Statistics for Computing MA4413

Lecture 14

Smaller Samples: The T Distribution

Kevin Burke

kevin.burke@ul.ie

The Central Limit Theorem

Let $X_1, X_2, \ldots, X_n \sim$ *any* distribution with:

- $\mu = E(X)$
- $\sigma = Sd(X)$

The power of the *central limit theorem* is that $\bar{X}$ is approximately normally distributed regardless of the distribution of the individual values, $X_1, X_2, \ldots, X_n$, i.e.,

$$\bar{X} \sim \text{Normal}\left(\mu,\ \sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}\right).$$

This works for large samples ($n > 30$) and, furthermore, in these large samples, we can also replace σ with s.

$$\sigma(\bar{X}) \approx s(\bar{X}) = \frac{s}{\sqrt{n}}.$$

Large Samples

Replacing σ with s is okay in large samples since s will be close to the true value σ in this case:

$$\Rightarrow \frac{\bar{X} - \mu}{S/\sqrt{n}} \approx \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = Z \sim \text{Normal}(0, 1).$$

Based on the above, we develop a confidence interval for μ via

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

which we know has probability $1 - \alpha$ of containing μ and probability α (the error probability) of not containing μ.

Small Samples

For **smaller samples** ($n \leq 30$), s varies more from sample to sample; we must account for this extra level of uncertainty.

In particular, it turns out that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \textbf{t distribution},$$

i.e., *not* a Normal$(0, 1)$ distribution.

The t distribution is symmetric like the Normal distribution but has longer tails (leading to wider confidence intervals) which reflects the extra uncertainty.

Use of the T Distribution

Unlike the central limit theorem, the result on the previous slide **does not hold** for $\overline{X}$ calculated from *any* sample, $X_1, X_2, \ldots, X_n$.

It relies on the **assumption** that the individual data values are **normally distributed**, i.e., $X_1, X_2, \ldots, X_n \sim \text{Normal}(\mu, \sigma)$.

In practice we must *check* that our small sample of data looks approximately normal (using a histogram and a Q-Q plot).

As long as the data looks reasonably normally distributed, we can apply the theory of the t distribution.

(note: for highly non-normal data there are so-called "non-parametric" methods)

Note on Normality

In order for the data to be normally distributed, it must be numeric.

- Technically continuous, but in practice we are not so strict.
- Often discrete data can still look reasonably normal.

On the other hand, **categorical data cannot be normally distributed**.

- $\{\text{Yes}, \text{No}\} = \{1, 0\} \Rightarrow$ frequencies of 1s and 0s could never look like a symmetric bell-shape.
- We *always* require large samples ($n > 30$) for categorical data to produce confidence intervals for p and $p_1 - p_2$.

  (in fact, it turns out that samples *much* larger than 30 may be required if the true proportion is near 0 or 1)

One Mean

$$\text{If} \quad X_1, X_2, \ldots, X_n \sim \text{Normal}(\mu, \sigma) \quad \text{and} \quad n \leq 30$$

$$\Rightarrow \quad \boxed{T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \text{T}(\nu)},$$

where $\text{T}(\nu)$ denotes a t distribution with ν *"degrees of freedom"*.
(note that ν is the Greek letter "nu")

The value of ν is the sample size minus one, i.e.,

$$\boxed{\nu = n - 1}.$$

One Mean: Confidence Interval

We calculate the confidence interval in the same way as before, except that a t value is used:

$$\boxed{\bar{x} \pm t_{\nu,\,\alpha/2}\,\frac{s}{\sqrt{n}}}$$

where, as before, $\alpha/2$ corresponds to a $(1-\alpha)\%$ confidence interval and $\nu = n - 1$ are the degrees of freedom for the t distribution.

The $t_{\nu,\,\alpha/2}$ values are always larger than the $z_{\alpha/2}$ values used previously. Hence, confidence intervals are wider and account for the extra uncertainty caused by using s in place of σ.

Using T Tables

Just as we must look up z values in the normal tables, we find t values in the **t tables**.

The t tables differ from the normal tables in that probabilities appear in the column headings and t values appear in the body of the table.

