1 Hypotheses test about μ if σ is not known

In this section we will introduce how to make decisions about a population mean, μ, when the standard deviation is not known. In order to develop a confidence interval for estimating μ and a test we need to study the sample distribution of a score, which does not involve the population standard deviation like the z-score.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

From the central limit theorem, we learn that if the population is normally distributed or the sample size is large and the true population mean $\mu = \mu_0$, then z has a standard normal distribution.
But the assumption that we know σ is too strong. Remember, we consider the case that we do not know μ, why would we know σ then?

For this reason we only consider the case where the population standard deviation σ is not known.
Since the z-score is based on the true population standard deviation we can not use it anymore.

1.1 t-distribution

We will consider the t-score, which we get when replacing the true standard deviation σ by its estimate s.

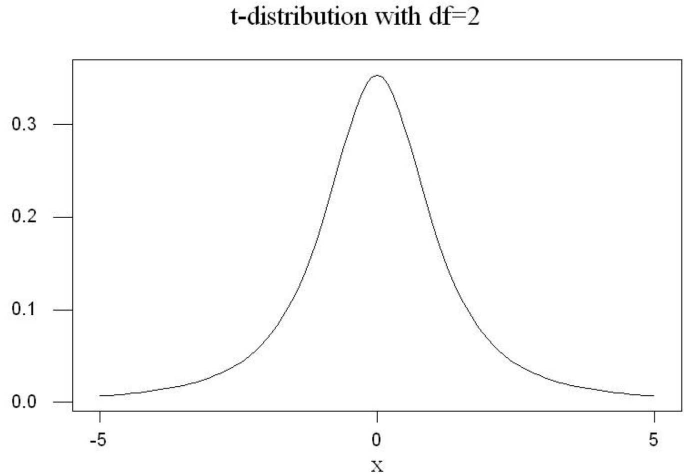
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Now we have to discuss the distribution of this score so we can determine probabilities and percentiles, which we need for making decisions in our test.

Definition:
Assume we have a normally distributed population with mean μ and standard deviation σ. Then

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has a t-distribution with $n - 1$ degrees of freedom ($df = n - 1$)



Critical values for the t-distribution can be found in table C in your textbook.

1.2 One sample t-confidence interval

To estimate a mean μ from a population using sample data, when the the population is either normal, or the sample size is large we use a "One sample t-confidence interval":

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

is a C confidence interval for μ, if t^* is the C critical value from a t-distribution with $df = n-1$.

Example:
The FBI compiles data on robbery and property crimes and publishes the information. A simple random sample of pickpocket offences yielded the losses, in dollars, described by the following summary statistics. $\bar{x} = 513.32$, $s = 262.23$, and $n = 25$.
We want to use this data to estimate μ, the mean loss in all pickpocket offenses in the U.S. using a 95% confidence interval.
Use table C to find t^*: $df = 25 - 1 = 24$ and the confidence level is 95%, therefore $t^* = 2.064$.
The confidence interval for μ is from

$$\bar{x} - t^* \frac{s}{\sqrt{n}} \quad \text{to} \quad \bar{x} + t^* \frac{s}{\sqrt{n}}$$

which is

$$513.32 - 2.064 \frac{262.23}{\sqrt{25}} \quad \text{to} \quad 513.32 + 2.064 \frac{262.23}{\sqrt{25}}$$

or \$405.07 to \$621.57.
We can be 95% confident that the mean loss of all pickpocket offenses is somewhere between \$405.07 and \$621.57.

With 95% confidence we give the chance that the sample we used for finding the confidence interval will result in an interval which actually includes the true unknown value of μ.

1.3 One sample t-test

The test we are about to introduce applies a t-score as the test statistic.
The test statistic will be

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

If $\mu = \mu_0$ is true, t_0 has a t-distribution with $n-1$ degrees of freedom.

t test about a mean μ, σ is unknown

1. Hypotheses:

   Choose μ_0.

Test type	**Hypotheses**
Upper tail	$H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$
Lower tail	$H_0 : \mu \geq \mu_0$ versus $H_a : \mu < \mu_0$
Two tail	$H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$

   Choose α.

