Comparison of Two Independent Samples

- In this chapter we continue our study of comparisons of two independent samples by introducing hypothesis testing.
- Question: How different do two samples have to be in order for us to infer that the populations that generated them are actually different?
- One way to approach this question is to compare the two sample means and to see how much they differ in comparison to the amount of difference we would expect to see due to chance.

Hypothesis Testing: The t Test

In Chapter 6 we saw that two means can be compared by using a confidence interval for the difference $\mu_1 - \mu_2$. Now we will explore another approach to the comparison of means: the procedure known as hypothesis testing. The general idea is to formulate as a hypothesis the statement that μ_1 and μ_2 differ and then to see whether the data provide sufficient evidence in support of that hypothesis.

The null and alternative hypotheses

The hypothesis that μ_1 and μ_2 are not equal is called an alternative hypothesis (or a research hypothesis) and is abbreviated H_A. It can be written as

$$H_A : \mu_1 \neq \mu_2.$$

Its antithesis is the null hypothesis,

$$H_0 : \mu_1 = \mu_2,$$

which asserts that μ_1 and μ_2 are equal.

A statistical test of hypothesis is a procedure for assessing the strength of evidence present in the data in support of H_A. The data are considered to demonstrate evidence for H_A if any discrepancies from H_0 (the opposite of H_A) could not be readily attributed to chance (i.e., to sampling error).

The t statistic

We consider the problem of testing the null hypothesis

$$H_0 : \mu_1 = \mu_2 \text{ or } H_0 : \mu_1 - \mu_2 = 0$$

against the alternative hypothesis

$$H_A : \mu_1 \neq \mu_2 \text{ or } H_A : \mu_1 - \mu_2 \neq 0.$$

The t test is a standard method of choosing between these two hypotheses. To carry out the t test, the first step is to compute the **test statistic**, which for a t test is defined as

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{\text{SE}_{\bar{Y}_1 - \bar{Y}_2}}.$$

Notice the structure of T: It is a measure of how far the difference between the sample means is from the difference we would expect to see if H_0 were true (zero difference), expressed in relation to the SE of the difference: the amount of variation we expect to see in differences of means from random samples.

Example: toluene and the brain

Abuse of substances containing toluene (e.g., glue) can produce various neurological symptoms. In an investigation of the mechanism of these toxic effects, researchers measured the concentrations of various chemicals in the brains of rats that had been exposed to a toluene-laden atmosphere, and also in unexposed control rats. The concentrations of the brain chemical norepinephrine (NE) in the medulla region of the brain, for six toluene-exposed rats and five control rats, are given in the following table.

Table 7.2.1 NE concentration (ng/gm)

	Toluene (Group 1)	Control (Group 2)
	543	535
	523	385
	431	502
	635	412
	564	387
	549	
n	6	5
$\bar{y}$	540.8	444.2
s	66.1	69.6
SE	27	31

The observed mean NE in the toluene group ($\bar{Y}_1 = 540.8$ ng/gm) is substantially higher than the mean in the control group ($\bar{Y}_2 = 444.2$ ng/gm). One might ask whether this observed difference indicates a real biological phenomenon (the effect of toluene) or whether the truth might be that toluene has no effect and that the observed difference between $\bar{Y}_1$ and $\bar{Y}_2$ reflects only chance variation.

- The SE for $\bar{Y}_1 - \bar{Y}_2$ is

$$\mathrm{SE}_{\bar{Y}_1-\bar{Y}_2} = \sqrt{\frac{66.1^2}{6} + \frac{69.6^2}{5}} = 41.195$$

and the value of the test statistic is

$$T = \frac{(540.8 - 444.2) - 0}{41.195} = 2.34.$$

- How shall we judge whether our data provide sufficient evidence for H_A? A lack of evidence for H_A (agreement with H_0) would be expressed by sample means that were similar and a resulting small test statistic T.
- It can be shown mathematically that **if H_0 is true, then the sampling distribution of the test statistic T is well approximated by a Student's t distribution with degrees of freedom given by**

