

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

Step 3: **Decision Rule :**

p-value approach: Compute p-value,

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

if $H_a : \mu > \mu_0$, $p\text{-value} = P(Z \geq z)$

if $H_a : \mu < \mu_0$, $p\text{-value} = P(Z \leq z)$

reject H_0 if $p\text{-value} < \alpha$

Critical value approach: Determine critical value(s) using α ,
reject H_0 against

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

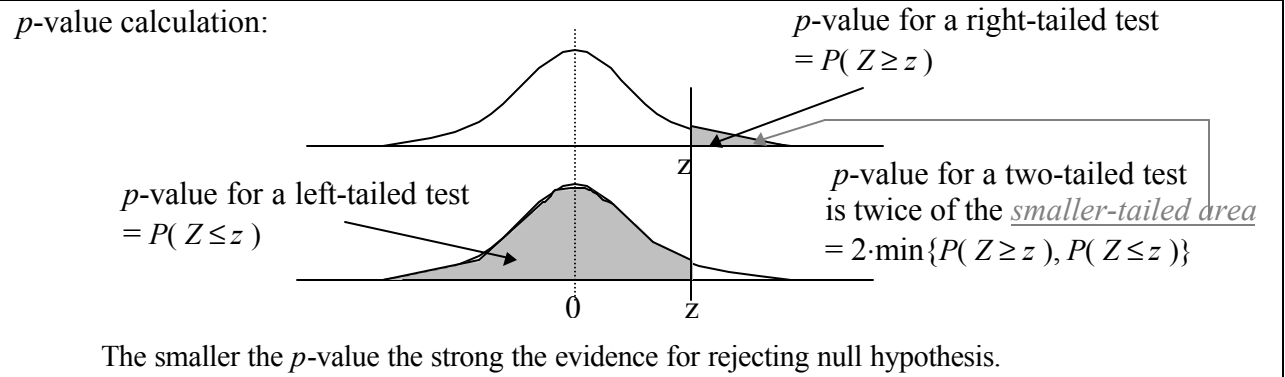
ii) $H_a : \mu > \mu_0$ if $z > z_\alpha$

iii) $H_a : \mu < \mu_0$ if $z < -z_\alpha$

Step 4: **Draw conclusion** (with a fixed significance level α)

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 $H_a : \mu \neq \mu_0$, $p\text{-value} = 2 \cdot \min\{P(T \geq t), P(T \leq t)\}$

if $H_a : \mu > \mu_0$, $p\text{-value} = P(T \geq t)$

if $H_a : \mu < \mu_0$, $p\text{-value} = P(T \leq t)$

reject H_0 if $p\text{-value} < \alpha$

Critical value approach: Determine critical value(s) using α ,
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 α)

* 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 be 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:**

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad \text{or} \quad \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (\text{for large } n_1, n_2)$$

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

Step 3: **Decision Rule :**

p-value approach: Compute p -value,
 if $H_a : \mu_1 \neq \mu_2$, $p\text{-value} = 2 \cdot \min\{P(Z \geq z), P(Z \leq z)\}$
 if $H_a : \mu_1 > \mu_2$, $p\text{-value} = P(Z \geq z)$
 if $H_a : \mu_1 < \mu_2$, $p\text{-value} = P(Z \leq z)$
reject H_0 if $p\text{-value} < \alpha$

Critical value approach: Determine critical value(s) using α ,
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 α)

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

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{or} \quad \bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (\text{for large } n_1, n_2)$$

Two Independent Samples *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.

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:

$$H_0 : \mu_1 - \mu_2 = D_0 \text{ v.s. } H_a : \mu_1 - \mu_2 \neq D_0$$

$$H_0 : \mu_1 - \mu_2 = D_0 \text{ v.s. } H_a : \mu_1 - \mu_2 > D_0$$

$$H_0 : \mu_1 - \mu_2 = D_0 \text{ v.s. } H_a : \mu_1 - \mu_2 < D_0$$

Step 2: **Compute *t* test statistic :**

$$\text{If } \sigma_1^2 \neq \sigma_2^2, \quad t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t\text{-distribution with d.f.} = \min(n_1, n_2)$$

a rough estimate see textbook
for a more accurate estimate

$$\text{If } \sigma_1^2 = \sigma_2^2, \quad t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \sim t \text{ distribution with d.f.} = n_1 + n_2 - 2,$$

$$\text{where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

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

Step 3: **Decision Rule :**

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

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

if $H_a : \mu_1 > \mu_2$, $p\text{-value} = P(T \geq t)$

if $H_a : \mu_1 < \mu_2$, $p\text{-value} = P(T \leq t)$

reject H_0 if $p\text{-value} < \alpha$

Critical value approach: Determine critical value(s) using α ,

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 α)