We look up the appropriate t value via:

- **Row** $\Rightarrow$ degrees of freedom ($\nu = n - 1$ for one mean).
- **Column** $\Rightarrow$ probability ($\alpha/2$ for confidence intervals).

Examples: $t_{3,\,0.025} = 3.182$, $t_{10,\,0.025} = 2.228$, $t_{14,\,0.005} = 2.977$ etc.

Note on T Tables

Recall that:

- $z_{0.1} = 1.64$
- $z_{0.025} = 1.96$
- $z_{0.005} = 2.58$

Note bottom row of t tables:

- $t_{\infty, 0.1} = 1.645$
- $t_{\infty, 0.025} = 1.96$
- $t_{\infty, 0.005} = 2.576$

This highlights the fact that when n is large, we proceed as before using z values (and don't require the individual data values to be normally distributed in this case).

Example: Life Time of Mechanical Components

Let's assume that a sample of 5 mechanical components were used until they failed. It was found that the average lifetime in the sample was 2.18 years and the standard deviation was 0.67 years.

Here we have $n = 5$, $\bar{x} = 2.18$ and $s = 0.67$.

We wish to produce a 95% confidence interval. As before there is 5% remaining $\Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$.

Since the sample is small, we use the t tables (assuming the data is reasonably normally distributed).

The t distribution we require has $\nu = n - 1 = 5 - 1 = 4$.

Example: Life Time of Mechanical Components

The 95% confidence interval is then

$$\bar{x} \pm t_{4,\,0.025}\,\frac{s}{\sqrt{n}}$$

$$2.18 \pm 2.776\left(\frac{0.67}{\sqrt{5}}\right)$$

$$2.18 \pm 2.776\,(0.2996)$$

$$2.18 \pm 0.8317$$

$$[1.35,\ 3.01]$$

We are 95% confident that the true mean lies in the above interval.

Question 1

A manufacturer of CPUs wishes to investigate the temperature of a type of CPU under certain conditions. A sample of 6 CPUs were randomly selected and left to run an intensive task for one hour. The temperature of each was then measured and the results are as follows:

38.3	38.9	39.2	39.2	39.6	41.0

a) Calculate $\bar{x}$ and s.

b) Calculate a 99% confidence interval for μ.

Difference Between Two Means

Previously we saw that for, $n_1 > 30$ and $n_2 > 30$, a confidence interval for the difference between two means, $\mu_1 - \mu_2$, is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

As with the confidence interval for one mean, we must replace the z value with a t value if one or both samples are small.

(note: both samples must be reasonably normally distributed)

There are two commonly used approaches:

- Unequal variances: no assumption about variances.
- Equal variances: assume $\sigma_1^2 = \sigma_2^2$ (must be checked).

Unequal Variances

For the **unequal variance** approach use the formula

$$\boxed{(\bar{x}_1 - \bar{x}_2) \pm t_{\nu,\,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

with degrees of freedom given by

$$\boxed{\nu = \frac{(a+b)^2}{\frac{a^2}{n_1-1} + \frac{b^2}{n_2-1}}},$$

where $a = \frac{s_1^2}{n_1}$ and $b = \frac{s_2^2}{n_2}$.

Equal Variances Assumed

An alternative (classical) approach is to assume that the true variances are equal, i.e., $\sigma_1^2 = \sigma_2^2$.

For the **equal variance** approach use the formula

$$\boxed{(\bar{x}_1 - \bar{x}_2) \pm t_{\nu,\,\alpha/2} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}},$$

where the **pooled variance** is

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

and the degrees of freedom are

$$\boxed{\nu = n_1 + n_2 - 2}.$$

Unequal Vs Equal Variances

- Unequal variance approach:
  - Referred to as *Welch's t test*.
  - The default method in R.
  - This method is preferable since it does not make the extra assumption of equal variances.
- Equal variance approach:
  - A more classical method.
  - Typically found in textbooks.
  - Equal variance assumption must be checked *first* using the **F test**.

Note: both methods assume that the two samples are approximately normally distributed.