2. Assumptions:
   Either the population is normally distributed or the sample size is large.

3. Test statistic:

   $$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad df = n - 1$$

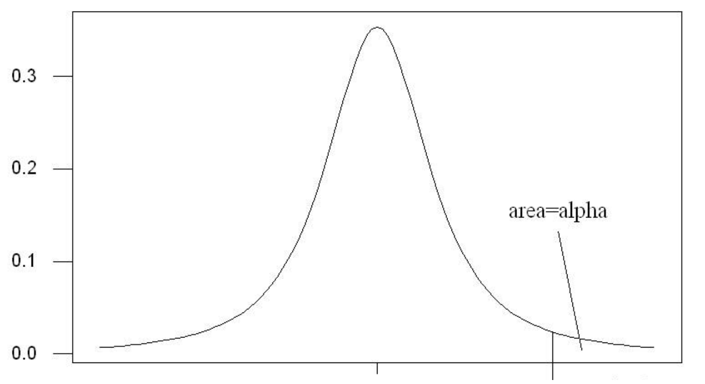
4. P-value:

Test type	**P-value**
Upper tail	$P(t > t_0)$
Lower tail	$P(t < t_0)$
Two tail	$2 \cdot P(t > abs(t_0))$

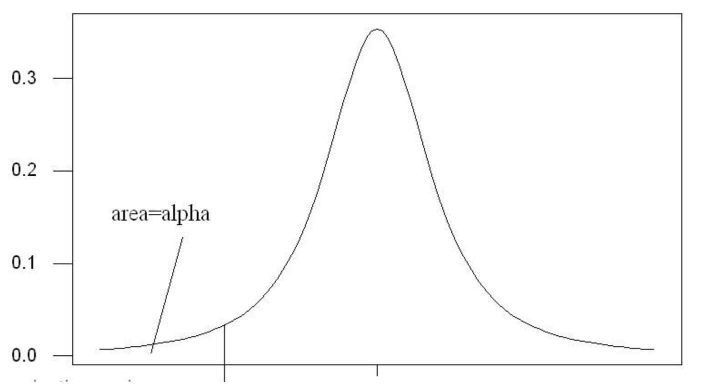
5. Decision:
   $p - value \leq \alpha$ then we reject H_0 and accept H_a.
   $p - value > \alpha$ then we do not reject H_0 (fail to reject H_0).

6. Context:
   Now put the finding into the context of the problem.

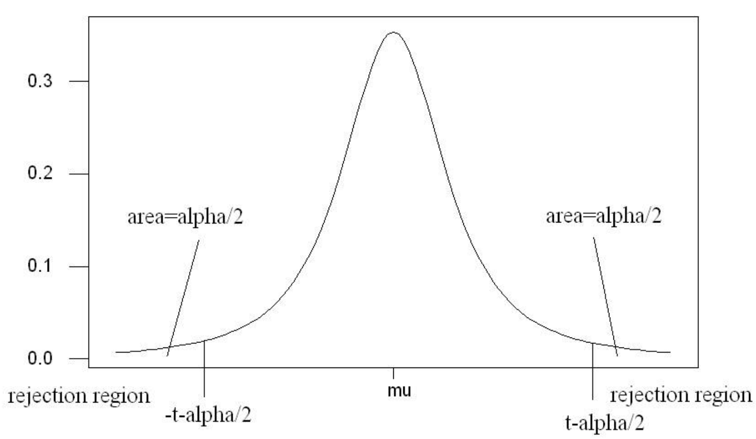
rejection region for uppertail test



rejection region for lowertail test



rejection region for 2-tail test



Definition:
The p-value of a statistical test is the probability to observe the value of the test statistic if in fact H_0 is true.

The p-value tells us how likely we would see the data from the sample if H_0 would be true. If this probability is small then we decide that H_0 can not be true (because if it would be true it is so unlikely to see this data.) and therefore reject H_0

When we conduct a statistical test in the steps outlined above, we know that the probability that our conclusion is wrong, when we reject H_0, is equal to α. This is a small probability we chose to be acceptable and we are quite certain that our decision is correct.

But if we fail to reject H_0, and decide that H_0 must be true, we don't know the probability for our conclusion being wrong. This is the reason we shall not commit ourselves to the decision that H_0 is correct.

