Confidence Intervals and Tests using the t-Distribution Cheat Sheet

Edexcel FS2

Mean of a normal distribution with unknown variance

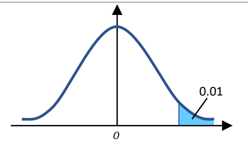
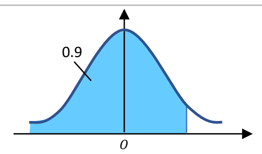
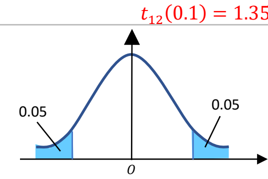
For a normally distributed random variable X, the sample mean is also distributed normally, which is denoted $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. As a result, as long as you know the population variance, you can construct a confidence interval for the population mean, μ.

- If the sample size, n, is large then you often will not know the population variance. Instead, the sample variance can be used as a good approximation.
- If n is small, the sample variance is often not sufficiently close to the population variance, and $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ or the z-score can no longer be modelled by the standard normal distribution $N(0, 1^2)$. Instead, we use the symbol t to denote $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$.

Because of this, we consider a different distribution for small values of n:

- If a random sample $X_1, X_2, \ldots, X_n$ is selected from a normal distribution with mean μ and unknown variance s^2, then $t = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ has a t_{n-1}-distribution where $S^2 = \frac{1}{n}(\sum x^2 - n\bar{x}^2)$, which is an unbiased estimator of σ^2.
- The t-distribution has degrees of freedom, $v = n - 1$, and as $v \to \infty$, the t-distribution becomes more and more similar to the standard normal distribution, $N(0, 1^2)$, so as expected, the t-distribution is only used when n is small.
- Much like with the F-distribution and the chi-squared distribution, the degrees of freedom affect the critical values, and these can be found using the tables in the formula book, or on some graphical calculators. Like when working with any distribution that uses tables, it is so important to draw a diagram to avoid simple mistakes.

Example 1: The random variable X has a t-distribution with 12 degrees of freedom. Determine values of t for which: i) $P(X > t) = 0.01$, ii) $P(X < t) = 0.9$, iii) $P(|X| > t) = 0.1$. (You will need the statistical table from the formula book)

As stated at the top of the table, the values in the table are those which a random variable that follows the t-distribution with v degrees of freedom will exceed with the probability shown on the top row. As we are looking for $P(X > t)$, we can read directly off the table.	 $t_{12}(0.01) = 2.681$	Read directly off the table where the $v = 12$ row intersects with the 0.01 column
For $P(X < t)$, we are looking at the area to the left of the line that we draw. We can use the fact that the area under the curve is 1 to work out the area that we need using the values in the table.	 $t_{12}(0.1) = 1.356$	If $P(X < t) = 0.9$, then $P(X > t) = 1 - 0.9 = 0.1$
$P(\|X\| > t) = P(X < -t) + P(X > t)$. We need to read off the value for $P(X > t)$ and use the symmetry of the graph to find the other t value, which will be the negative of the one we have found. Remember as this is a two tailed problem the total of the two tails must add up to the probability given in the question.	 So we read off where $P(X > t) = 0.05$, by looking at the intersection of the $v = 12$ row with the 0.05 column and find $t_{12}(0.05) = 1.782$. The t-value in the question is therefore 1.782 and the other tail is found below $t = -1.782$.	$P(X < -t) + P(X > t) = 0.1$, so each tail has an area of 0.05. Due to the symmetry of the graph, the t-values will be a negative of each other.

You should also be able to calculate the probability of a random variable that follows a t-distribution is greater/less than a given number.

Example 2: The random variable Y has a t_6-distribution. Determine $P(Y < 1.440)$

From looking at the $v = 6$ row, we can see that 1.440 is in the 0.10 column, but we need to pay attention to the direction of the inequalities	From the table, $P(Y > 1.440) = 0.1$ However, we are looking for $P(Y < 1.440)$, so using the fact that the area under the curve is 1, we know that $P(Y < 1.440) = 1 - P(Y > 1.440) = 1 - 0.1 = 0.9$

You can also use the t-distribution to find a confidence interval for the mean of a normal distribution where the sample variance is unknown. As shown before, a sample taken from a normal distribution with an unknown variance $t = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ has a t_{n-1}-distribution:

For a small sample size n from a normal distribution $N(\mu, \sigma^2)$ with an unknown mean and variance:

- The $100(1-\alpha)\%$ confidence **limits** for the population mean, μ, are
$$\bar{x} \pm t_{n-1}\left(\frac{\alpha}{2}\right) \times \frac{s}{\sqrt{n}}$$
- The $100(1-\alpha)\%$ confidence **interval** for the population mean, μ, is
$$\left(\bar{x} - t_{n-1}\left(\frac{\alpha}{2}\right) \times \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1}\left(\frac{\alpha}{2}\right) \times \frac{s}{\sqrt{n}}\right)$$

Example 3: A sample of 8 biscuits are taken by a quality control inspector in a factory and the weight in grams are measured. The weights are as follows: 15, 14.5, 16.5, 14.7, 13.9, 15.4, 13.6, 16.7. Assuming the weight of the biscuits are normally distributed, find a 95% confidence interval for the mean weight of biscuits in the factory.