Example: Salary

The salaries (in thousands) of graduates from two universities are as follows:

University 1	32.1	32.4	33.2	33.3	33.6
University 2	35.7	36.3	39.4	40.5	

Here $n_1 = 5$ and $n_2 = 4$. Hence, we need to apply the small sample theory.

Whether we assume equal variances or not, we first need to calculate $\bar{x}_1$, s_1, $\bar{x}_2$ and s_2.

Example: Salary (University 1)

						$\sum$
x_1	32.1	32.4	33.2	33.3	33.6	164.6
x_1^2	1030.41	1049.76	1102.24	1108.89	1128.96	5420.26

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{164.6}{5} = 32.92.$$

$$s_1^2 = \frac{\sum x_1^2 - n_1 \bar{x}_1^2}{n_1 - 1} = \frac{5420.26 - 5(32.92^2)}{4} = 0.407.$$

$$s_1 = \sqrt{0.407} = 0.638.$$

Example: Salary (University 2)

					$\sum$
x_2	35.7	36.3	39.4	40.5	151.9
x_2^2	1274.49	1317.69	1552.36	1640.25	5784.79

$$\bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{151.9}{4} = 37.975.$$

$$s_2^2 = \frac{\sum x_2^2 - n_2 \bar{x}_2^2}{n_2 - 1} = \frac{5784.79 - 4(37.975^2)}{3} = 5.4625.$$

$$s_2 = \sqrt{5.4625} = 2.337.$$

Example: Salary (Unequal Variances)

For the **unequal variances** approach we need to calculate:

$$a = \frac{s_1^2}{n_1} = \frac{0.407}{5} = 0.0814, \qquad b = \frac{s_2^2}{n_2} = \frac{5.4625}{4} = 1.3656.$$

$$\begin{aligned}\Rightarrow \nu = \frac{(a+b)^2}{\frac{a^2}{n_1-1} + \frac{b^2}{n_2-1}} = \frac{(0.0814+1.3656)^2}{\frac{0.0814^2}{5-1} + \frac{1.3656^2}{4-1}} &= \frac{1.447^2}{\frac{0.0814^2}{4} + \frac{1.3656^2}{3}} \\ &= \frac{2.0938}{0.6233} \\ &= 3.36.\end{aligned}$$

Since only whole number ν values appear in the tables, we round this to $\nu = 3$.

Example: Salary (Unequal Variances)

A 95% confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{3,\,0.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(32.92 - 37.975) \pm 3.182\sqrt{\frac{0.407}{5} + \frac{5.4625}{4}}$$

$$-5.055 \pm 3.182\sqrt{0.0814 + 1.3656}$$

$$-5.055 \pm 3.182\sqrt{1.447}$$

$$-5.055 \pm 3.182\,(1.203)$$

$$-5.055 \pm 3.828$$

$$[-8.883,\ -1.227]$$

Example: Salary (Equal Variances)

For the **equal variances** approach we should first apply the F test.

We will defer this for the moment and carry on as if the assumption $\sigma_1^2 = \sigma_2^2$ is reasonable here (shortly we will see that it isn't).

$$
\begin{aligned}
s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} &= \frac{(5 - 1)\, 0.407 + (4 - 1)\, 5.4625}{5 + 4 - 2} \\
&= \frac{4\,(0.407) + 3\,(5.4625)}{7} \\
&= \frac{18.0155}{7} = 2.574.
\end{aligned}
$$

In this case the degrees of freedom are

$$\nu = n_1 + n_2 - 2 = 5 + 4 - 2 = 7.$$

Example: Salary (Equal Variances)

A 95% confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{7,0.025} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$(32.92 - 37.975) \pm 2.365 \sqrt{\frac{2.574}{5} + \frac{2.574}{4}}$$

$$-5.055 \pm 2.365 \sqrt{0.5148 + 0.6435}$$

$$-5.055 \pm 2.365 \sqrt{1.1583}$$

$$-5.055 \pm 2.365\,(1.076)$$

$$-5.055 \pm 2.545$$

$$[-7.60, -2.51]$$

Equal Variance Assumption

In the example just covered, note that $s_1^2 = 0.407$ and $s_2^2 = 5.4635$.