Remark:
From Table C in the appendix we can find upper and lower bounds for the p-value for the different degrees of freedom.

Example 1 A study is conducted to look at the time students exercise in average. A researcher claims that in average students exercise less than 15 hours per month.
In a random sample of size $n = 115$ he finds that the mean time students exercise is $\bar{x} = 11.3$h/month with $s = 6.43$h/month.

1. $H_0 : \mu \geq 15$ versus $H_a : \mu < 15$ at a significance level of $\alpha = 0.05$.
2. Since $n = 115$ we conclude that the sample size is large and the t-test is the appropriate tool.
3. The test statistic is
$$t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{11.3 - 15}{\frac{6.4}{\sqrt{115}}} = -6.20$$
4. This is a lower tail test, so the P-value= $P(t < t_0) = P(t > abs(t_0))$, use $df = 100$ in table C (always go down). Since 6.20 is larger than the largest value in this row conclude that the P-value< 0.0005.
5. Decision: Since P-value$< \alpha = 0.05$, we reject H_0 at the significance level 0.05.
6. At significance level of 5% the data provide sufficient evidence that the mean time students exercise is less than 15h/month.

Example 2

1. Let $df = 40$ and $t_0 = 2$, then the P-value for an upper tailed test falls between 0.025 and 0.05 (diagram), because 2 fall between 1.684 and 2.021 and therefore the P-value falls between 0.025 and 0.05.
2. Let $df = 40$ and $t_0 = 2$, then the P-value for a two-tailed test falls between 0.05 and 0.1 (diagram).
3. Let $df = 40$ and $t_0 = 2$, then the P-value for a lower tailed test is greater than 0.5 (diagram).
4. Let $df = 40$ and $t_0 = -2$, then the P-value for a lower tailed test falls between 0.025 and 0.05 (diagram).

2 Comparing Two Population Means

Most statistical studies are designed to compare two or more populations.

Examples:

- Is the new developed drug effective. Compare it with a control.
- Which company pays a better salary, compare!
- Comparing two strategies in a game. Is one better than the other?

Suppose we are investigating two populations, we will use the following notation:

Notation:

	population		sample		
	mean	standard deviation	size	mean	standard deviation
population 1	μ_1	σ_1	n_1	$\bar{x_1}$	s_1
population 2	μ_2	σ_2	n_2	$\bar{x_2}$	s_2

We will use samples from the 2 populations for drawing conclusions about the difference in μ_1 and μ_2 (these are the population means).
In this case we need to distinguish the following two situations:

1. independent samples – example: height of males and females

   Using one random sample of males and one random sample of females, where there is no relationship between measurements in the two samples, then the values in one sample are independent from the values in the other sample.

2. paired samples – example: blood pressure in the morning and evening

   A good way of choosing samples would be to measure the blood pressure of all participants in the morning as well as the evening.

   Sample one would consist of the measurements taken in the morning and sample 2 would consist of measurements taken in the evening. The two samples are paired through the individuals, because it is reasonable to assume that the blood pressure measured in the evening somewhat depends on the blood pressure measured in the SAME person in the morning.

   The samples are paired.

   The two samples are connected, they are NOT independent.

2.1 t-Procedure for Two Paired Samples

In order to control extraneous factors in some studies you can use paired samples. In this case for every measurement in the sample from population 1 you find a matching measurement from population 2. Typical example, are measurements before therapy in sample 1, and measurements for the same individuals after therapy in sample 2.
In this case it is always $n = n_1 = n_2$.

Example 3

- Compare the resting pulse and pulse after exercise.

  To control for all other influences, you take both measurements in every individual in a sample.

We are interested in the difference in the population means $\mu_d = \mu_1 - \mu_2$. By studying this difference we can decide if the means are equal, or which i greater than the other. I for example $\mu_1 - \mu_2 > 0$, than $\mu_1 > \mu_2$.

In the case of paired samples we will turn our data into ONE sample. For this we will be able to apply the one-sample t-test.

To turn the two samples into one, we will find the pairwise differences.

$$\text{sample 1 value} - \text{sample 2 value}$$

This creates **one** sample of n differences.
For this sample of differences we then obtain the sample mean and the sample standard deviation, $\bar{x}_d$ and s_d.

