Sections 5.4, 5.5 and 5.6: CIs for the difference between two proportions and two means

**The 100(1 − α)% Confidence Interval for a difference in two proportions: p1-p2, large samples.**

Let $X$ be the number of successes in $n_x$ independent Bernoulli trials with success probability $p_x$, and let $Y$ be the number of successes in $n_y$ independent Bernoulli trials with success probability $p_y$, so that $X \sim \text{Bin}(n_x, p_x)$ and $Y \sim \text{Bin}(n_y, p_y)$. Assume large samples, same as for CI for one proportion.

Define $\tilde{n}_x = n_x + 2$, $\tilde{n}_y = n_y + 2$

Then, the 100(1 − α)% CI for the difference $p_x - p_y$ is

$$\tilde{p}_x - \tilde{p}_y \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}_x(1-\tilde{p}_x)}{\tilde{n}_x} + \frac{\tilde{p}_y(1-\tilde{p}_y)}{\tilde{n}_y}}$$

**NOTE1:** If the lower limit of the CI is less than −1, replace it with −1. If the upper limit of the CI is greater than 1, replace it with 1.

**NOTE2:** There is a traditional CI as well. It is a generalization of the one for a single proportion. It uses the above formula with standard values for $n_x$ and $n_y$, and standard estimators of the population proportion: $\hat{p}_x$ and $\hat{p}_y$. MINITAB uses this standard CI.

**Confidence interval for paired data: section 5.7 read yourself.**
CIs for a difference of means in two populations, independent samples.

Case 1: Population st. devs. $\sigma_x$ and $\sigma_y$ known:
Let $X_1, X_2, \ldots, X_{n_x}$ be a large random sample of size $n_x$ from a population with mean $\mu_X$ and standard deviation $\sigma_X$ and let $Y_1, Y_2, \ldots, Y_{n_y}$ be a large random sample of size $n_y$ from a population with mean $\mu_Y$ and standard deviation $\sigma_Y$. If the sample sizes are not both large, we have to assume normal populations. Then a level $100(1 - \alpha)$% CI for $\mu_X - \mu_Y$ is

$$\bar{X} - \bar{Y} \pm z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_y}}.$$ 

Case 2: Population st. devs. $\sigma_x$ and $\sigma_y$ NOT known:
Let $X_1, X_2, \ldots, X_{n_x}$ be a random sample of size $n_x$ from a normal population with mean $\mu_X$ and standard deviation $\sigma_X$ and let $Y_1, Y_2, \ldots, Y_{n_y}$ be a random sample of size $n_y$ from a normal population with mean $\mu_Y$ and standard deviation $\sigma_Y$.

**A: Populations do not necessarily have the same variance.** Then, a level $100(1 - \alpha)$% CI for $\mu_X - \mu_Y$ is

$$\bar{X} - \bar{Y} \pm t_{v, \alpha/2} \sqrt{\frac{s_X^2}{n_x} + \frac{s_Y^2}{n_y}}.$$ 

The number of degrees of freedom, $v$, is given by (rounded down to the nearest integer):

$$v = \frac{\left(\frac{s_X^2}{n_x} + \frac{s_Y^2}{n_y}\right)^2}{\frac{s_X^2}{n_x}/(n_x - 1) + \frac{s_Y^2}{n_y}/(n_y - 1)}.$$ 

**NOTE:** When the sample sizes are large, we can replace the percentile from t-distribution with the corresponding percentile from standard normal distribution.

**B: Populations have nearly the same variance.** Then, a level $100(1 - \alpha)$% CI for $\mu_X - \mu_Y$ is

$$\bar{X} - \bar{Y} \pm t_{n_x + n_y - 2, \alpha/2} \cdot s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}.$$ 

The quantity $s_p$ is the pooled standard deviation, given by

$$s_p = \sqrt{\frac{(n_x - 1)s_X^2 + (n_y - 1)s_Y^2}{n_x + n_y - 2}}.$$ 

**NOTE:** When the sample sizes are large, we can replace the percentile from t-distribution with the corresponding percentile from standard normal distribution.
Example 1. Let $\mu_X$ denote true average tread life for a premium brand of radial tire and let $\mu_Y$ denote the true average tread life for an economy brand of the same size. Compute 90% CI for the difference of the mean thread lives for the two types of tires: $\mu_X - \mu_Y$. The following information was computed from the samples: $n_X=50$, $\bar{x} = 43,000$, $s_X=2200$, and $n_Y=50$, $\bar{y} = 37,000$, $s_Y=1500$.

Example 2. Suppose $\mu_X$ and $\mu_Y$ are true mean stopping distances in m at 50 mph for cars of a certain type equipped with two different types of braking systems. Assume normal distribution of the stopping distances. Find a 95% CI for $\mu_X - \mu_Y$ using the following statistics: $n_X=6$, $\bar{x} = 116$, $s_X=5.0$, and $n_Y=6$, $\bar{y} = 129$, $s_Y=5.5$.

Example 3. A random sample of 5726 telephone numbers from a certain region taken in March 2002 yielded 1105 that were unlisted, and 1 year later a sample of 5384 yielded 980 unlisted numbers. Find a 98% CI for the difference in true proportions of unlisted numbers between the two years.

**EXAMPLES’ SOLUTIONS IN MINITAB**

**EXAMPLE 1:** Use: Stat, Basic Statistics, 2 sample t: choose "summarized data"

Options: confidence at 95.0

Results: Two-Sample T-Test and CI

<table>
<thead>
<tr>
<th>Sample</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>43000</td>
<td>2200</td>
<td>311</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>37000</td>
<td>1500</td>
<td>212</td>
</tr>
</tbody>
</table>

Difference = mu (1) - mu (2)

Estimate for difference: 6000

90% CI for difference: (5374, 6626)

T-Test of difference = 0 (vs not =): T-Value = 15.93  P-Value = 0.000  DF = 86

**EXAMPLE 2:** Make sure you have put confidence at 95.0 in Options

Two-Sample T-Test and CI

<table>
<thead>
<tr>
<th>Sample</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>116.00</td>
<td>5.00</td>
<td>2.0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>129.00</td>
<td>5.50</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Difference = mu (1) - mu (2)

Estimate for difference: -13.00

95% CI for difference: (-19.86, -6.14)

T-Test of difference = 0 (vs not =): T-Value = -4.28  P-Value = 0.002  DF = 9

**EXAMPLE 3:** Use: Stat, Basic Statistics, 2 Proportions: choose "summarized data"

Options: confidence at 95.0

Results: Test and CI for Two Proportions

<table>
<thead>
<tr>
<th>Sample</th>
<th>X</th>
<th>N</th>
<th>Sample p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1105</td>
<td>5726</td>
<td>0.192979</td>
</tr>
<tr>
<td>2</td>
<td>980</td>
<td>5384</td>
<td>0.182021</td>
</tr>
</tbody>
</table>

Difference = p (1) - p (2)

Estimate for difference: 0.0109586

98% CI for difference: (-0.00627091, 0.0281881)

Test for difference = 0 (vs not =): Z = 1.48  P-Value = 0.139

Fisher's exact test: P-Value = 0.145