Example:

A study of teenage suicide included a sample of 96 boys and 123 girls between ages of 12 and 16 years selected scientifically from admissions records to a private psychiatric hospital. Suicide attempts were reported by 18 of the boys and 60 of the girls. We assume that the girls constitute a simple random sample from a population of similar girls and likewise for the boys. Construct a 99 percent confidence interval for the difference between the two proportions.

Solution:

(1) Given

$n_1 = 123 \qquad n_2 = 96$

$\hat{p}_1 = .4878 \qquad \hat{p}_2 = .1875$

(2) Calculation

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

$$(.3003) \pm 2.575 \sqrt{\frac{(.4878)(.5122)}{123} + \frac{(.1875)(.8125)}{96}}$$

$$(.3003) \pm 2.575\ (.0602)$$

$$(.1453,\ .4553)$$

Discussion: We interpret this interval that the difference between the two population proportions, $p_1 - p_2$, is .3003. We are 99% confident that the true value of the difference between the two population proportions lies between .1435 and .4553.

Example:

What is the prevalence of anemia in developing countries?

	African Women	Women from Americas
Sample size	2100	1900
Number with anemia	840	323
Sample proportion	$840/2100 = 0.40$	$323/1900 = 0.17$

Find a 95% confidence interval for the difference in proportions of all African women with anemia and all women from the Americas with anemia.

Solution. Let's start by simply defining some notation. Let:

- n_1 = the number of African women sampled = 2100
- n_2 = the number of women from the Americas sampled = 1900
- y_1 = the number of African women with anemia = 840
- y_2 = the number of women from the Americas with anemia = 323

Based on these data, we can calculate two sample proportions. The proportion of African women sampled who have anemia is:

$$\hat{p}_1 = \frac{840}{2100} = 0.40$$

And the proportion of women from the Americas sampled who have anemia is:

$$\hat{p}_2 = \frac{323}{1900} = 0.17$$

Now, letting:

- p_1 = the proportion of all African women with anemia
- p_2 = the proportion of all women from the Americas with anemia

we are then interested in finding a 95% confidence interval for $p_1 - p_2$, the difference in the two population proportions. We need to derive a formula for the confidence interval before we can actually calculate it!

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

Solution. Substituting in the numbers that we know into the formula for a 95% confidence interval for $p_1 - p_2$, we get:

$$\mathbf{(0.40 - 0.17) \pm 1.96\sqrt{\frac{0.40(0.60)}{2100} + \frac{0.17(0.83)}{1900}}}$$

which simplifies to:

$$0.23 \pm 0.027 = (0.203, 0.257)$$

We can be 95% confident that there are between 20.3% and 25.7% more African women with anemia than women from the Americas with anemia.

Example:

A social experiment conducted in 1962 involved $n = 123$ three- and four-year-old children from poverty-level families in some locality, Michigan. The children were randomly assigned either to (1) a treatment group receiving two years of preschool instruction, or to (2) a control group receiving no preschool instruction. The participants were followed into their adult years. Here is a summary of the data:

	Arrested for some crime	
	Yes	**No**
Control	32	30
Preschool	19	42

Find CI at 95% Confidence for difference of proportions.

Solution. Of the $n_1 = 62$ children serving as the control group, 32 were later arrested for some crime, yielding a sample proportion of:

$$\hat{p}_1 = 0.516$$

And, of the $n_2 = 61$ children receiving preschool instruction, 19 were later arrested for some crime, yielding a sample proportion of:

$$\hat{p}_2 = 0.311$$

A 95% confidence interval for $p_1 - p_2$ is therefore:

$$(0.516 - 0.311) \pm 1.96\sqrt{\frac{0.516(0.484)}{62} + \frac{0.311(0.689)}{61}}$$

which simplifies to:

$$0.205 \pm 0.170 = (0.035, 0.375)$$

We can be 95% confident that between 3.5% and 37.5% more children not having attended preschool were arrested for a crime by age 19 than children who had received preschool instruction.

Example: Independent Proportions

A survey was given to a sample of college students. They were asked whether they think same sex marriage should be legal. Of the 251 females in the sample, 185 said "yes." Of the 199 males in the sample, 107 said "yes." Let's construct a 95% confidence interval for the difference of the proportion of females and males who responded "yes."

$$\hat{p}_f = \frac{185}{251} = .737$$

$$\hat{p}_m = \frac{107}{199} = .538$$

In order to keep $\hat{p}_1 - \hat{p}_2$ positive, let's say that females are group 1 and males are group 2.

$$z_{.05/2} = 1.96$$

$$(\hat{p}_1 - \hat{p}_2) \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(.737 - .538) \pm 1.96\sqrt{\frac{.737(1-.737)}{251} + \frac{.538(1-.538)}{199}} = .199 \pm 1.96(.0450) = .199 \pm .0881 = [.111, .287]$$

We are 95% confident that in the population the difference between the proportion of females and males who respond "yes" to the question concerning same sex marriage is between .111 and .287.

