Hypothesis Testing

Tests of hypotheses

Confidence interval:	Form an interval (on the basis of data) of plausible values for a population parameter.
Test of hypothesis:	Answer a yes or no question regarding a population parameter.

Examples:

⟶ Do the two strains have the same average response?

⟶ Is the concentration of substance X in the water supply above the safe limit?

⟶ Does the treatment have an effect?

Example

We have a quantitative assay for the concentration of antibodies against a certain virus in blood from a mouse.

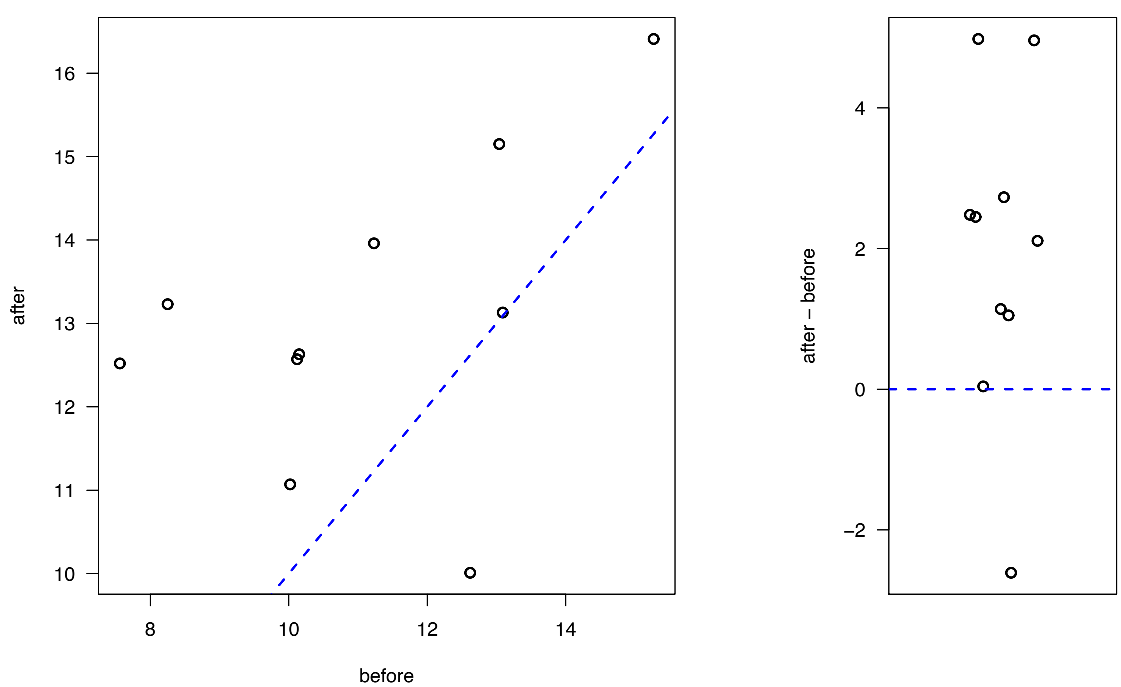
We apply our assay to a set of ten mice before and after the injection of a vaccine. (This is called a "paired" experiment.)

Let X_i denote the differences between the measurements ("after" minus "before") for mouse i.

We imagine that the X_i are independent and identically distributed Normal(μ, σ).

$\longrightarrow$ Does the vaccine have an effect? In other words: Is $\mu \neq 0$?

The data



Hypothesis testing

We consider two hypotheses:

Null hypothesis, H_0: $\mu = 0$ Alternative hypothesis, H_a: $\mu \neq 0$

Type I error: Reject H_0 when it is true (false positive)

Type II error: Fail to reject H_0 when it is false (false negative)

We set things up so that a Type I error is a worse error (and so that we are seeking to prove the alternative hypothesis). We want to control the rate (the significance level, α) of such errors.

$\longrightarrow$ Test statistic: $T = (\overline{X} - 0)/(S/\sqrt{10})$

$\longrightarrow$ We reject H_0 if $|T| > t^\star$, where $t^\star$ is chosen so that $\Pr(\text{Reject } H_0 \mid H_0 \text{ is true}) = \Pr(|T| > t^\star \mid \mu = 0) = \alpha$.

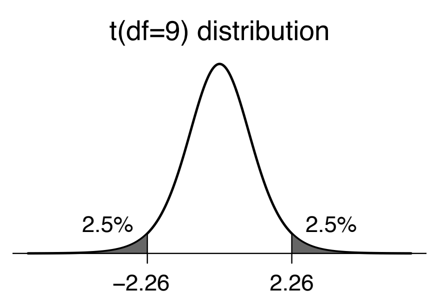
(generally α = 5%)

Example (continued)

Under H_0 (i.e., when $\mu = 0$),

$T = (\overline{X} - 0)/(S/\sqrt{10}) \sim t(df = 9)$

We reject H_0 if $|T| > 2.26$.