Similar to the one-sample t-score we can now determine the distribution of the t-score based on $\bar{x}_d$ and s_d:
If the differences come from a normal distribution, then the t-score

$$t = \frac{\bar{x}_d - (\mu_1 - \mu_2)}{s_d/\sqrt{n}}$$

is t-distributed with $df = n - 1$.

Paired t-Test for Comparing Two Population Means

1. Hypotheses:

Test type	Hypotheses
Upper tail	$H_0 : \mu_d \leq d_0 \Leftrightarrow \mu_1 - \mu_2 \leq d_0$ versus $H_a : \mu_d > d_0 \Leftrightarrow \mu_1 - \mu_2 > d_0$
Lower tail	$H_0 : \mu_d \geq d_0 \Leftrightarrow \mu_1 - \mu_2 \geq d_0$ versus $H_a : \mu_d < d_0 \Leftrightarrow \mu_1 - \mu_2 < d_0$
Two tail	$H_0 : \mu_d = d_0 \Leftrightarrow \mu_1 - \mu_2 = d_0$ versus $H_a : \mu_d \neq d_0 \Leftrightarrow \mu_1 - \mu_2 \neq d_0$

2. Assumption: n is large or the distribution of differences is approximately normal.

3. Test statistic:

$$t_0 = \frac{\bar{x}_d - d_0}{s_d/\sqrt{n}}$$

   with $df = n - 1$.

4. P-value:

Test type	P-value
Upper tail	$P(t > t_0)$
Lower tail	$P(t < t_0)$
Two tail	$2 \cdot P(t > abs(t_0))$

5. Decision: If P-value$\leq \alpha$, then reject H_0.
   If P-value$> \alpha$ then do not reject H_0.

6. Interpretation: Put decision into context.

Paired t-Confidence Interval for $\mu_d = \mu_1 - \mu_2$

Assumption: n is large or the population distribution of differences is approximately normal.

The C Confidence Interval for μ_d:

$$\bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}}$$

and t^* is the C critical value of the t-distribution with $n-1$ degrees of freedom (Table C).

Example 4 Two sun blockers are compared for their effectiveness. For the comparison half of the backs of participants are treated with sun blocker 1 and the other half with sun blocker 2. After exposing the participants to two hours of sun and waiting for another hour the two halves are rated for the redness.
The following table gives the data obtained

person	x_1	x_2	$d = x_1 - x_2$
1	2	2	0
2	7	5	2
3	8	4	4
4	3	1	2
5	5	3	2

Therefore $\bar{d} = 2$ and $s_d = \sqrt{2}$ (confirm these results).
To estimate the difference in the mean rating for sun blockers 1 and 2, we will use a 95% confidence interval. The point estimator is $\bar{x_1} - \bar{x_2} = \bar{d} = 2$. That is, it seems that the mean rating for sun blocker 1 is higher than for sun blocker 2. For the confidence interval: $t^4_{\alpha/2} = 2.776$.

$$2 \pm \underbrace{2.776 \frac{\sqrt{2}}{\sqrt{5}}}_{=ME=1.756}$$

giving [0.244, 3.756]. We are 95% confident that the difference in the mean rating for sunblocker 1 and 2 falls between 0.244 and 3.756. Since zero does not fall within the ci, we are 95%

confident that the difference is not 0 and one is better than the other. Since confidence interval falls above 0, we are 95% confident that the mean rating is higher for sun blocker 1, indicating that it is less effective (the area was more red than for the other sun blocker.)
Instead we could have conducted a test

1. Hypotheses: $H_0 : \mu_d = \mu_1 - \mu_2 = 0$ vs. $H_a : \mu_1 - \mu_2 \neq 00$. $\alpha = 0.05$

2. Assumptions: Random sample (we chose participants randomly, sample size is not large, so we have to assume that the difference in rating is normally distributed.

3. Test statistic: with $d_0 = 0$

$$t_0 = \frac{\bar{x}_d - d_0}{\frac{s_d}{\sqrt{n}}} = \frac{2 - 0}{\frac{\sqrt{2}}{\sqrt{8}}} = 3.16$$

   with 4 df.

