```
5  CI =
6
7      10.0798
8      13.3523
```

The argument `'alpha'` in the function `ttest` corresponds to α giving the $1-\alpha$ coverage probability for the confidence interval.

Difference of two means

Suppose we have two independent simple random samples from two populations. We have m samples $(X_1, \ldots, X_m)$ from the first population which has a $\text{Normal}(\mu_X, \sigma^2)$ distribution and we have n samples $(Y_1, \ldots, Y_n)$ from the second population which has a $\text{Normal}(\mu_Y, \sigma^2)$ distribution. We would like to estimate the difference in the means and be able to quantify our uncertainty about this difference.

To construct a confidence interval for $\mu_X - \mu_Y$, we need to know the distribution of $\bar{X} - \bar{Y}$. We know that $\bar{X}$ and $\bar{Y}$ are independent with $\bar{X} \sim \text{Normal}(\mu_X, \sigma^2/m)$ and $\bar{Y} \sim \text{Normal}(\mu_Y, \sigma^2/n)$. Therefore

$$\mathbb{E}(\bar{X}-\bar{Y}) = \mathbb{E}(\bar{X}) - \mathbb{E}(\bar{Y}) = \mu_X - \mu_Y \qquad \text{Var}(\bar{X}-\bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y})$$

$$\bar{X} - \bar{Y} \sim \text{Normal}\left(\mu_X - \mu_Y, \sigma^2\left(\tfrac{1}{m} + \tfrac{1}{n}\right)\right)$$

so

$$\frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim \text{Normal}(0,1).$$

(variance ↑, mean ↑)

As before, it is unrealistic to assume that we know σ^2 and so it will need to be estimated. If S_X^2 and S_Y^2 are the sample variance estimators applied to the first and second sample, respectively, then we can form the *pooled variance* estimator by

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}.$$

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2$$

$$(m-1)S_X^2 = \sum_{i=1}^{m}(X_i - \bar{X})^2$$

Exercise: Show S_p^2 is an unbiased estimator of σ^2.

(unbiased estimators of σ^2)

$$\mathbb{E}[S_p^2] = \frac{(m-1)}{m+n-2}\mathbb{E}[S_X^2] + \frac{(n-1)}{m+n-2}\mathbb{E}[S_Y^2] = \frac{(m-1)\sigma^2}{(m+n-2)} + \frac{(n-1)\sigma^2}{(m+n-2)} = \sigma^2.$$

Replacing σ^2 by the pooled variance estimator S_p^2, leads to a statistic that has a t_{m+n-2}-distribution

$$\frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t_{m+n-2},$$

$$1-\alpha = \mathbb{P}\left(t_{\alpha/2;\,m+n-2} \leq \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \leq t_{1-\alpha/2;\,m+n-2}\right)$$

Rearranging as usual, we obtain

$$1-\alpha = \mathbb{P}\left(\bar{X}-\bar{Y}-t_{1-\alpha/2;\,m+n-2}\,S_p\sqrt{\tfrac{1}{m}+\tfrac{1}{n}} \leq \mu_X-\mu_Y \leq \bar{X}-\bar{Y}-t_{\alpha/2;\,m+n-2}\,S_p\sqrt{\tfrac{1}{m}+\tfrac{1}{n}}\right)$$

or more compactly (and using the symmetry of quantiles of the t-distribution)

$$(\bar{X}-\bar{Y}) \pm t_{1-\alpha/2;\,m+n-2}\,S_p\sqrt{\frac{1}{m}+\frac{1}{n}}$$

as a $1-\alpha$ stochastic confidence interval for the difference in means. The numerical confidence interval is

$$(\bar{x}-\bar{y}) \pm t_{1-\alpha/2;\,m+n-2}\,s_p\sqrt{\frac{1}{m}+\frac{1}{n}},$$

where

$$s_p^2 = \frac{(m-1)s_X^2+(n-1)s_Y^2}{m+n-2}.$$

Remark: If our two ~~samples~~ populations have different variances, then it is still possible to construct an approximate confidence interval for the difference of the two means. The statistic

$$\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}}}$$

has approximately a t-distribution with the degrees of freedom determine by the *Welch* approximation.

