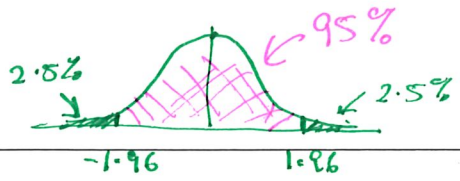


As the standard normal distribution is symmetric about 0, the quantiles satisfy $-z_{\alpha/2} = z_{1-\alpha/2}$.

Hence a stochastic $1-\alpha$ confidence interval for μ in this case is

$$\left(\bar{X} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\ , \bar{X} + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\ \right),$$

$z_{1-\alpha/2}$: $1-\alpha/2$ quantile of $N(0,1)$

which is often *abbreviated* to

$$\bar{X} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}$$

So for example, in 95% of the simple random samples from Normal(μ, σ^2), μ will be within $1.96 \times \sigma/\sqrt{n}$ of $\bar{X}$.

Exercise: Suppose that we wish to determine the average time it takes to write a 2 Gb file to a hard-drive 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 mean μ but known standard deviation $\sigma = 1$ s. We have the following data:

$$7.2\ s, 8.3\ s, 7.8\ s, 8.1\ s, 7.5\ s\,.$$

Construct a numerical (numerical) 95% confidence interval for the unknown mean.

We calculate

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

From the tabulated values of the standard normal cdf,

$z_{0.975} = 1.96$, so $z_{0.975} \times \sqrt{\frac{\sigma^2}{n}} = 1.96 \times \sqrt{\frac{1}{5}} \simeq 0.88\,s$.

MATLAB norminv R qnorm

The (numerical) 95% confidence interval is

$$7.78 \pm 0.88\,s \quad \text{or} \quad (6.90\,s\ ,\ 8.66\,s)$$

This is great. We were able to say something about an unknown parameter μ based on our sample. Unfortunately, this is practically useless since there is no reason why we would know what σ^2 is.

Impact of Unknown Variance

For a random sample from a normal distribution with known variance σ^2, we have seen that the estimator corresponding to the *sample mean* $\bar{X}$ is normally distributed. From this we can construct a confidence interval for the unknown mean μ. How can we proceed when σ^2 is unknown?

It is natural to consider replacing σ^2 by the unbiased estimator of σ^2 given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\,.$$

$\mathbb{E}[S^2] = \sigma^2$ unbiased.

We will denote the estimate of σ^2 obtained in this way by s^2, that is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

$\bar{X}$ estimator of μ
$\bar{x}$ estimate of μ
S^2 estimator of σ^2
s^2 estimate of σ^2.

For a simple random sample of size n from a Normal(μ, σ^2) distribution, we now have

Students t distribution

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathrm{t}_{n-1},$$

where t_{n-1} denotes the t-distribution with $n-1$ *degrees of freedom*. For each sample size we have a different t-distribution. For small degrees of freedom (that is small sample sizes), the t-distribution has much fatter tails than the standard normal distribution. However, as we increase the degrees of freedom (that is as the sample size increases) the t-distribution converges to the standard normal distribution.