4. P-value: This is a 2-tailed test, so the p-value=$2P(t > 3.16)$. Find the two numbers in the row with df=4, which enclose 3.16. They are 2.999 and 3.747, with 2-tailed P of 0.04 and 0.02 respectively.

   Therefore: 0.002< $P - value < 0.04$

5. Decision: Since the P-value is less than α=0.05, we reject H_0 and accept H_a.

6. Interpretation: At significance level of 5% the data provide sufficient evidence that the mean ratings for the two sun blockers are not the same.

Example 5 The effect of exercise on the amount of lactic acid in the blood was examined. Blood lactate levels were measured in eight males before and after playing three games of racquetball. Since the samples have to be considered paired, we first have to obtain the differences between the paired measurements (see table).

Player	**Before**	**After**	**Difference**
1	13	18	-5
2	20	37	-17
3	17	40	-23
4	13	35	-22
5	13	30	-17
6	16	20	-4
7	15	33	-18
8	16	19	-3

Now we will only need the differences. We get $\bar{x}_d = -13.63$, $s_d = 8.28$, $n = 8$
Let's test if the increase in lactate levels is significantly greater than 0 at significance level of 0.05, that is

1. Assumptions: It is reasonable to assume that the difference lactic acid levels (before - after) is normally distributed.

2. Hypotheses: $H_0 : \mu_d = \mu_{before} - \mu_{after} \geq 0$ vs. $H_a : \mu_d < 0$. $\alpha = 0.05$

3. Test statistic: with $d_0 = 0$

$$t_0 = \frac{\bar{x}_d - d_0}{\frac{s_d}{\sqrt{n}}} = \frac{-13.63}{\frac{8.28}{\sqrt{8}}} = -4.65597$$

with 7 df.

4. P-value: This is a lower tailed test, so the p-value=$P(t < t_0) = P(t < -4.65597)$ Because of symmetry this is

   P-value=$P(t > abs(t_0)) = P(t > 4.66)$ for a t-distribution with 7 df. From Table C we get 0.001 < P-value< 0.0025.

5. Decision: Since the P-value is less than α=0.05, we reject H_0 and accept H_a.

6. Interpretation: At significance level of 5% the data indicate that the lactate level is after three games of racquet ball significantly higher than before.

Let's give an estimate (95% Confidence interval) for the increase in lactate level through three games of racquetball in males. Find t^* in table C, by finding $df = 8 - 1 = 7$, and C=95%.

$$\bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}} = -13.63 \pm 2.365 \frac{8.28}{\sqrt{8}} = -13.63 \pm 6.938$$

or $(-20.568; -6.696)$. Based on the sample data, we can be 95 % confident that the mean decrease in lactate level falls between 6.692 and 20.568 after three racquetball games.

2.2 2 Sample t-Procedure for Two Independent Samples

When estimating the difference in the **population** means, $\mu_1 - \mu_2$, based on sample data the point estimator that comes first to my mind is the difference in the **sample** means, $\bar{x}_1 - \bar{x}_2$. In order to do inferential statistics using this difference we have to investigate the sampling distribution of this statistic. (Remember the value of $\bar{x}_1 - \bar{x}_2$ will be different for each set of samples, but $\mu_1 - \mu_2$ is the unknown fixed number, we would like to learn about.
We have to discuss the sampling distribution of the point estimator $\bar{x}_1 - \bar{x}_2$ of $\mu_1 - \mu_2$. Based on these results we will be able to obtain a confidence interval and a test for $\mu_1 - \mu_2$.

Sampling Distribution of $\bar{x}_1 - \bar{x}_2$ from two independent samples.

- For the mean: $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$, so that $\bar{x}_1 - \bar{x}_2$ is an unbiased estimate for $\mu_1 - \mu_2$.

- For the variance:
$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- For the standard deviation:
$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Shape: If n_1 and n_2 are both large or both populations are normally distributed, then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is (approximately) normal.

Conclusion:

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is under above assumptions standard normally distributed.
For inferential statistics it does not help since for evaluation you would need to know the population standard deviations σ_1 and σ_2.
So again we will have to consider the t-score.