Exercise: Suppose that we have two hard-drives of the same model, and we wish to determine an estimate for the average difference in time it takes to write a 2 Gb file to each of the two hard-drives we are testing, as well as to quantify the uncertainty inherent in the estimate. We will assume that write times are Normally distributed, with unknown means μ_X and μ_Y and common variance. We have the following data:

$$(x_1,x_2,x_3,x_4,x_5) = (7.2\ s, 8.3\ s, 7.8\ s, 8.1\ s, 7.5\ s)$$
$$(y_1,y_2,y_3,y_4,y_5) = (7.6\ s, 7.3\ s, 8.1\ s, 7.1\ s, 7.0\ s)\,.$$

Construct a 95% confidence interval for the unknown difference in means, $\mu_X-\mu_Y$.

A company decided to upgrade the computer software in its production process. It must decide between two different software packages (called A and B). The supplier of Package B claims that the production time using their software is less than the production time using Package A. In order to test this claim, the information systems manager performed a trial where employees using both packages were randomly selected. Employees were timed to see how long they needed with Package A and Package B during the production process. The average and standard deviation time (in seconds) taken by employees in using the software packages was calculated and is presented below.

	Package A	Package B
Number using package	15	20
Average time taken by employees (seconds)	26	23.5
Sample standard deviation (seconds)	2.4	2.6

Calculate a 95% confidence interval, for the difference in mean production times using Package A and Package B from the sample data in the table above.

We want a 95% CI for $\mu_A - \mu_B$ where

μ_A - mean time taken by employees with package A

μ_B - " " B

$\bar{x}_A = 26$, $s_A = 2.4$, $n_A = 15$

$\bar{x}_B = 23.5$, $s_B = 2.6$, $n_B = 20$

(pooled variance) $s_p^2 = \dfrac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$

$= \dfrac{14 \times 2.4^2 + 19 \times 2.6^2}{33} = 6.3358$

$s_p = \sqrt{6.3358} = 2.5171$

$t_{0.975;\,33} = 2.0345$ s.e. $(\bar{x}_A - \bar{x}_B) = s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 2.5171 \times \sqrt{\frac{1}{15} + \frac{1}{20}}$

$= 0.8598$

$\bar{x}_A - \bar{x}_B = 26 - 23.5 = 2.5$

$2.5 \pm 2.0345 \times 0.8598 \Rightarrow 2.5 \pm 1.7657$ (sec) is the 95% CI for $\mu_A - \mu_B$.

$\bar{x} = \frac{1}{5}(7.2 + 8.3 + 7.8 + 8.1 + 7.5) = 7.78$

$s_x^2 = \frac{1}{(5-1)} \cdot \left((7.2-7.78)^2 + (8.3-7.78)^2 + (7.8-7.78)^2 + (8.1-7.78)^2 + (7.5-7.78)^2\right)$

$= 0.1970$

$\bar{x} = 7.78 \qquad s_x^2 = 0.1970 \qquad m = 5$

(bit weird)

$\bar{y} = 7.42 \qquad s_y^2 = 0.1970 \qquad n = 5$

$$s_p^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2} = \frac{4 \times (0.1970) + 4 \times (0.1970)}{8} = (0.1970).$$

The critical value we need from the t-distribution is

$0.95 = 1 - \alpha$

$\Rightarrow \alpha = 0.05$

$t_{1-\alpha/2;\, m+n-2} = t_{0.975;\, 8} = 2.306$

The (numerical) 95% confidence interval is

$$\bar{x} - \bar{y} \pm t_{0.975;8} \times s_p \sqrt{\frac{1}{5} + \frac{1}{5}}$$

$$(7.78 - 7.42) \pm 2.306 \times \sqrt{0.1970} \times \sqrt{2/5}$$

95% CI 0.36 ± 0.6473 (sec)

This proceedure is implemented in MATLAB with the function `ttest2`.