Consider the ratio

$$F = \frac{\text{larger variance}}{\text{smaller variance}} = \frac{5.4635}{0.407} = 13.42$$

which shows that s_2^2 is 13.42 times larger than s_1^2.

It seems unlikely that the true variances (σ_1^2 and σ_2^2) are equal. If this was the case then we would expect $F \approx 1$.

F Test

We can formally **test the hypothesis**

$$\boxed{\sigma_1^2 = \sigma_2^2}$$

using the **F test**.

Although we have not yet covered hypothesis testing, it is necessary to introduce the F test at this point.

Hence, we will only mention the basic details now.

F Test Procedure

1. Calculate $$\boxed{F = \frac{\text{larger variance}}{\text{smaller variance}} = \frac{s^2_{\text{larger}}}{s^2_{\text{smaller}}}}.$$

2. Find *critical value* F_{ν_1,ν_2} in the **F tables** where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ correspond to s^2_{larger} and s^2_{smaller} respectively.

3. If $F > F_{\nu_1,\nu_2}$ then we *reject the hypothesis* that $\sigma_1^2 = \sigma_2^2$.

F Tables

To find the critical value, F_{ν_1,ν_2}, go to:

- **Column** $\Rightarrow \nu_1 = n_1 - 1$ corresponding to s^2_{larger}.
- **Row** $\Rightarrow \nu_2 = n_2 - 1$ corresponding to s^2_{smaller}.

You will see *four* critical values for each ν_1 - ν_2 combination;
$\Rightarrow$ **select the value in brackets**.

Examples: $F_{1,3} = 17.4$, $F_{6,4} = 9.20$, $F_{5,7} = 5.29$ etc.

Example: Salary

Going back to the salary example, we had

$$F = \frac{s^2_{\text{larger}}}{s^2_{\text{smaller}}} = \frac{5.4635}{0.407} = 13.42.$$

$$\left.\begin{array}{lcl} s^2_{\text{larger}} = 5.4635 & \Rightarrow & \nu_1 = 4 - 1 = 3 \\ s^2_{\text{smaller}} = 0.407 & \Rightarrow & \nu_2 = 5 - 1 = 4 \end{array}\right\} \Rightarrow F_{3,4} = 9.98.$$

(note: index “1” denotes the sample corresponding to s^2_{larger})

Since $13.42 > 9.98$, we reject the hypothesis that $\sigma_1^2 = \sigma_2^2$.

In other words, the equal variance assumption is *not* appropriate here.

Paired Samples

We have dealt with the case of two *independent* groups where the difference between the means is estimated.

We now consider **paired** samples, i.e., *dependent* groups.

Each data value in group one has a unique match in group two
$\Rightarrow$ the measurements come in *pairs*.

Most commonly, these are *before and after* measurements.

Example: Training Program

Five individuals were subjected to a variety of fitness tests and given an overall fitness score. These individuals then followed a 6-week training program and their fitness levels were tested again. The results are as follows:

Individual	Before Program	After Program
1	68	75
2	45	50
3	83	78
4	77	85
5	60	57

Calculating Differences

"Before" and "After" pairs are dependent (i.e., relate to the same individual) $\Rightarrow$ *cannot* use the approach for independent samples.

In fact, the case of paired samples is very easy to deal with. We simply define a new variable:

$$\boxed{\text{Difference} = \text{After} - \text{Before}}$$

and apply the single mean formula.

Note: The calculated differences need to be approximately normal to use the t distribution (but Before and After do not need to be).