As a result, if H_0 is true, there's a 5% chance that you'll reject it!

For the observed data:

$\bar{x} = 1.93$, $s = 2.24$, $n = 10$ $T = (1.93 - 0) / (2.24/\sqrt{10}) = 2.72$

$\longrightarrow$ Thus we reject H_0.

The goal

$\longrightarrow$ We seek to prove the alternative hypothesis.

$\longrightarrow$ We are happy if we reject H_0.

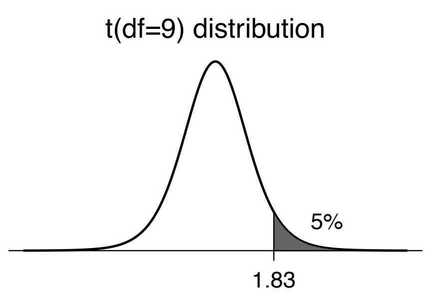
$\longrightarrow$ In the case that we reject H_0, we might say:
Either H_0 is false, or a rare event occurred.

Example (continued)

What if we knew that antibody levels could not decrease in truth?

$\longrightarrow$ We would use a one-tailed (or one-sided) test.

Null hypothesis, H_0: $\mu = 0$ Alternative hypothesis, H_a: $\mu > 0$

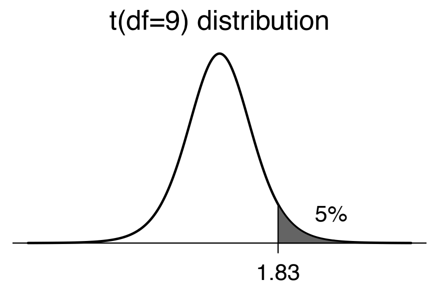


We reject H_0 if $T > 1.83$.

One-tailed vs two-tailed tests

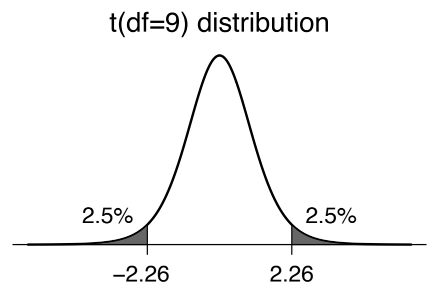
If you are trying to prove that a treatment improves things, you want a one-tailed (or one-sided) test.

You'll reject H_0 only if $T > t^\star$.



If you are just looking for a difference, use a two-tailed (or two-sided) test.

You'll reject H_0 if $T < t^\star$ or $T > t^\star$.



Another example

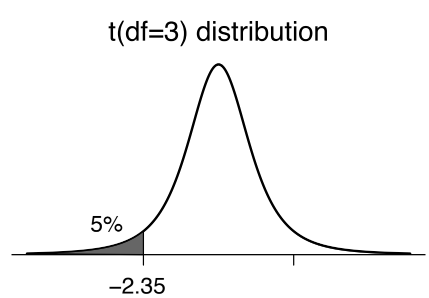
Question: is the concentration of substance X in the water supply above the safe level?

$X_1, X_2, \ldots, X_4 \sim$ iid Normal(μ, σ).

$\longrightarrow$ We want to test H_0: $\mu \geq 6$ (unsafe) versus H_a: $\mu < 6$ (safe).

Test statistic: $T = \dfrac{\overline{X} - 6}{S/\sqrt{4}}$

If we wish to have the significance level α = 5%, the rejection region is $T < t^\star = –2.35$.



P-values

P-value: $\longrightarrow$ the smallest significance level (α) for which you would fail to reject H_0 with the observed data.

$\longrightarrow$ the probability, if H_0 was true, of receiving data as extreme as what was observed.

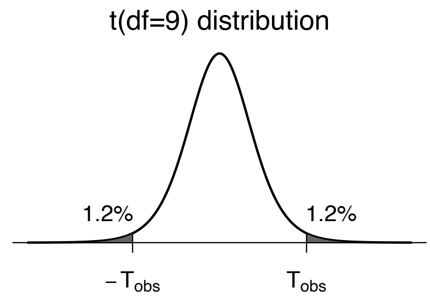
$X_1, \ldots, X_{10} \sim$ iid Normal(μ, σ), H_0: $\mu = 0$; H_a: $\mu \neq 0$.

$\bar{x} = 1.93$; $s = 2.24$

$T_{obs} = \frac{1.93-0}{2.24/\sqrt{10}} = 2.72$

P-value = $\Pr(|T| > T_{obs}) = 2.4\%$.

```
2*pt(-2.72,9)
```



Another example

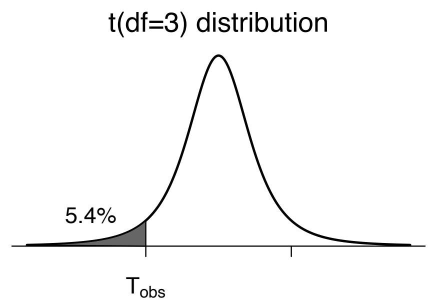
$X_1, \ldots, X_4 \sim$ Normal(μ, σ) H_0: $\mu \geq 6$; $H_a : \mu < 6$.