R: ttest

```
1  x = [7.2    8.3    7.8    8.1    7.5];
2  y = [7.6    7.3    8.1    7.1    7.0];
3  [H,P,CI,STATS] = ttest2(x,y);
4  CI
5
6  CI =
7
8     -0.2873    1.0073
9
10 STATS
11
12 STATS =
13
14   struct with fields:
15
16      tstat: 1.2824
17         df: 8
18         sd: 0.4438
```

Confidence intervals for proportions

One of the variables recorded in Desharnais's survey is the number of years experience of the team undertaking the project. Of the 79 projects, 20 were completed by teams with only one year experience. Assuming this survey is representative of the population of information systems development projects, we could construct a confidence interval for the proportion of projects completed by teams with one year experience.

Suppose $X_1, X_2, \ldots, X_n$ is a simple random sample with $X_i \sim \text{Bernoulli}(p)$. The natural estimator for p is

$$\widehat{P} = \frac{1}{n} \sum_{i=1}^{n} X_i .$$

$\mathbb{E}[\hat{P}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n} p = p$

$\mathrm{Var}(\hat{P}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}$

$= \frac{1}{n^2} \times n\,p(1-p)$

By the central limit theorem,

$$\hat{P} \sim_{approx} \text{Normal}\left(p\ ,\ \frac{p(1-p)}{n}\right),$$

Therefore, we have

We need $np \geq 10$ and $n(1-p) \geq 10$ for this approximation

$$\mathbb{P}\left(z_{\alpha/2} \leqslant \frac{\sqrt{n}(\hat{P}-p)}{\sqrt{p(1-p)}} \leqslant z_{1-\alpha/2}\right) \approx 1-\alpha.$$

As $\hat{P}$ is consistent estimator of p, we may replace p in the denominator to obtain

$$\mathbb{P}\left(z_{\alpha/2} \leqslant \frac{\sqrt{n}(\hat{P}-p)}{\sqrt{\hat{P}(1-\hat{P})}} \leqslant z_{1-\alpha/2}\right) \approx 1-\alpha.$$

Rearranging, and using the symmetry of standard normal quantiles, we have

$$\mathbb{P}\left(\hat{P} - z_{1-\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leqslant p \leqslant \hat{P} + z_{1-\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) \approx 1-\alpha,$$

which gives as an approximate $1-\alpha$ stochastic confidence interval for p:

$$\hat{P} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

Returning to our example at the start of this section, suppose we are constructing a 90% confidence interval for proportion of projects completed by teams with one year's experience. Each project will either be completed by a team with one year's experience or by a team with a different amount of experience. Think of the random variables

$$X_i = \begin{cases} 1, & \text{project } i \text{ completed by a team with one year's experience} \\ 0, & \text{else.} \end{cases}$$

Our estimate of this proportion is $\hat{p} = 20/79 \approx 0.2532$. So the numerical 90% confidence interval for the proportion is

$s.e.(\hat{p}) = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} = \sqrt{\frac{0.253 \times 0.747}{79}} \simeq 0.0489$

$0.90 = 1-\alpha \Rightarrow \alpha = 0.1 \qquad z_{1-\alpha/2} = z_{0.95} = 1.645$

$0.2532 \pm 1.645 \times 0.0489 \Rightarrow 0.2532 \pm 0.0804$ 90% CI of proportion.

Confidence intervals constructed in this way are only approximate in the sense that the coverage probability for these intervals is appoximately $1-\alpha$. How close the coverage probability is to the desired level depends on how well close the distribution of $\hat{P}$ is to the normal distribution. A general rule of thumb is that both np and $n(1-p)$ should be at least 10. The MATLAB function `binofit` will produce a confidence interval for a proportion using a different approach to the one described above.