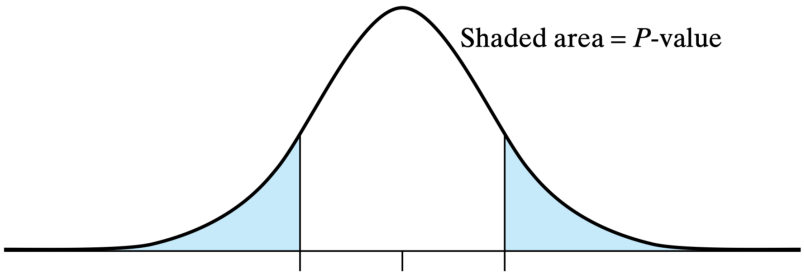
$$\nu = \frac{(\mathrm{SE}_1^2 + \mathrm{SE}_2^2)^2}{\mathrm{SE}_1^4/(n_1 - 1) + \mathrm{SE}_2^4/(n_2 - 1)}.$$

The essence of the t test procedure is to identify where the observed value T falls in the Student's t distribution.

- If T is near the center, then the data are regarded as compatible with H_0 because the observed difference between $\bar{Y}_1 - \bar{Y}_2$ and the null difference of zero can readily be attributed to chance variation caused by sampling error. (H_0 predicts that the sample means will be equal, since H_0 says that the population means are equal.)
- If, on the other hand, T falls in the far tail of the t distribution, then the data are regarded as evidence for H_A , because the observed deviation cannot be readily explained as being due to chance variation.

The p-value

- To judge whether an observed value T is "far" in the tail of the t distribution, we need a quantitative yardstick for locating T within the distribution.
- The (two-sided) p-value of the t test is the area under Student's t curve in the double tails beyond $-T$ and T.

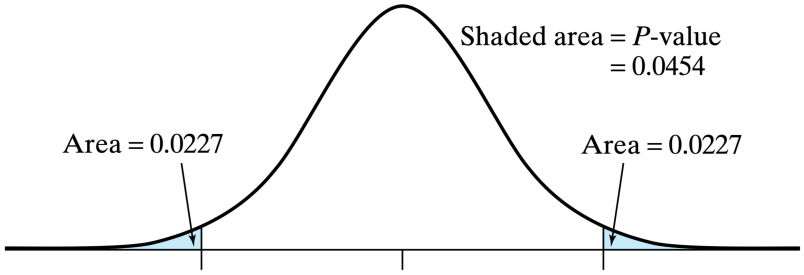


Example: toluene and the brain

For the brain NE data, the value of T is 2.34. We can ask, "If H_0 were true so that one would expect $\bar{Y}_1 - \bar{Y}_2 = 0$, on average, what is the probability that $\bar{Y}_1 - \bar{Y}_2$ would differ from zero by as many as 2.34 SEs?". The p-value answers this question. The formula

$$\nu = \frac{(\mathrm{SE}_1^2 + \mathrm{SE}_2^2)^2}{\mathrm{SE}_1^4/(n_1 - 1) + \mathrm{SE}_2^4/(n_2 - 1)}$$

yields 8.47 degrees of freedom for these data. Thus, the p-value is the area under the t curve (with 8.47 degrees of freedom) beyond ± 2.34. This area, which was found using a computer is shown in the following figure to be 0.0454.



The **p-value** for a hypothesis test is the probability, computed under the condition that **the null hypothesis is true**, of the test statistic being **at least as extreme as the value of the test statistic that was actually obtained**.

From the definition of p-value, it follows that the p-value is **a measure of compatibility between the data and H_0** and thus measures the **evidence for H_A**: A large p-value (close to 1) indicates a value of T near the center of the t distribution (lack of evidence for H_A), whereas a small p-value (close to 0) indicates a value of T in the far tails of the t distribution (evidence for H_A).

Drawing conclusion from a t test

- Making a decision requires drawing a definite line between sufficient and insufficient evidence. The threshold value, on the p-value scale, is called the **significance level** of the test and is denoted by the Greek letter α (alpha).
- The value of α is chosen by whoever is making the decision. Common choices are $\alpha = 0.10$, **0.05**, and 0.01.
- If the p-value of the data is less than α, the data are judged to provide statistically significant evidence in favor of H_A ; we also may say that H_0 **is rejected**.
- If the p-value of the data is greater than or equal to α, we say that the data provide insufficient evidence to claim that H_A is true, and thus H_0 **is not rejected**.

Example: toluene and the brain

- For the brain NE experiment, suppose we choose to make a decision at the 5% significance level, $\alpha = 0.05$.