Note that this confidence interval does not contain 0, therefore it is not likely that the difference between females and males is 0. We conclude that there is a difference between the proportion of males and females in the population who respond "yes."

Example: Independent Means (Pooled Standard Deviations)

A team of pediatricians want to estimate with 95% confidence the mean difference between males and females in terms of BMI at 26 months of age. For $n = 1,175$ males, the mean BMI is 16.46 with a standard deviation of 1.17. For $n = 1,250$ females, the mean BMI is 16.29 with a standard deviation of 1.25. (Data derived from http://www.cdc.gov/growthcharts/zscore.htm table 8)

The standard deviations of the two groups are similar, neither is more than twice of the other. Thus, pooled standard deviation methods will be used.

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(1175-1)1.17^2+(1250-1)1.25^2}{1175+1250-2}} = 1.21$$

$$df = n_1 + n_2 - 2 = 1175 + 1250 - 2 = 2423$$

For a 95% confidence interval: $t_{df=2423} \approx 1.96$

$$(\overline{x}_1 - \overline{x}_2) \pm t\,(s_p)\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = (16.46 - 16.29) \pm 1.96\,(1.21)\sqrt{\frac{1}{1175}+\frac{1}{1250}} = 0.17 \pm 0.096 = [0.074, 0.2$$

We are 95% confident that in the population, the mean difference between males and females is between 0.074 and 0.266.

Note that this confidence interval does not contain 0. Therefore, we could conclude that it is unlikely that the mean BMI of males and females in the population are the same.

Example: Independent Means (Unpooled Standard Deviations)

Medical researchers want to estimate with 95% confidence the difference in the average weights of boys at 2 years and 10 years. In $n = 200$ 2 year old boys, the mean weight was 12.671 kg with a standard deviation of 1.433 kg. In $n = 150$ 10 year old boys, the mean weight was 32.088 kg with a standard deviation of 6.633 kg. (Data derived from http://www.cdc.gov/growthcharts/html_charts/wtage.htm)

The 95% confidence interval for the difference in means is [18.3288, 20.5052]. We are 95% confident that in the population the mean difference in weight between 10 year old and 2 year old boys is between 18.3288 kg and 20.5052 kg.

Two Means

Situation: Population variances are known

Example:

A research team is interested in the difference between serum uric acid levels in patients with and without Down's syndrome. In a large hospital for the treatment of the mentally retarded, a sample of 12 individuals with Down's syndrome yielded a mean of $\bar{x}_1 = 4.5$ mg/100 ml. In a general hospital a sample of 15 normal individuals of the same age and sex were found to have a mean value of $\bar{x}_2 = 3.4$ mg/100 ml. If it is reasonable to assume that the two populations of values are normally distributed with variances equal to 1 and 1.5, find the 95 percent confidence interval for $\mu_1 - \mu_2$.

(1) Given

$$n_1 = 12,\ \bar{x}_1 = 4.5,\ \sigma_1^2 = 1$$
$$n_2 = 15,\ \bar{x}_2 = 3.4,\ \sigma_2^2 = 1.5$$

(2) Calculations

- The point estimate for $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$

  $\bar{x}_1 - \bar{x}_2 = 4.5 - 3.4 = 1.1$

- The standard error is

  $$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{1}{12} + \frac{1.5}{15}} = .4282$$

- The 95% confidence interval is

  $$(\bar{x}_1 - \bar{x}_2) \pm z_{(1-\alpha/2)} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

  $$1.1 \pm 1.96\ (.4282)$$

  $$(.26,\ 1.94)$$

Discussion: As this is a z-interval, we know that the correct value of z to use is 1.96. We interpret this interval that the difference between the two population means is 1.1 and we are 95% confident that the true mean lies between 0.26 and 1.94.