* 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

$$\text{If } \sigma_1^2 \neq \sigma_2^2, \quad \bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \quad \text{If } \sigma_1^2 = \sigma_2^2, \quad \bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \cdot \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}.$$

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:

Filtered 0.9 1.1 1.2 0.8 1.6 0.9 0.7 1.0 0.9
Nonfiltered 1.5 0.9 1.6 0.5 1.4 1.9 1.0 1.3 1.2 1.6

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.

Group Statistics					
	CIGTYPE	N	Mean	Std. Deviation	Std. Error Mean
TAR	Filtered	9	1.0111	.2667	8.889E-02
	Non-filtered	10	1.2900	.4067	.1286

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

Independent Samples Test									
		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower Upper
TAR	Equal variances assumed	1.530	.233	-1.745	17	.099	-.2789	.1599	-.6162 5.839E-02
	Equal variances not assumed			-1.784	15.637	.094	-.2789	.1564	-.6110 5.319E-02

Hypothesis: $H_0 : \mu_1 = \mu_2$ v.s. $H_a : \mu_1 \neq \mu_2$ (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.

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

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?

$H_0 : \delta = 0$ v.s. $H_a : \delta \neq 0$ or $H_0 : \mu_d = 0$ v.s. $H_a : \mu_d \neq 0$ (Two-sided test)

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

$\bar{x}_d = -.4583$, (average of the differences)

$s_d = .6171$, (standard deviation of the differences)

$$t = \frac{-.4583}{.6171 / \sqrt{12}} = -2.573, \text{ degrees of freedom} = 12 - 1 = 11.$$

Critical value is $-t_{.05} = -1.796$, and **p -value** $< .025$ (approximated using t -distribution table, since t score of 2.201 has a p -value of .025)

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

Conclusion: _____

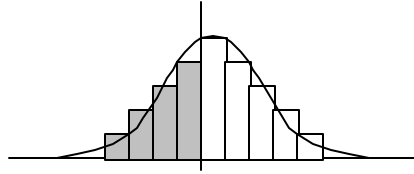
If one wish to test if there is a significance difference in weight loss, then a two-tailed test would be used, and p -value of a two-tailed test would be less than .05 (i.e., twice of .025), and the critical values would be $-t_{.025} = -2.201$, and $t_{.025} = 2.201$.

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

(To use SPSS, see SPSS link on our class webpage.)

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** $\sqrt{\frac{p(1-p)}{n}}$.

Note: In $\hat{p} = \frac{x}{n}$, x is the sum of data in the sample contains 1's and 0's. Therefore, \hat{p} is the same as sample mean, and the confidence interval estimate for mean and the hypothesis test for mean can all be used for estimating and hypothesis testing for proportion, except that the standard error in the formula can be replaced by $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

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

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (\text{or } \bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \text{ if data are 1's and 0's})$$

Required Sample size to achieve a margin of error B :

$$n = \left(\frac{z_{\alpha/2}}{B} \right)^2 \hat{p}(1-\hat{p}), \text{ where } \hat{p} \text{ is from a pilot study or prior study.}$$

Or

$$n = \left(\frac{z_{\alpha/2}}{B} \right)^2 \cdot .25, \text{ where } .25 \text{ is from } .5 \times .5 \text{ that would maximize } n$$

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.

$$\hat{p} = 122/1000 = .122, \alpha = .05, z_{\alpha/2} = z_{.025} = 1.96$$

$$\begin{aligned} \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &\Rightarrow .122 \pm 1.96 \cdot \sqrt{\frac{.122(1-.122)}{1000}} \\ &\Rightarrow .122 \pm .020 \Rightarrow (10.2\%, 14.2\%) \quad \text{How to interpret this interval?} \end{aligned}$$