- We found that the p-value of these data is 0.0454. This means that one of two things happened: Either (1) H_0 is true and we got a strange set of data just by chance or (2) H_0 is false.
- If H_0 is true, the kind of discrepancy we observed between $\bar{Y}_1$ and $\bar{Y}_2$ would happen only about 4.5% of the time.
- Because the p-value, 0.0454, is less than 0.05, we **reject** H_0 and conclude that the data provide statistically significant evidence in favor of H_A.
- The strength of the evidence is expressed by the statement that the p-value is 0.0454.
- **Conclusion**: The data provide sufficient evidence at the 0.05 level of significance (p-value $= 0.0454$) that toluene affects NE concentration.

Example: fast plants

In Chapter 6 we saw that the mean height of fast plants was smaller when ancy was used than when water (the control) was used. The following table summarizes the data.

Table 7.2.3 Fourteen-day height of control and of ancy plants

	Control	Ancy
n	8	7
$\bar{y}$	15.9	11.0
s	4.8	4.7

The difference between the sample means is $15.9 - 11.0 = 4.9$. The SE for the difference is

$$\text{SE}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{4.8^2}{8} + \frac{4.7^2}{7}} = 2.46.$$

Suppose we choose to use $\alpha = 0.05$ in testing

$$H_0 : \mu_1 - \mu_2 = 0$$

against the alternative hypothesis

$$H_A : \mu_1 - \mu_2 \neq 0.$$

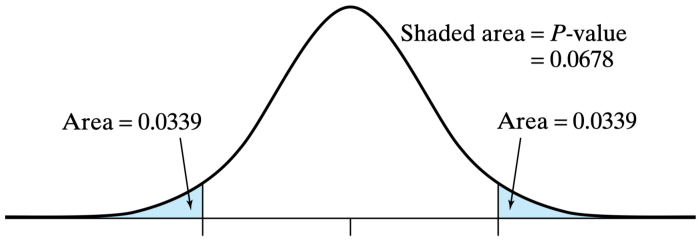
The value of the test statistic is

$$T = \frac{(15.9 - 11.0) - 0}{2.46} = 1.99.$$

Using the formula, we find the degrees of freedom to be:

$$\nu = \frac{(1.7^2 + 1.8^2)^2}{1.7^4/(8-1) + 1.8^4/(7-1)} = 12.8.$$

The p-value for the test is the probability of getting a t statistic that is at least as far away from zero as 1.99.



- Because the p-value is greater than α, we have insufficient evidence for H_A; thus, we **do not reject** H_0. That is, these data do not provide sufficient evidence to conclude that μ_1 and μ_2 differ; the difference we observed between $\bar{Y}_1$ and $\bar{Y}_2$ could easily have happened by chance.
- **Conclusion**: The data do not provide sufficient evidence (p-value $= 0.0678$) at the 0.05 level of significance to conclude that ancy and water differ in their effects on fast plant growth.

- Note carefully the phrasing of the conclusion for hypothesis testing. We do not say that there is evidence for the null hypothesis, but only that there is insufficient evidence against it.
- When we do not reject H_0, this indicates a lack of evidence that H_0 is false, which is not the same thing as evidence that H_0 is true.
- In other words, nonrejection of H_0 is not the same as acceptance of H_0. (To avoid confusion, it may be best **not** to use the phrase "accept H_0" at all.)
- In testing a hypothesis, the researcher starts out with the assumption that H_0 is true and then asks whether the data contradict that assumption.

In this section we have considered tests of the form $H_0 : \mu_1 - \mu_2 = 0$ against $H_A : \mu_1 - \mu_2 \neq 0$; this is the most common pair of hypotheses. However, it may be that we wish to test that μ_1 differs from μ_2 by some specific, nonzero amount, say c. To test $H_0 : \mu_1 - \mu_2 = c$ against $H_A : \mu_1 - \mu_2 \neq c$ we use the t test with test statistic given by

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - c}{\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2}}.$$

From this point on, the test proceeds as before (i.e., as for the case when $c = 0$).

Relationship between test and confidence interval

There is a close connection between the confidence interval approach and the hypothesis testing approach to the comparison of μ_1 and μ_2. The t test and the confidence interval use the same three quantities $\bar{Y}_1 - \bar{Y}_2$, $\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2}$, and $t_\nu(\alpha/2)$ but manipulate them in different ways.