Find the sample mean and variance	Using a calculator gives $\bar{x} = 15.0375$, $s = 1.1211$
Put the values you have found into the formula $\left(\bar{x} - t_{n-1}\left(\frac{\alpha}{2}\right) \times \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1}\left(\frac{\alpha}{2}\right) \times \frac{s}{\sqrt{n}}\right)$	$\bar{x} = 15.0375, s = 1.211, n = 8, \alpha = 0.05$, $t_7(0.025) = 2.365$. So the confidence interval is $15.0375 \pm (2.365) \times \frac{1.211}{\sqrt{8}}$ $= (14.025, 16.050)$

Hypothesis test for the mean of a normal distribution with unknown variance

To conduct a hypothesis test of a normal distribution with unknown variance, follow the same steps as when you were testing a mean of a normal distribution with known variance:

1. State the null hypothesis H_0 and the alternative hypothesis H_1
2. Specify the significance level, α, and the number of degrees of freedom, v.
3. Write down the critical region
4. Calculate $\bar{x}, s^2$ and t- remember these are sample parameters
5. Write the conclusion, explaining if the result is significant, and then link it back to the original problem

Example 4: The last time a town took a census, the mean height of adult women was 164cm. After increasing the amount of food produced, a sample of 6 women were measured and the mean height was 167.5cm. The heights of the women are assumed to be normally distributed, and an estimate for the standard deviation of the heights of women based on the sample of 6 is 3cm. Test, at the 5% significance level whether the women in the town have gotten taller.

State the hypotheses	$H_0: \mu = 164, H_1: \mu > 164$
State the significance level and the degrees of freedom	$\alpha = 0.05$ (one-tailed test), $v = 6 - 1 = 5$
Find the critical value on the table by looking at the value in the 5th row and the column labelled 0.05. We need a positive t-value as we are doing a right-tailed test	The critical value t_5 is 2.015, so the critical region is $t \geq 2.015$.
Use the values of s and $\bar{x}$ given to calculate t.	$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{167.5 - 164}{\frac{3}{\sqrt{6}}} = 2.858$
Accept or reject the null hypothesis	As 2.015<2.858, the result is statistically significant and H_0 is rejected.
Put the conclusion in terms of the original problem	There is sufficient evidence to suggest that the mean height of the women has increased and they have gotten taller

Make sure to pay attention to the wording in the question- if you are looking whether the mean has 'changed' instead of increased or decreased, you may need to do a 2-tailed test

The paired t-test

It is often useful to test the 'before' and 'after' of an experiment, for example how effective a treatment is for increasing reaction times. For these types of tests, a result in one sample is paired with the result in another sample and is therefore referred to as paired. In paired experiments we focus on the difference, D, between the results, which we assume is normally distributed, and as we are unlikely to know the standard deviation of the population, we model with the t-distribution. Often we take $H_0: \mu_D = 0$, as in there is no difference between the two populations.

With $\bar{D}$ is the mean of the differences between the samples:

$$\frac{\bar{D} - \mu_D}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

The paired t-test proceeds in a very similar way to the t-test itself, but before calculating the test statistic we must calculate the difference between the two samples.

Example 5: A class of students take a maths test, scored out of 50, and then listen to a seminar about improving problem solving skills. They then take a different maths test of the same difficulty, and their results are shown below:

Student	A	B	C	D	E	F	G	H	I	J	K
Score 1	40	45	37	33	49	27	36	42	37	44	39
Score 2	42	47	43	30	46	35	36	45	41	41	40
Difference	2	2	6	-3	-3	7	0	3	4	-3	1

Test, at the 5% significance level, whether or not the seminar improved the student's maths test performances.

State your hypotheses, significance level, the number of degrees of freedom and therefore the critical value	$H_0: \mu_d = 0, H_1: \mu_d > 0$ Significance level= 0.05, one tailed test $v = 11 - 1 = 10$ From the table, the critical value $t_{10}(0.05) = 1.812$, thus the critical region is $t \geq 1.812$
Find the differences (added onto the table) and calculate $\bar{d}$ and s^2	$\bar{d} = \frac{\sum d}{n} = \frac{16}{11} = 1.4545$ $s^2 = \frac{\sum d^2 - n\bar{d}^2}{n-1} = \frac{146 - 11(1.45)^2}{10}$ $= 14.4545$
Calculate the test statistic, we put $\mu_D = 0$ as that is what we are testing	$t = \frac{\bar{d} - \mu_D}{\frac{s}{\sqrt{n}}} = \frac{1.4545 - 0}{\frac{\sqrt{14.4545}}{\sqrt{11}}}$ $= 1.2688$
Write your conclusion	$1.2688 < 1.812$, so the result isn't statistically significant and we do not reject H_0. There is insufficient evidence to suggest that the seminar improved the student's maths test performances.