Situation: Population variances are unknown but can be assumed to be equal

Example:

Given that,

$$n_1 = 13, \quad \bar{x}_1 = 21.0, \quad s1 = 4.9$$

$$n_2 = 17, \quad \bar{x}_2 = 12.1, \quad s2 = 5.6$$

Calculations

- The point estimate for $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$

$$\bar{x}_1 - \bar{x}_2 = 21.0 - 12.1 = 8.9$$

- The pooled estimate of the variance is

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

$$s_p^2 = \frac{(13 - 1)(4.9)^2 + (17 - 1)(5.6)^2}{13 + 17 - 2}$$

$$s_p^2 = \frac{288.12 + 501.76}{28}$$

$$s_p^2 = 28.21$$

- The standard error is

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$= \sqrt{\frac{28.21}{13} + \frac{28.21}{17}} = 1.9569$$

- The 95% confidence interval is

$$(x_1 - x_2) \pm t_{(1-\alpha/2)} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$8.9 \pm 2.0484\ (1.9569)$$

$$8.9 \pm 4.0085$$

$$(4.9,\ 12.9)$$

Discussion: The correct value of t to use for a 95% confidence interval with 28 degrees of freedom is 2.0484. We interpret this interval that the difference between the two-population means is estimated to be 8.9 and we are 95% confident that the true value lies between 4.9 and 12.9.

Two Variance

Example:

Let's return to the example, in which the feeding habits of two-species of net-casting spiders are studied. The species, the *deinopis* and *menneus*, coexist in eastern Australia. The following summary statistics were obtained on the size, in millimeters, of the prey of the two species:

Adult DINOPIS	Adult MENNEUS
$n = 10$	$m = 10$
$\overline{x} = 10.26$ mm	$\overline{y} = 9.02$ mm
$s_X^2 = (2.51)^2$	$s_Y^2 = (1.90)^2$

Estimate, with 95% confidence, the ratio of the two population variances.

Solution. In order to estimate the ratio of the two population variances, we need to obtain two F-values from the F-table, namely:

$$F_{0.025}(9,9) = 4.03 \text{ and } F_{0.975}(9,9) = \frac{1}{F_{0.025}(9,9)} = \frac{1}{4.03}$$

Then, the 95% confidence interval for the ratio of the two population variances is:

$$\frac{1}{4.03}\left(\frac{2.51^2}{1.90^2}\right) \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq 4.03\left(\frac{2.51^2}{1.90^2}\right)$$

Simplifying, we get:

$$0.433 \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq 7.033$$

That is, we can be 95% confident that the ratio of the two population variances is between 0.433 and 7.033. (Because the interval contains the value 1, we cannot conclude that the population variances differ.)

Small Sample Example:

We previously considered a subsample of n=10 participants attending the 7th examination of the Offspring cohort in the Framingham Heart Study. The following table contains descriptive statistics on the same continuous characteristics in the subsample stratified by sex.

	Men			Women		
Characteristic	**n**	**Sample Mean**	**s**	**n**	**Sample Mean**	**s**
Systolic Blood Pressure	6	117.5	9.7	4	126.8	12.0
Diastolic Blood Pressure	6	72.5	7.1	4	69.5	8.1
Total Serum Cholesterol	6	193.8	30.2	4	215.0	48.8
Weight	6	196.9	26.9	4	146.0	7.2
Height	6	70.2	1.0	4	62.6	2.3
Body Mass Index	6	28.0	3.6	4	26.2	2.0

Suppose we wish to construct a 95% confidence interval for the difference in mean systolic blood pressures between men and women using these data. We will again arbitrarily designate men group 1 and women group 2. Since the sample sizes are small (i.e., $n_1 < 30$ and $n_2 < 30$), the confidence interval formula with t is appropriate. However, we will first check whether the assumption of equality of population variances is reasonable. The ratio of the sample variances is $9.7^2/12.0^2 = 0.65$, which falls between 0.5 and 2, suggesting that the assumption of equality of population variances is reasonable. The solution is shown below.

First, we compute Sp, the pooled estimate of the common standard deviation:

$$S_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

Substituting:

$$S_p = \sqrt{\frac{(6-1)9.7^2+(4-1)12.0^2}{6+4-2}} = \sqrt{112.81} = 10.6$$

Note that again the pooled estimate of the common standard deviation, Sp, falls in between the standard deviations in the comparison groups (i.e., 9.7 and 12.0). The degrees of freedom (df) = $n_1+n_2-2 = 6+4-2 = 8$. From the t-Table t=2.306. The 95% confidence interval for the difference in mean systolic blood pressures is:

$$(\bar{x}_1 - \bar{x}_2) \pm t\, S_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Substituting:

$$(117.5-126.8)\pm 2.306\,(10.6)\sqrt{\tfrac{1}{6}+\tfrac{1}{4}}=-9.3\pm 2.306(6.84)$$