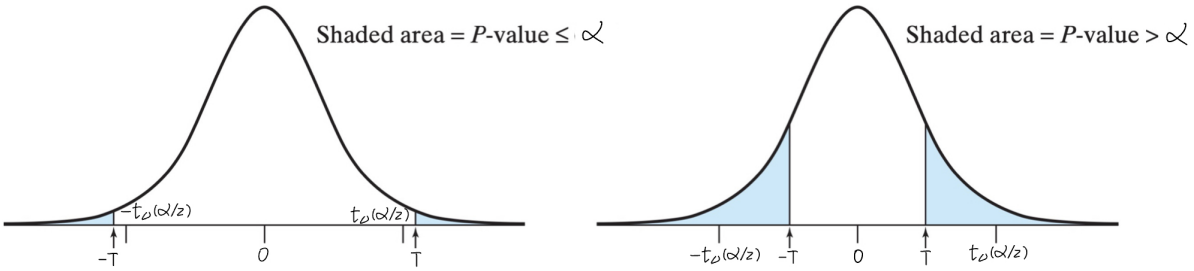
The p-value is less than or equal to α if and only if the test statistic T is in the tail of of the t distribution, at or beyond $\pm t_\nu(\alpha/2)$. Thus we lack significant evidence for $H_A : \mu_1 - \mu_2 \neq 0$ if and only if $|T| \leq t_\nu(\alpha/2)$, i.e.,

$$\frac{|\bar{Y}_1 - \bar{Y}_2|}{\text{SE}_{\bar{Y}_1 - \bar{Y}_2}} \leq t_\nu(\alpha/2).$$

This is equivalent to

$$(\bar{Y}_1 - \bar{Y}_2) - t_\nu(\alpha/2) \times \text{SE}_{\bar{Y}_1 - \bar{Y}_2} \leq 0 \leq (\bar{Y}_1 - \bar{Y}_2) + t_\nu(\alpha/2) \times \text{SE}_{\bar{Y}_1 - \bar{Y}_2}.$$

Thus we have shown that we lack significant evidence for $H_A : \mu_1 - \mu_2 \neq 0$ if and only if the confidence interval for $\mu_1 - \mu_2$ includes zero.



Rejection region

The rejection region for a hypothesis test is the set of values of the test statistic for which we reject the null hypothesis. It is determined based on the desired significance level (α), the degrees of freedom ν, and the alternative hypothesis H_A. If the test statistic falls within the rejection region, it provides significant evidence against the null hypothesis H_0.

To test $H_0 : \mu_1 - \mu_2 = 0$ against $H_A : \mu_1 - \mu_2 \neq 0$ at the significance level α, we reject H_0 at the significance level α if $|T| > t_\nu(\alpha/2)$. The corresponding rejection region is $\{T : |T| > t_\nu(\alpha/2)\}$.

Example: crawfish lengths

Biologists took samples of the crawfish species Orconectes sanborii from two rivers in central Ohio, the Upper Cuyahoga River (CUY) and East Fork of Pine Creek (EFP), and measured the length (mm) of each crawfish captured.

Table 7.3.1 Crawfish data: length (mm)

	CUY	EFP
n	30	30
$\bar{y}$	22.91	21.97
s	3.78	2.90

For these data the two SEs are $3.78/\sqrt{30} = 0.69$ and $2.90/\sqrt{30} = 0.53$ for CUY and EFP, respectively. The degrees of freedom are

$$\nu = \frac{(0.69^2 + 0.53^2)^2}{0.69^4/(30-1) + 0.53^4/(30-1)} = 54.4.$$

The quantity needed for a t test with $\alpha = 0.05$ is

$$\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{0.69^2 + 0.53^2} = 0.87.$$

The test statistic is

$$T = \frac{(22.91 - 21.97) - 0}{0.87} = \frac{0.94}{0.87} = 1.08.$$

The p-value for this test (found using a computer) is 0.2850, which is greater than 0.05, so we do not reject H_0. (A quick look at t Table, using $\mathrm{df} = 50$, shows that the p-value is between 0.20 and 0.40.)