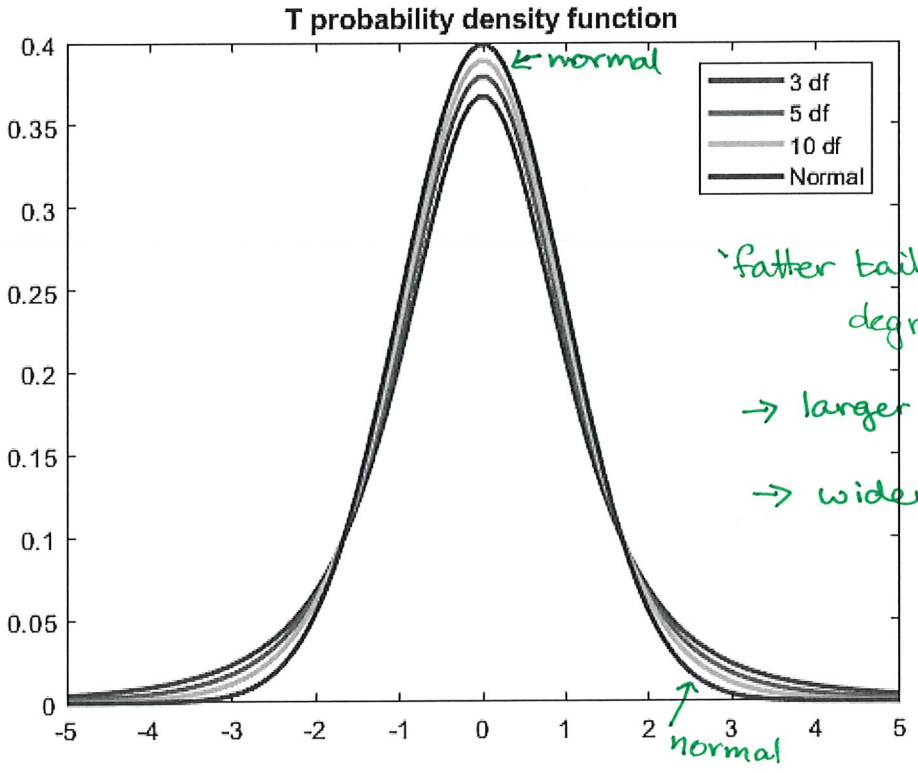


Figure 6.1: The probability density function of the t-distribution with 3, 5 and 10 degrees of freedom together with the standard normal probability density function.

As with the normal distribution, the cumulative distribution function of the t-distribution does not have a simple form and so we will need to refer to tables. There is an added complication since there is a different t-distribution for each sample size. Instead of making a book of tables for the t-distribution only the important (**critical**) values of the t-distribution are tabulated.

With this in mind, lets now construct our confidence interval for μ in the realistic setting

where σ^2 is unknown. We know that

$$\mathbb{P}\left(t_{\alpha/2;\,n-1} \leqslant \frac{\bar{X}-\mu}{S/\sqrt{n}} \leqslant t_{1-\alpha/2;\,n-1}\right) = 1-\alpha\,,$$

where $t_{\gamma;\,n-1}$ is the γ-quantile of the t_{n-1} distribution. Rearranging, we have

$$\mathbb{P}\left(\bar{X} - t_{1-\alpha/2;\,n-1}\frac{S}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + t_{1-\alpha/2;\,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha\,.$$

Like the standard normal distribution, the t-distribution is symmetric about 0 so the quantiles satisfy $-t_{\alpha/2;\,n-1} = t_{1-\alpha/2;\,n-1}$. Hence a stochastic $1-\alpha$ confidence interval for μ when σ^2 is unknown is

$$\left(\bar{X} - t_{1-\alpha/2;n-1}\frac{S}{\sqrt{n}}\,,\, \bar{X} + t_{1-\alpha/2;n-1}\frac{S}{\sqrt{n}}\right).$$

The numerical $(1-\alpha)$ confidence interval, which is what we actually calculate from sample data, is

$$\bar{x} \pm t_{1-\alpha/2;\,n-1}\frac{s}{\sqrt{n}}\,.$$

Some terminology:

- The quantity $s/\sqrt{n}$ is an estimate of the standard deviation of $\bar{X}$. It is called the *standard error* of sample mean and is sometimes denoted by $\mathrm{se}(\bar{x})$.
- The quantity $t_{1-\alpha/2;\,n-1}s/\sqrt{n}$ is the *margin of error* of our estimate $\bar{x}$.

In this section we have worked under the assumption that our simple random sample was from a $\mathsf{Normal}(\mu, \sigma^2)$ distribution. Distributions that arise in practice are rarely exactly normal and so it is important to understand how well our inferential methods perform under deviations from the normal distribution. Due to the central limit theorem, this depends in part on the size of the sample. The general recommendations are as follows:

- For small sample sizes ($n < 15$) we can use methods based on the t-distribution if the data are close to symmetric and there are no *outliers*.
- For moderate sample sizes ($15 \leqslant n < 40$) we can use methods based on the t-distribution as long as there are no outliers or strong skewness in the data.
- For large sample sizes ($40 \leqslant n$) we can use methods based on the t-distribution even in the presence of skewness, though outliers may still affect results.

Let's now consider an example of the material covered in this chapter so far.

Example: Jean-Marc Desharnais[1] surveyed 10 organisations on 81 management information systems development projects completed between 1983 and 1988. Of those projects in the survey, the sample mean for time to completion was 11.7160 months and the sample standard deviation was 7.3997 months. We would like to construct a 95% confidence interval for the mean length of time to completion for a project completed. A histogram of the data is given in Figure 6.2.