Difference between means of two independent normal distributions

It is possible to find a confidence interval for the difference between two means from normal distributions with unknown but equal variances by finding a pooled estimate of variance.

- If a random sample of n_x observations is taken from a normal distribution with unknown variance σ^2 and an independent sample of n_y observations is taken from a normal distribution that also has unknown variance σ^2, then a pooled estimate for σ^2 is:
$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}, \text{ where } s_x^2 = \frac{\sum x^2 - n_x\bar{x}^2}{n_x - 1} \text{ and } s_y^2 = \frac{\sum y^2 - n_y\bar{y}^2}{n_y - 1}$$
- If a random sample of n_x observations is taken from a normal distribution that has unknown variance σ^2 and an independent sample of n_y observations that is taken from a normal distribution with equal variance, then
$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{S_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t_{n_x+n_y-1}, \text{ where } S_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$$
- The confidence limits for the difference between two means from independent normal distributions X and Y, when the variances are equal but unknown are given by
$$(\bar{x} - \bar{y}) \pm t_c s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$$
Where s_p is the pooled estimate of the population variance and t_c is the relevant value taken from the t-distribution tables, and therefore the confidence interval is given by
$$\left((\bar{x} - \bar{y}) - t_c s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}, (\bar{x} - \bar{y}) + t_c s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}\right)$$

Example 6: A packet of seeds were sown, with 10 being sown into normal compost, brand A, and 15 being sown into a new brand of compost with additional nutrients, brand B. After 6 weeks of growth, the heights of the plants were measured in cm with the following results:

Brand A: 10.2, 11, 10.5, 9.7, 11.1, 9.3, 10.4, 11.3, 10.3, 9.4

Brand B: 12.4, 12.7, 11.9, 11.3, 13.3, 13.6, 12.9, 11.7, 12.2, 12.5, 11.8, 13, 13.4, 12.6, 12.8

Calculate a 90% confidence interval for the difference between the two mean heights. You can assume that the variables are normally distributed and have the same variance.

Calculate $n_x, n_y, \bar{x}, \bar{y}, s_x^2$ and s_y^2.	For brand A: $n_x = 10, \bar{x} = 10.32, s_x^2 = 0.484$ For brand B: $n_y = 15, \bar{y} = 12.473, s_y^2 = 0.4735$
Calculate s_p^2, the pooled estimate for σ^2.	$s_p^2 = \frac{9 \times 0.484 + 14 \times 0.4735}{10 + 15 - 2} = 0.4776$ $s_p = \sqrt{0.4776} = 0.6911$
Find the relevant value from the t-table, remembering that confidence intervals are two tailed so half the percentage.	$t_c = t_{23}(0.05) = 1.714$
State the confidence limits	$(12.473 - 10.32) \pm 1.714 \times 0.6911\sqrt{\frac{1}{15} + \frac{1}{10}}$ $= 2.153 \pm 0.4836$ $= (1.6694, 2.6366)$

Hypothesis test for the difference between means

Doing a hypothesis test for the difference of means is similar to every other hypothesis test you have done. Doing a hypothesis test for two independent distributions with unknown variances require you to use the t-distribution.

Example 7:

A class of 16 students, 8 boys and 8 girls have their heights measured. After analysis it was found that $\bar{x} = 152$cm, $\bar{y} = 134$cm, $s_x^2 = 43\text{cm}^2$, $s_y^2 = 51\text{cm}^2$ where x is the height of a boy and y is the height of a girl. Conduct a two-sample t-test at the 10% significance level to determine whether the mean height of boys is greater than the mean height of girls by more 5cm.

State your hypotheses, significance level, the number of degrees of freedom and therefore the critical value	$H_0: \mu_x = \mu_y + 5, H_1: \mu_x > \mu_y + 5$ Significance level= 0.1, one tailed test $v = 8 + 8 - 2 = 14$ From the table, the critical value $t_{14}(0.1) = 1.345$, thus the critical region is $t \geq 1.345$
Calculate s_p^2, the pooled estimate for σ^2. Square root to obtain s_p.	$s_p^2 = \frac{7 \times 43 + 7 \times 51}{8 + 8 - 2} = 47$ $s_p = \sqrt{47} = 6.8556\ldots$
Calculate t observing that $\mu_x - \mu_y = 5$ from the hypothesis	$t = \frac{(152 - 134) - 5}{\sqrt{47} \times \sqrt{\frac{1}{8} + \frac{1}{8}}} = 3.792$ (4s.f.)
Conclude	$3.792 > 1.345$ so t is in the critical region so reject H_0 as there is sufficient evidence to suggest that the mean height of boys in the class is greater than the mean height of girls by more than 5cm.

cc BY NC ND