If we construct a 95% confidence interval for $\mu_1 - \mu_2$ we get

$$0.94 \pm 2.004 \times 0.87$$

or $(-2.68, 0.80)$. The confidence interval includes zero, which is consistent with not having significant evidence for $H_A : \mu_1 - \mu_2 \neq 0$ in the t test.

Significance Level versus p-value

Students sometimes find it hard to distinguish between significance level α and p-value. For the t test, both α and the p-value are tail areas under Student's t curve. But α is an arbitrary prespecified value; it can be (and should be) chosen before looking at the data. By contrast, the p-value is determined from the data; indeed, giving the p-value is a way of describing the data.

Type I and type II errors

We have seen that α can be interpreted as a probability:

$$\alpha = P(\text{finding significant evidence for } H_A) \text{ if } H_0 \text{ is true.}$$

- Claiming that data provide evidence that significantly supports H_A when H_0 is true is called a **Type I error**. The probability of making a Type I error is α.
- If H_A is true, but we do not observe sufficient evidence to support H_A, then we have made a **Type II error**.

Table 7.3.2 Possible outcomes of testing H_0

		True situation	
		H_0 true	H_A true
OUR DECISION	Lack of significant evidence for H_A	Correct	Type II error
	Significant evidence for H_A	Type I error	Correct

Power

The probability of making a Type II error is denoted by

$$\beta = P(\text{lack of significant evidence for } H_A) \text{ if } H_A \text{ is true.}$$

The chance of not making a Type II error when H_A is true (the chance of having significant evidence for H_A when H_A is true) is called the **power** of a statistical test:

$$\text{power} = 1 - \beta = P(\text{finding significant evidence for } H_A) \text{ if } H_A \text{ is true.}$$

Thus, the power of a t test is a measure of the sensitivity of the test, or the ability of the test procedure to detect a difference between μ_1 and μ_2 when such a difference really does exist.

One-sided t Test

- The t test described in the preceding sections is called a **two-sided** t test because the null hypothesis is rejected if T falls in either tail of the Student's t distribution and the p -value of the data is a two-sided area under Student's t curve.
- In some studies it is apparent from the beginning (before the data are collected) that there is only one reasonable direction of the comparison between μ_1 and μ_2. In such situations it is appropriate to formulate a **one-sided hypothesis test**.
- **left-sided hypothesis tests**:

$$H_0 : \mu_1 - \mu_2 \geq 0 \text{ against } H_A : \mu_1 - \mu_2 < 0.$$

- **Right-sided hypothesis tests**:

$$H_0 : \mu_1 - \mu_2 \leq 0 \text{ against } H_A : \mu_1 - \mu_2 > 0.$$

Left-sided t test

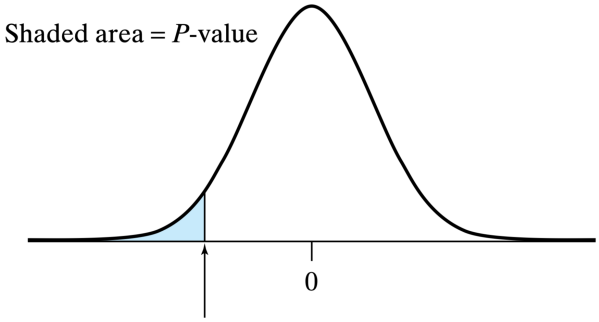
- The test statistic is the same as the two-sided t test:

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{\text{SE}_{\bar{Y}_1 - \bar{Y}_2}}.$$

- The p-value is the probability of getting a t statistic, with ν degrees of freedom, that is **less than** $-T$,

$$P(t_\nu < -T),$$

which is the area under Student's t curve in the left tail beyond $-T$.



Right-sided t test

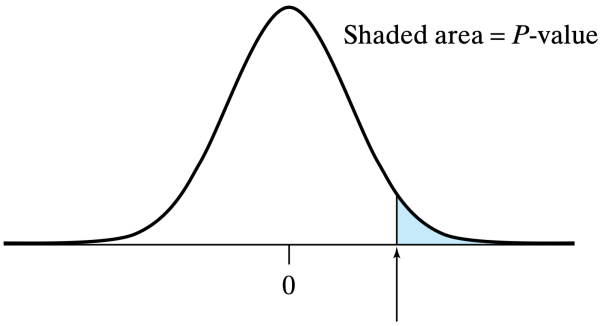
- The test statistic is the same as the two-sided t test:

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2}}.$$

- The p-value is the probability of getting a t statistic, with ν degrees of freedom, that is **greater than** T,

$$P(t_\nu > T),$$

which is the area under Student's t curve in the right tail beyond T.



