>
<td>5.90</td>
</tr>
<tr>
<td>5.50</td>
<td>5.50</td>
</tr>
<tr>
<td>5.00</td>
<td>5.00</td>
</tr>
</tbody>
</table>

Hypothesis to be tested:

\[ H_0 : \mu_1 = \mu_2 = \mu_3 \]
\[ H_a : \mu_i \neq \mu_j \text{ for at least one pair of } (i,j), i,j = 1, 2, 3. \]

For comparing three independent random samples with two-sample t-test, three t-tests would be needed. Each test suffers a Type I Error rate at \( \alpha \) level. If three tests were used simultaneously, the Type I Error will be inflated.

Sources of variation can be used to form a statistic to test the difference between means.

Total variability = Variability Between Groups + Variability Within Groups
\[ = s^2_B + s^2_W \]
**F test** for difference between means from independent random samples

Assume that there are $k$ population means to be compared.

Notations:

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
<th>...</th>
<th>Group $k$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>$\mu_1$</td>
<td>$\mu_2$</td>
<td>...</td>
<td>$\mu_k$</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>$\sigma_1$</td>
<td>$\sigma_2$</td>
<td>...</td>
<td>$\sigma_k$</td>
</tr>
<tr>
<td><strong>Sample</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>$\bar{x}_1$</td>
<td>$\bar{x}_2$</td>
<td>...</td>
<td>$\bar{x}_k$</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>$s_1$</td>
<td>$s_2$</td>
<td>...</td>
<td>$s_k$</td>
</tr>
<tr>
<td>Sample Size</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>...</td>
<td>$n_k$</td>
</tr>
</tbody>
</table>

**Hypothesis:**

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$

$H_a : \mu_i \neq \mu_j$ for at least one pair of $(i, j)$, $i, j = 1, 2, \ldots, k$.

**Test Statistic:**

$$F = \frac{s^2_w}{s^2_B}$$

Assumptions:

1) Normally distributed populations
2) Independent random samples
3) Equal variances, i.e., $\sigma^2_1 = \sigma^2_2 = \ldots = \sigma^2_k = \sigma^2$

Under $H_0$, $F$ statistic follows an $F$-distribution, $F_{df_1, df_2}$, with degrees of freedom, $df_1 = k - 1$, $df_2 = n - k$.

(Table 5, page A-11.) $k$ is number of groups, $n = n_1 + n_2 + \ldots + n_k$.

Within groups variability: $s^2_w = \frac{(n_1 - 1)s^2_1 + (n_2 - 1)s^2_2 + \ldots + (n_k - 1)s^2_k}{n - k}$, where $n = n_1 + n_2 + \ldots + n_k$, $s^2_w$ is a good estimate of $\sigma^2$.

Between groups variability: $s^2_B = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \ldots + n_k(\bar{x}_k - \bar{x})^2}{k - 1}$, where $\bar{x}$ is the overall mean.

**Decision Rule:** If $F > F_{a, df_1, df_2}$, the null hypothesis would be rejected. This implies that the difference between group means is significant.

**Conclusion:**

____
Example:
The birth weight of an infant has been hypothesized to be associated with the smoking status of the mother during the first trimester of pregnancy. The mothers are divided into four groups according to smoking habit, and the sample of birth weights in pounds within each group is given as follow:

Group 1: (Mother is a nonsmoker)
7.5  6.9  7.4  9.2  8.3  7.6

Group 2: (Mother is an ex-smoker but not during the pregnancy)
5.8  7.1  8.2  7.1  7.8

Group 3: (Mother is a current smoker and smoke less than 1 pack per day)
5.9  6.2  5.8  4.7  7.2  6.2

Group 4: (Mother is a current smoker and smoke more than 1 pack per day)
6.2  6.8  5.7  4.9  6.2  5.8  5.4

Hypothesis:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_a : \mu_i \neq \mu_j$ for at least one pair of $(i, j)$, $i, j = 1, 2, 3, 4$.

Test Statistic:

$$F = \frac{s^2_w}{s^2_B} = \frac{5.537}{.609} = 9.088$$

<table>
<thead>
<tr>
<th>Sample</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\bar{x}_1 = 7.8167$</td>
<td>$\bar{x}_2 = 7.2$</td>
<td>$\bar{x}_3 = 6.0$</td>
<td>$\bar{x}_4 = 5.8571$</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>$s_1 = .8134$</td>
<td>$s_2 = .9138$</td>
<td>$s_3 = .8075$</td>
<td>$s_4 = .6161$</td>
</tr>
<tr>
<td>Sample Size</td>
<td>$n_1 = 6$</td>
<td>$n_2 = 5$</td>
<td>$n_3 = 6$</td>
<td>$n_4 = 7$</td>
</tr>
</tbody>
</table>

Over all mean, $\bar{x} = 6.7222$

Under $H_0$, $F$ statistic follows an $F$-distribution, $F_{df_1, df_2}$, with degrees of freedom, $df_1 = 4 - 1 = 3$, $df_2 = 24 - 4 = 20$. (Table 5, page A-11.)

Within groups variability:

$$s^2_w = \frac{(6-1)s^2_1 + (5-1)s^2_2 + (6-1)s^2_3 + (7-1)s^2_4}{24-4} = .609,$$

where $n = 6 + 5 + 6 + 7 = 24$, $s^2_w$ is a good estimate of $\sigma^2$.