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?

$$n = \left(\frac{z_{\alpha/2}}{B} \right)^2 \cdot .25 = \left(\frac{1.96}{.03} \right)^2 \cdot .25 = 1067.11 \Rightarrow 1068$$

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) $H_0 : p = p_0$ v.s. $H_a : p \neq p_0$ (Two-sided test)
- ii) $H_0 : p = p_0$ v.s. $H_a : p > p_0$ (Right-sided test)
- iii) $H_0 : p = p_0$ v.s. $H_a : p < p_0$ (Left-sided test)

Step 2: **Compute z test statistic:**
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Step 3: **Decision Rule :**

p-value approach: Compute p-value,

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

if $H_a : \mu > \mu_0$, $p\text{-value} = P(Z \geq z)$

if $H_a : \mu < \mu_0$, $p\text{-value} = P(Z \leq z)$

reject H_0 if $p\text{-value} < \alpha$

Critical value approach: Determine critical value(s) using α ,

reject H_0 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 α)

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:
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.122 - .10}{\sqrt{\frac{.10(1-.10)}{1000}}} = 2.32 \quad p\text{-value} = 2 \times .01 = .02$$

Decision Rule: Reject null hypothesis if $p\text{-value} < .05$.

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

$$\hat{p}_1 = \frac{x_1}{n_1} \quad \text{and} \quad \hat{p}_2 = \frac{x_2}{n_2}, \quad \text{and overall sample proportion of successes } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

$$z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad z \text{ has a standard normal distribution if } n_1 \text{ and } n_2 \text{ are large.}$$

Step 3: **Decision Rule:**

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

if $H_a : p_1 \neq p_2$, $p\text{-value} = 2 \cdot \min\{P(Z \geq z), P(Z \leq z)\}$

if $H_a : p_1 > p_2$, $p\text{-value} = P(Z \geq z)$

if $H_a : p_1 < p_2$, $p\text{-value} = P(Z \leq z)$

reject H_0 if $p\text{-value} < \alpha$

Critical value approach: Determine critical value(s) using α ,
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?

$$\hat{p}_1 = 551/1500 = .367 \text{ (from 1989)}$$

$$\hat{p}_2 = 652/2000 = .326 \text{ (from 1999)}$$

$$\hat{p} = (551 + 652)/(1500 + 2000) = .344 \text{ (overall percentage of smokers)}$$

Hypothesis:

$$H_0: p_1 = p_2 \quad \text{v.s.} \quad H_a: p_1 > p_2$$

Test Statistic:

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

$$p\text{-value} = 0.006$$

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 $p\text{-value} = 0.006 < 0.05$, 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:

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

$$\Rightarrow .041 \pm .032$$

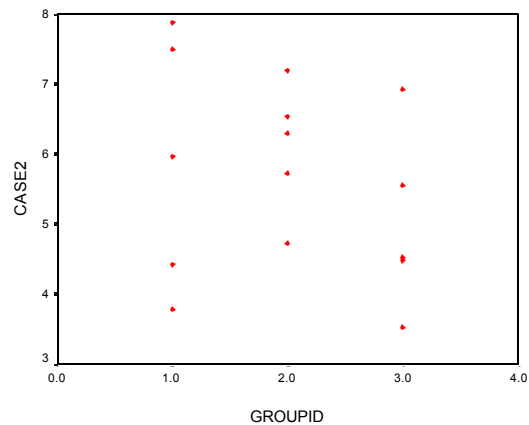
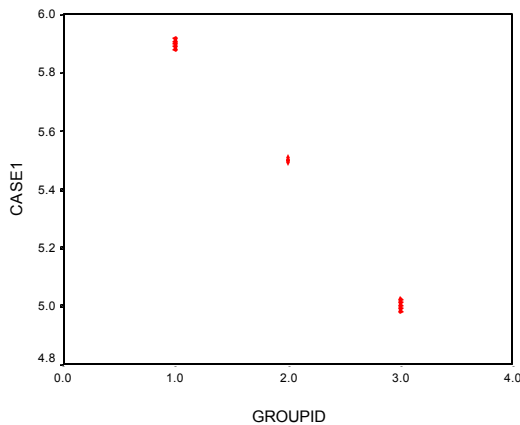
$$\Rightarrow (.009, .073)$$