Then simplifying further:

$$-9,3\pm 15.77$$

So, the 95% confidence interval for the difference is (-25.07, 6.47)

Interpretation: Our best estimate of the difference, the point estimate, is -9.3 units. The standard error of the difference is 6.84 units and the margin of error is 15.77 units. We are 95% confident that the difference in mean systolic blood pressures between men and women is between -25.07 and 6.47 units. In this sample, the men have lower mean systolic blood pressures than women by 9.3 units. Based on this interval, we also conclude that there is no statistically significant difference in mean systolic blood pressures between men and women, because the 95% confidence interval includes the null value, zero.

Example:

In a Study, participants attend clinical examinations approximately every four years. Suppose we want to compare systolic blood pressures between examinations (i.e., changes over 4 years). The data below are systolic blood pressures measured at the sixth and seventh examinations in a subsample of n=15 randomly selected participants.

Subject #	Examination 6	Examination 7
1	168	141
2	111	119
3	139	122
4	127	127
5	155	125
6	115	123
7	125	113
8	123	106
9	130	131

10	137	142
11	130	131
12	129	135
13	112	119
14	141	130
15	122	121

Therefore,

$$X_d = \frac{\Sigma X}{n} = \frac{-79.0}{15} = -5.3$$

and

$$s_d = \sqrt{\frac{\Sigma(\text{Differences} - \bar{X}_d)^2}{n-1}} = \sqrt{\frac{2296.95}{14}} = \sqrt{164.07} = -12.8$$

We can now use these descriptive statistics to compute a 95% confidence interval for the mean difference in systolic blood pressures in the population. Because the sample size is small (n=15), we use the formula that employs the t-statistic. The degrees of freedom are df=n-1=14. From the table of t-scores (see Other Resource on the right), t = 2.145. We can now substitute the descriptive statistics on the difference scores and the t value for 95% confidence as follows:

$$\bar{X}_d = \pm t \frac{s_d}{\sqrt{n}}$$

$$-5.3 \pm 2.145 \frac{12.8}{\sqrt{15}} = -5.3 \pm 2.145\,(3.3) = -5.3 \pm 7.1$$

So, the 95% confidence interval for the difference is (-12.4, 1.8).

Interpretation:

We are 95% confident that the mean difference in systolic blood pressures between examinations 6 and 7 (approximately 4 years apart) is between -12.4 and 1.8. The null (or no effect) value of the CI for the mean difference is zero. Therefore, based on the 95% confidence interval we can conclude that there is no statistically significant difference in blood pressures over time, because the confidence interval for the mean difference includes zero.

Example:

The following table contains data on prevalent cardiovascular disease (CVD) among participants who were currently non-smokers and those who were current smokers at the time of the fifth examination in the Framingham Offspring Study.

	Free of CVD	History of CVD	Total
Non-Smoker	2,757	298	3,055
Current Smoker	663	81	744
Total	3,420	379	3,799

The point estimate of prevalent CVD among non-smokers is $298/3{,}055 = 0.0975$, and the point estimate of prevalent CVD among current smokers is $81/744 = 0.1089$. When constructing confidence intervals for the risk difference, the convention is to call the exposed or treated group 1 and the unexposed or untreated group 2. Here smoking status defines the comparison groups, and we will call the current smokers group 1 and the non-smokers group 2. A confidence interval for the difference in prevalent CVD (or prevalence difference) between smokers and non-smokers is given below. In this example, we have far more than 5 successes (cases of prevalent CVD) and failures (persons free of CVD) in each comparison group, so the following formula can be used:

$$\hat{p}_1 - \hat{p}_2 \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Substituting we get:

$$(0.1089 - 0.0975) \pm 1.96\sqrt{\frac{0.1089(1-0.1089)}{744} + \frac{0.0975(1-0.0975)}{3055}}$$

This simplifies to

$$0.0114 \pm 1.96\,(0.0126) = 0.0114 \pm 0.0247$$

So the 95% confidence interval is (-0.0133, 0.0361),

Interpretation: We are 95% confident that the difference in proportion the proportion of prevalent CVD in smokers as compared to non-smokers is between -0.0133 and 0.0361. The null value for the risk difference is zero. Because the 95% confidence interval includes zero, we conclude that the difference in prevalent CVD between smokers and non-smokers is not statistically significant.