[1] The file `deshardnais.xlsx` is on Blackboard. The original data from Desharnais's Masters thesis is available from `http://promise.site.uottawa.ca/SERepository/datasets-page.html`.

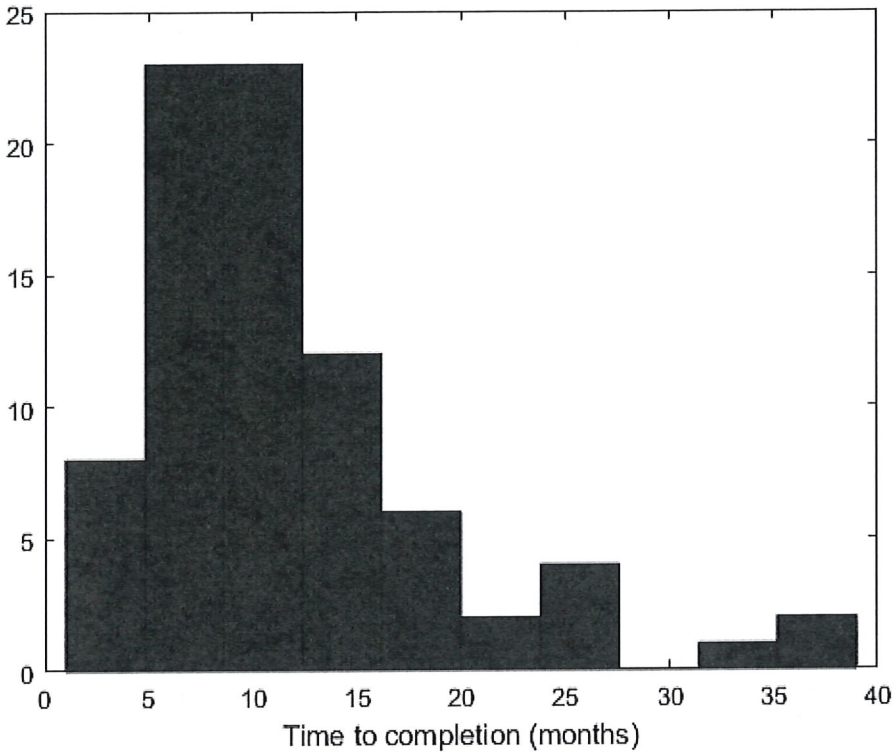


Figure 6.2: Time in months for project completion of 81 management information system development projects.

We are told

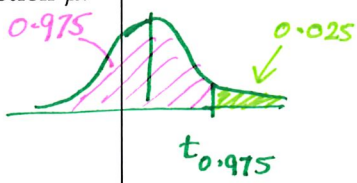
$$\bar{x} = 11.7160 \qquad s = 7.3997 \qquad n = 81$$

We want to construct a 95% confidence interval for the mean time to completion μ.

The critical value from the t-distribution in this instance is

$0.95 = 1-\alpha$, $\alpha = 0.05$

$$t_{1-\alpha/2;n-1} = t_{0.975;80} = 1.990$$



The (numerical) 95% confidence interval for μ is

$$\bar{x} \pm t_{0.975;80} \times \frac{s}{\sqrt{81}} = 11.7160 \pm 1.990 \times \frac{7.3997}{\sqrt{81}} = 11.7160 \pm 1.636 \text{ (months)}$$

margin of error

The following code can be used to construct a confidence interval for the mean in MATLAB. First place the file 'desharnais.xlsx' in your current working directory, and then execute the following code.

```
1 desh = readtable('desharnais.xlsx');
2 [H,P,CI]=ttest(desh.Length,0,'alpha',0.05);
3 CI
4
```

```
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 $\mathsf{Normal}(\mu_X, \sigma^2)$ distribution and we have n samples $(Y_1, \ldots, Y_n)$ from the second population which has a $\mathsf{Normal}(\mu_Y, \sigma^2)$ 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 $\bar{X} \sim \mathsf{Normal}(\mu_X, \sigma^2/m)$ and $\bar{Y} \sim \mathsf{Normal}(\mu_Y, \sigma^2/n)$. Therefore

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

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

so

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

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

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

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

$$\frac{(\bar{X} - \bar{Y}) - }{} \sim t_{m+n-2}\,,$$