Example: Training Program

Individual	Before Program	After Program	Difference x	x^2
1	68	75	7	49
2	45	50	5	25
3	83	78	-5	25
4	77	85	8	64
5	60	57	-3	9
		$\sum$	12	172

$$\bar{x} = \frac{\sum x}{n} = \frac{12}{5} = 2.4.$$

$$s^2 = \frac{\sum x^2 - n\,\bar{x}^2}{n-1} = \frac{172 - 5(2.4^2)}{4} = 35.8 \quad \Rightarrow s = \sqrt{35.8} = 5.98.$$

Example: Training Program

Note that $\nu = n - 1 = 4$ for the t value. Thus, the 95% confidence interval is

$$\bar{x} \pm t_{4,\,0.025}\,\frac{s}{\sqrt{n}} \qquad \Rightarrow \qquad 2.4 \pm 2.776\left(\frac{5.98}{\sqrt{5}}\right)$$

$$2.4 \pm 2.776\,(2.6743)$$

$$2.4 \pm 7.4239$$

$$[-5.02,\, 9.82]$$

We are 95% confident that the true mean (of the differences) lies in the above interval which includes $\mu = 0$. Thus, the training program is not successful, i.e., it does not improve fitness.

R Code: One Mean

Confidence intervals for means are calculated using `t.test` in R.

By default a 95% confidence interval is calculated.

```
x = c(38.3, 38.9, 39.2, 39.2, 39.6, 41.0)

t.test(x)
```

For other confidence levels, use the `conf.level` option.

```
t.test(x,conf.level=0.99)
t.test(x,conf.level=0.9)
t.test(x,conf.level=0.8)
```

Note: The output includes a *hypothesis test* in addition to the confidence interval. This topic will be covered in the next lecture.

R Code: Two Means

We can also compare means in two independent samples.

By default the unequal variance approach is used (i.e., Welch's t test).

```
x1 = c(32.1, 32.4, 33.2, 33.3, 33.6)
x2 = c(35.7, 36.3, 39.4, 40.5)
t.test(x1,x2,conf.level=0.95)
```

Note: The confidence interval is slightly different to that of slide 22 since we rounded ν to 3 whereas R can use the exact value $\nu = 3.36$.

For the equal variance approach set `var.equal = TRUE`.

```
t.test(x1,x2,conf.level=0.95,var.equal=TRUE)
```

R Code: F Test

To test equality of variances, we use the F test. In R, this is achieved via the `var.test` function.

```
var.test(x1,x2,conf.level=0.95)
```

A 95% confidence interval for the *ratio* of true variances, $\sigma_1^2 \,/\, \sigma_2^2$, is calculated. If this interval includes the value 1, then the hypothesis $\sigma_1^2 \,/\, \sigma_2^2 = 1$ is supported, i.e., $\sigma_1^2 = \sigma_2^2$.

Note: When we carry out the F test by hand, we put the larger sample variance on top. Recall that s_2^2 was the bigger variance for this data. Thus, we can swap `x1` and `x2` to match our previous work.

```
var.test(x2,x1,conf.level=0.95)
```

R Code: Paired Samples

We can deal with paired samples using the option `paired = TRUE`.

```
x1 = c(68,45,83,77,60)
x2 = c(75,50,78,85,57)
t.test(x1,x2,conf.level=0.95,paired=TRUE)
```

For “After” – “Before”, swap `x1` and `x2`.

```
t.test(x2,x1,conf.level=0.95,paired=TRUE)
```

Of course we can also manually define the difference variable.

```
x = x2 - x1
t.test(x,conf.level=0.95)
```

R Code: Large Samples

Recall that when samples are large, $t_{\nu,\alpha/2} \approx z_{\alpha/2}$. (see slide 10)

Thus, the `t.test` function can be used for samples of *all sizes*.

- If samples are small then the appropriate t value will be used. In this case the samples should be approximately normal.
- If samples are large then the t value will automatically become a z value. In this case the samples do not need to be normal.

R Code: One Proportion

Confidence intervals for proportions are calculated using `prop.test`.

```
x = 359
n = 500

prop.test(x,n,conf.level=0.95)
```

The above relates to 359 individuals who use Android devices out of a sample of 500 (see Q1 of Tutorial7).

Note: R uses a slightly different method for calculating confidence intervals for proportions to what we use in this course - so the results will be different.

R Code: Two Proportions

We can also compare proportions in two independent samples.

```
x = c(95,103)
n = c(130,150)

prop.test(x,n,conf.level=0.95)
```

The above relates to 95 out of 130 individuals compared with 103 out of 150 individuals (see slide 20 of Lecture13).