Between groups variability:

$$s^2_B = \frac{6(\bar{x}_1 - \bar{x})^2 + 5(\bar{x}_2 - \bar{x})^2 + 6(\bar{x}_3 - \bar{x})^2 + 7(\bar{x}_4 - \bar{x})^2}{4-1} = 5.537$$

Decision Rule: If $F > F_{0.05, 3, 20} = 3.10$, the null hypothesis would be rejected. This implies that the difference between group means is significant.

Conclusion: The test statistic $F = 9.088 > F_{0.05, 3, 20} = 3.10$, null hypothesis is rejected. There is significant difference between group means.
**SPSS steps:**
Prepare data: Create a variable for birth weights and another ID variable for identifying group.
Perform test:
1) **Analyze ⇒ Compare Means ⇒ One-way ANOVA**
2) Select birth-weight variable as the **dependent** variable and group ID variable as **factor** and click OK.

**SPSS output:**

```
ANOVA

BWEIGHT

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>16.611</td>
<td>3</td>
<td>5.537</td>
<td>9.088</td>
<td>.001</td>
</tr>
<tr>
<td>Within Groups</td>
<td>12.185</td>
<td>20</td>
<td>.609</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>28.796</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

**Post Hoc Analysis (Multiple Comparisons)**
After the first two steps above, click **Post Hoc…** button and select desired comparison procedures. The following output tables are from Bonferroni and Tukey’s-b options.

**Multiple Comparisons**

<table>
<thead>
<tr>
<th>(I) SMOKEST</th>
<th>(J) SMOKEST</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non smoker</td>
<td>Ex-smoker</td>
<td>.6167</td>
<td>.4727</td>
<td>1.000</td>
<td>-.7668 - 2.0002</td>
</tr>
<tr>
<td>Smoker &lt;1</td>
<td>Ex-smoker</td>
<td>1.8167*</td>
<td>.4507</td>
<td>.004</td>
<td>.4975 - 3.1358</td>
</tr>
<tr>
<td>Smoker &gt;1</td>
<td>Ex-smoker</td>
<td>1.9595*</td>
<td>.4343</td>
<td>.001</td>
<td>.6844 - 3.2307</td>
</tr>
<tr>
<td>Ex-smoker</td>
<td>Non smoker</td>
<td>-.6167</td>
<td>.4727</td>
<td>1.000</td>
<td>-2.0002 - .7668</td>
</tr>
<tr>
<td>Smoker &lt;1</td>
<td>Non smoker</td>
<td>1.2000</td>
<td>.4727</td>
<td>.117</td>
<td>-.1835 - 2.5835</td>
</tr>
<tr>
<td>Smoker &gt;1</td>
<td>Non smoker</td>
<td>1.3429*</td>
<td>.4570</td>
<td>.049</td>
<td>5.020E-03 - 2.6807</td>
</tr>
<tr>
<td>Smoker &lt;1</td>
<td>Ex-smoker</td>
<td>-1.8167*</td>
<td>.4507</td>
<td>.004</td>
<td>-3.1358 - -.4975</td>
</tr>
<tr>
<td>Smoker &gt;1</td>
<td>Ex-smoker</td>
<td>-1.2000</td>
<td>.4727</td>
<td>.117</td>
<td>-2.5835 - .1835</td>
</tr>
<tr>
<td>Non smoker</td>
<td>Ex-smoker</td>
<td>-1.1283</td>
<td>.4343</td>
<td>1.000</td>
<td>-1.4140 - 1.1283</td>
</tr>
<tr>
<td>Ex-smoker</td>
<td>Smoker &gt;1</td>
<td>-1.9595*</td>
<td>.4343</td>
<td>.001</td>
<td>-3.2307 - -.6884</td>
</tr>
<tr>
<td>Smoker &lt;1</td>
<td>Ex-smoker</td>
<td>-1.3429*</td>
<td>.4570</td>
<td>.049</td>
<td>-2.6807 - 5.0200E-03</td>
</tr>
<tr>
<td>Smoker &gt;1</td>
<td>Ex-smoker</td>
<td>-.1429</td>
<td>.4343</td>
<td>1.000</td>
<td>-1.1283 - 1.1283</td>
</tr>
<tr>
<td>Non smoker</td>
<td>Ex-smoker</td>
<td>-1.4140</td>
<td>.4343</td>
<td></td>
<td>-1.1283 - 1.1283</td>
</tr>
</tbody>
</table>

*: The mean difference is significant at the .05 level.
BWEIGHT

<table>
<thead>
<tr>
<th>SMOKEST</th>
<th>N</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tukey B</td>
<td>7</td>
<td>5.8571</td>
<td></td>
</tr>
<tr>
<td>Smoker &gt;1</td>
<td>6</td>
<td>6.0000</td>
<td></td>
</tr>
<tr>
<td>Smoker &lt;1</td>
<td>6</td>
<td>7.2000</td>
<td></td>
</tr>
<tr>
<td>Ex-smoker</td>
<td>5</td>
<td></td>
<td>7.8167</td>
</tr>
<tr>
<td>Non smoker</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 5.915.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Excel Two Independent Samples t-Test Example

Step 1: enter data of the two samples into column A (sample 1) and column B (sample 2)

Step 2: (testing for equal variances) Click on Tools and select Data Analysis and choose “F-test Two Sample for Variances” and click OK, and select the range of data for the two sets of data to be tested and click OK.

Step 3: (if equal variance is assumed) Click on Tools and select Data Analysis and choose “t-test: Two Samples Assumed Equal Variances” and click OK, and select the range of data for the two sets of data to be tested and click OK.