The t-score

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

It can be shown that this t-score t is approximately t-distributed with degrees of freedom close to $df = \min(n_1 - 1, n_2 - 1)$. When using this distribution for finding confidence intervals and p-values the results are conservative, i.e. the confidence interval is wider than it has to be given the confidence level and the p-values found will be greater than the true one.
This is all the information we need to put together a

Two–sample t-Test for Comparing Two Population Means

1. Hypotheses:

Test type	Hypotheses
Upper tail	$H_0 : \mu_1 - \mu_2 \leq d_0$ versus $H_a : \mu_1 - \mu_2 > d_0$
Lower tail	$H_0 : \mu_1 - \mu_2 \geq d_0$ versus $H_a : \mu_1 - \mu_2 < d_0$
Two tail	$H_0 : \mu_1 - \mu_2 = d_0$ versus $H_a : \mu_1 - \mu_2 \neq d_0$

2. Assumption: Random samples and n_1 and n_2 are large or both populations are normally distributed.

3. Test statistic:

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

   with $df = \min(n_1 - 1, n_2 - 1)$.

4. P-value/Rejection Region:

Test type	P-value
Upper tail	$P(t > t_0)$
Lower tail	$P(t < t_0)$
Two tail	$2 \cdot P(t > abs(t_0))$

5. Decision: If P-value$\leq \alpha$, reject H_0.
   If P-value$> \alpha$, fail to reject H_0.

6. Interpretation: Put result into context.

Notice that the only difference between this test and the test introduced before is the choice of the test statistic. This is due to different assumptions. For the first test we assumed that the population standard deviations are the same. This assumption was dropped for the second test. In conclusion we get a different test statistic with a different distribution(df).

Example 6
A company wanted to show, that their vitamin supplement decreases the recovery time from a common cold. They selected randomly 70 adults with a cold. 35 of those were randomly selected to receive the vitamin supplement. The other group received a placebo. The data on the recovery time for both samples is shown below.

sample	1 placebo	2 vitamin
sample size	35	35
sample mean	6.9	5.8
sample standard deviation	2.9	1.2

1. Hypotheses: Now test the claim of the company: $H_0 : \mu_p - \mu_v \leq 0$ versus $H_a : \mu_p - \mu_v > 0$ at a significance level of α=0.05.

2. Assumptions: The sample sizes are large enough (greater than 30).

3. Test statistic: with $d_0 = 0$

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{6.9 - 5.8}{\sqrt{\frac{2.9^2}{35} + \frac{1.2^2}{35}}} = \frac{1.1}{0.53} = 2.07$$

   and $df = \min(34, 34) = 34$

4. P-value: Since this is an upper tail test the p-value=$P(t > t_0) = P(t > 2.07)$ for a t-distribution with 34 df. From Table C we get (for df=30) 0.02 $\leq$P-value$\leq$ 0.025.

5. Decision: Since the P-value is less than α=0.05, we reject H_0 and accept H_a.

6. Interpretation: We find that the recovery time under vitamin treatment is significantly shorter than the recovery time taking placebo at a significance level of 0.05.

Beside testing for difference or trend of the means it is also beneficial to have a confidence interval:

Two–sample t-Confidence Interval for Comparing Two Population Means

Assumption: n_1 and n_2 are large or both populations are approximately normal distributed.

The C Confidence Interval for $\mu_1 - \mu_2$:

$$\bar{x}_1 - \bar{x}_2 \pm t_C^{df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

with $df = \min(n_1 - 1, n_2 - 1)$

and t_C^{df} is the critical value of the t-distribution with the given number of degrees of freedom (Table C).

Continue Example:
Calculate a 95% Confidence Interval for the difference in recovery time $\mu_1 - \mu_2$!
The degrees of freedom are 34 (use df= 30 in table C):

$$\bar{x}_1 - \bar{x}_2 \pm t_C^{df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$6.9 - 5.8 \pm 2.042 \cdot 0.53$$

$$1.1 \pm 1.076$$

or (0.024 ; 2.176). The 95% confidence interval lies entirely above 0. So that 0 is with a confidence of 95% less than $\mu_1 - \mu_2$. We can state with confidence 0.95 that the recovery time without vitamin treatment takes longer than with vitamin treatment.