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

	<u>CASE I</u>				<u>CASE II</u>		
	1	2	3		1	2	3
	5.90	5.51	5.01		5.90	6.31	4.52
	5.92	5.50	5.00		4.42	3.54	6.93
	5.89	5.50	4.99		7.51	4.73	4.48
	5.91	5.49	4.98		7.89	7.20	5.55
	<u>5.88</u>	<u>5.50</u>	<u>5.02</u>		<u>3.78</u>	<u>5.72</u>	<u>3.52</u>
Average	5.90	5.50	5.00		5.90	5.50	5.00



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 α 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 **B**etween Groups + Variability **W**ithin Groups

$$= s_B^2 + s_W^2$$

F test for difference between means from independent random samples

Assume that there are k population means to be compared.

Notations:

		Group 1	Group 2	...	Group k
Population	Mean	μ_1	μ_2	...	μ_k
	Standard Deviation	σ_1	σ_2	...	σ_k
Sample	Mean	\bar{x}_1	\bar{x}_2	...	\bar{x}_k
	Standard Deviation	s_1	s_2	...	s_k
	Sample Size	n_1	n_2	...	n_k

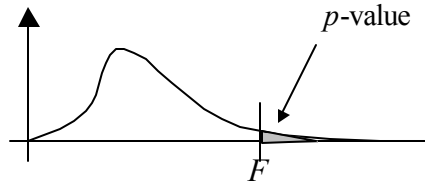
Hypothesis:

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

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

Test Statistic:

$$F = \frac{s_B^2}{s_W^2}$$



Assumptions:

- 1) Normally distributed populations
- 2) Independent random samples
- 3) Equal variances, i.e., $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$

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, $n = n_1 + n_2 + \dots + n_k$.

Within groups variability:
$$s_W^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_k - 1)s_k^2}{n - k},$$

where $n = n_1 + n_2 + \dots + n_k$, s_W^2 is a good estimate of σ^2 .

Between groups variability:
$$s_B^2 = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \dots + n_k(\bar{x}_k - \bar{x})^2}{k - 1},$$

where \bar{x} is the overall mean.

Decision Rule : If $F > F_{\alpha, df_1, df_2}$, 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:

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

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

Test Statistic:

$$F = \frac{s_B^2}{s_W^2} = \frac{5.537}{.609} = 9.088$$

		Group 1	Group 2	Group 3	Group 4
Sample	Mean	$\bar{x}_1 = 7.8167$	$\bar{x}_2 = 7.2$	$\bar{x}_3 = 6.0$	$\bar{x}_4 = 5.8571$
	Standard Deviation	$s_1 = .8134$	$s_2 = .9138$	$s_3 = .8075$	$s_4 = .6161$
	Sample Size	$n_1 = 6$	$n_2 = 5$	$n_3 = 6$	$n_4 = 7$
Over all mean, $\bar{x} = 6.7222$					

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

Within groups variability: $s_W^2 = \frac{(6-1)s_1^2 + (5-1)s_2^2 + (6-1)s_3^2 + (7-1)s_4^2}{24-4}$

$$s_W^2 = \frac{(6-1).8134^2 + (5-1).9138^2 + (6-1).8075^2 + (7-1).6161^2}{24-4} = .609,$$

where $n = 6 + 5 + 6 + 7 = 24$, s_W^2 is a good estimate of σ^2 .

Between groups variability: $s_B^2 = \frac{6(\bar{x}_1 - \bar{x})^2 + 5(\bar{x}_2 - \bar{x})^2 + 6(\bar{x}_3 - \bar{x})^2 + 7(\bar{x}_4 - \bar{x})^2}{4-1}$

$$s_B^2 = \frac{6(7.8167 - 6.7222)^2 + 5(7.2 - 6.7222)^2 + 6(6.0 - 6.7222)^2 + 7(5.8571 - 6.7222)^2}{4-1} = 5.537$$

Decision Rule : If $F > F_{.05, 3, 20} = 3.10$, the null hypothesis would be rejected. This implies that the difference between group means is significant.