Example: niacin supplementation

Consider a feeding experiment with lambs. The observation Y will be weight gain in a 2-week trial. Ten animals will receive diet 1, and 10 animals will receive diet 2, where Diet 1 = Standard ration + Niacin and Diet 2 = Standard ration. On biological grounds it is expected that niacin may increase weight gain; there is no reason to suspect that it could possibly decrease weight gain. An appropriate alternative would be

$$H_A: \text{ Niacin is effective in increasing weight gain } (\mu_1 - \mu_2 > 0),$$

which is a right-sided hypothesis test. Suppose that we have $\bar{Y}_1 = 14$ lb, $\bar{Y}_2 = 10$ lb, $\text{SE}_{\bar{Y}_1 - \bar{Y}_2} = 2.2$ lb, and $\nu = 18$ and that we choose the significance level $\alpha = 0.05$.

The test statistic is thus

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{\text{SE}_{\bar{Y}_1 - \bar{Y}_2}} = \frac{(14 - 10) - 0}{2.2} = 1.82.$$

The (right-sided) p-value for the test is the probability of getting a t statistic, with 18 degrees of freedom, that is as large or larger than 1.82. This upper tail probability (found with a computer) is 0.043. If we did not have a computer or graphing calculator available, we could use t Table to bracket the p-value. From t Table, we see that the p-value would be bracketed as follows:

$$0.04 < p\text{-value} < 0.05.$$

Since p-value $< \alpha = 0.05$, we reject H_0 and conclude that there is some evidence that niacin is effective.

Choosing the form of H_A

- When is it legitimate to use a directional H_A, and therefore perform a one-sided test?
- It is legitimate to use a directional alternative H_A only if H_A is formulated **before seeing the data** and there is no scientific interest in results that deviate in a manner opposite to that specified by H_A.
- A researcher who uses a directional alternative when it is not justified pays the price of a doubled risk of Type I error.

Student's t: Conditions and Summary

The t test and confidence interval procedures we have described are appropriate if the following conditions hold:

- Conditions on the design of the study
  - It must be reasonable to regard the data as **random samples** from their respective populations. The populations must be large relative to their sample sizes. The observations within each sample must be **independent**.
  - The two samples must be **independent** of each other.
- Condition on the form of the population distributions
  - The sampling distributions of $\bar{Y}_1$ and $\bar{Y}_2$ must be **(approximately) normal**. This can be achieved via normality of the populations or by appealing to the Central Limit Theorem if the populations are nonnormal but the sample sizes are large.

H_0	H_A	Test statistic	Rejection region	p-value
$\mu_1 - \mu_2 = c$	$\mu_1 - \mu_2 \neq c$	$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - c}{\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2}} \overset{H_0}{\sim} t_\nu$ with $\nu = \frac{(\mathrm{SE}_1^2 + \mathrm{SE}_2^2)^2}{\mathrm{SE}_1^4/(n_1-1) + \mathrm{SE}_2^4/(n_2-1)}$	$\lvert T \rvert > t_\nu(\alpha/2)$	$2 \times P(t_\nu > \lvert T \rvert)$
$\mu_1 - \mu_2 \geq c$	$\mu_1 - \mu_2 < c$		$T < -t_\nu(\alpha)$	$P(t_\nu < T)$
$\mu_1 - \mu_2 \leq c$	$\mu_1 - \mu_2 > c$		$T > t_\nu(\alpha)$	$P(t_\nu > T)$

How are H_0 and H_A chosen

- **Typically the alternative hypothesis is a statement that the researcher is trying to establish**; thus H_A is also referred to as the research hypothesis. For example, if we are testing a new drug against a standard drug, the research hypothesis is that the new drug is better than the standard drug.
- Here are other examples: If we are comparing men and women on some attribute, the usual null hypothesis is that there is no difference, on average, between men and women; if we are studying a measure of biodiversity in two environments, the usual null hypothesis is that the biodiversities of the two environments are equal, on average; if we are studying two diets, the usual null hypothesis is that the diets produce the same average response.