Conclusion: The test statistic $F = 9.088 > F_{.05, 3, 20} = 3.10$, and p -value < 0.001 , since $F_{.001, 3, 20} = 8.10 < 9.088$, 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					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	16.611	3	5.537	9.088	.001
Within Groups	12.185	20	.609		
Total	28.796	23			

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

Dependent Variable: BWEIGHT

Bonferroni

(I) SMOKEST	(J) SMOKEST	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Non smoker	Ex-smoker	.6167	.4727	1.000	-.7668	2.0002
	Smoker <1	1.8167*	.4507	.004	.4975	3.1358
	Smoker >1	1.9595*	.4343	.001	.6884	3.2307
Ex-smoker	Non smoker	-.6167	.4727	1.000	-2.0002	.7668
	Smoker <1	1.2000	.4727	.117	-.1835	2.5835
	Smoker >1	1.3429*	.4570	.049	5.020E-03	2.6807
Smoker <1	Non smoker	-1.8167*	.4507	.004	-3.1358	-.4975
	Ex-smoker	-1.2000	.4727	.117	-2.5835	.1835
	Smoker >1	.1429	.4343	1.000	-1.1283	1.4140
Smoker >1	Non smoker	-1.9595*	.4343	.001	-3.2307	-.6884
	Ex-smoker	-1.3429*	.4570	.049	-2.6807	-5.0200E-03
	Smoker <1	-.1429	.4343	1.000	-1.4140	1.1283

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

BWEIGHT				
Subset for alpha = .05				
SMOKEST	N	1	2	
Tukey B ^{a,b} Smoker >1	7	5.8571		
Smoker <1	6	6.0000		
Ex-smoker	5		7.2000	
Non smoker	6		7.8167	

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 Assuming Equal Variances” and click OK, and select the range of data for the two sets of data to be tested and click OK.

	A	B	C	D	E	F	G	H	I	J
1	0.9	1.5	F-Test Two-Sample for Variances					t-Test: Two-Sample Assuming Equal Variances		
2	1.1	0.9								
3	1.2	1.6		Variable 1	Variable 2			Variable 1	Variable 2	
4	0.8	0.5	Mean	1.011111	1.29			Mean	1.011111	1.29
5	1.6	1.4	Variance	0.071111	0.165444			Variance	0.071111	0.165444
6	0.9	1.9	Observations	9	10			Observations	9	10
7	0.7	1.0	df	8	9			Pooled Variance	0.121052	
8	1.0	1.3	F	0.429819				Hypothesized Mean	0	
9	0.9	1.2	P(F<=f) one-tail	0.124272				df	17	
10		1.6	F Critical one-tail	0.295148				t Stat	-1.74457	
11								P(T<=t) one-tail	0.049555	
12								t Critical one-tail	1.739606	
13								P(T<=t) two-tail	0.09911	
14								t Critical two-tail	2.189819	

Variable 1

Variable 2

p-value of F-test

(if two-tailed test) p-value of t-Test

Multiple Comparisons using Confidence Interval Estimate for difference of two means with s_w^2 as the pooled estimate of common variance for multiple comparisons and with **Bonferroni correction**:

$$\boxed{\bar{x}_1 - \bar{x}_2 \pm t_{\alpha^*/2} \cdot \sqrt{\frac{s_w^2}{n_1} + \frac{s_w^2}{n_2}}} \text{ where } \alpha^* = \frac{\alpha}{\text{Number of pairs to be compared}} .$$

Number of pairs to be compared is $\binom{k}{2} = \frac{k!}{2!(k-2)!}$

Example : (from the previous problem about smoking mothers)

For comparing “Smoke < 1” & “Smoke > 1”

$\alpha = 0.05$, $\alpha^* = 0.05 / \{4! / (2!2!)\} = .05/6 = .008$, $t_{\alpha^*/2} = t_{.004} \approx 2.9$, (degrees of freedom = 20)
 $s_w^2 = .609$.

$$5.8571 - 6.000 \pm 2.9 \cdot \sqrt{\frac{.609}{6} + \frac{.609}{5}} \Rightarrow (-1.5133, 1.2275)$$

This interval does contain zero. It implies the difference between the means of “Smoke < 1” & “Smoke > 1” groups is insignificant.