$\bar{x} = 5.51$; $s = 0.43$

$T_{obs} = \frac{5.51-6}{0.43/\sqrt{4}} = -2.28$

P-value = $\Pr(T < T_{obs} \mid \mu = 6) = 5.4\%$.

```
pt(-2.28, 3)
```



$\longrightarrow$ The P-value quantifies how likely it is to get data as extreme as the data observed, assuming the null hypothesis was true.

Recall: We want to prove the alternative hypothesis (i.e., reject H_0, receive a small P-value)

Hypothesis tests and confidence intervals

$\longrightarrow$ The 95% confidence interval for μ is the set of values, μ_0, such that the null hypothesis $H_0 : \mu = \mu_0$ would not be rejected by a two-sided test with α = 5%.

The 95% CI for μ is the set of plausible values of μ. If a value of μ is plausible, then as a null hypothesis, it would not be rejected.

For example:

9.98 9.87 10.05 10.08 9.99 9.90   assumed to be iid Normal(μ,σ)

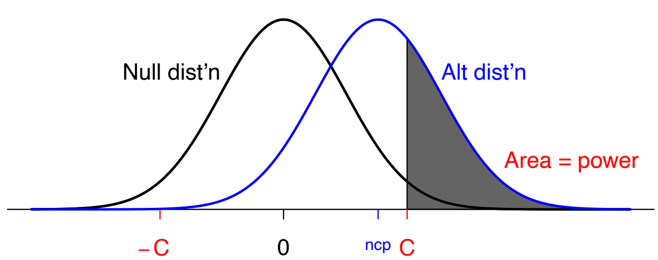
$\bar{x}$ = 9.98; s = 0.082; n = 6; `qt(0.975,5)` = 2.57

The 95% CI for μ is

$$9.98 \pm 2.57 \times 0.082 / \sqrt{6} = 9.98 \pm 0.086 = (9.89, 10.06)$$

Power

The power of a test = Pr(reject H_0 | H_0 is false).



The power depends on:
- The null hypothesis and test statistic
- The sample size
- The true value of μ
- The true value of σ

Why "fail to reject"?

If the data are insufficient to reject H_0, we say,

The data are insufficient to reject H_0.

We shouldn't say, *We have proven H_0.*

- $\longrightarrow$ We may only have low power to detect anything but extreme differences.
- $\longrightarrow$ We control the rate of type I errors ("false positives") at 5% (or whatever), but we may have little or no control over the rate of type II errors.

Testing the difference between two means

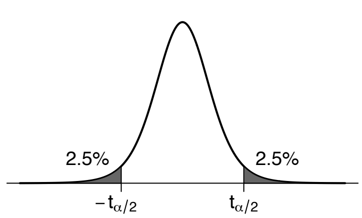
Strain A: $X_1, \ldots, X_n \sim$ iid Normal(μ_A, σ_A)

Strain B: $Y_1, \ldots, Y_m \sim$ iid Normal(μ_B, σ_B)

Test $H_0 : \mu_A = \mu_B$ vs $H_a : \mu_A \neq \mu_B$

Test statistic: $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_A^2}{n} + \frac{S_B^2}{m}}}$

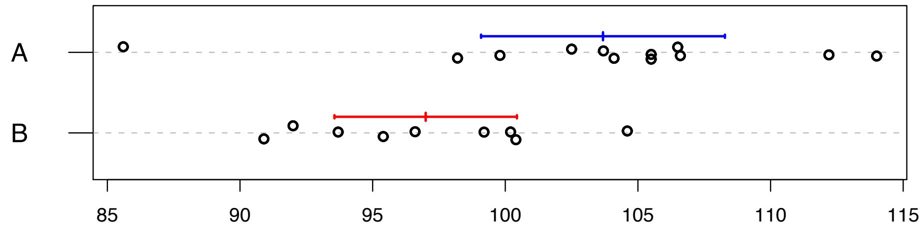
Reject H_0 if $|T| > t_{\alpha/2}$



If H_0 is true, then T follows (approximately) a t distr'n with k d.f.

k according to the nasty formula from a previous lecture.

Example



Strain A: n = 12, sample mean = 103.7, sample SD = 7.2

Strain B: n = 9, sample mean = 97.0, sample SD = 4.5

$$\widehat{\text{SD}}(\overline{X} - \overline{Y}) = \sqrt{\frac{7.2^2}{12} + \frac{4.5^2}{9}} = 2.57$$

T = (103.7 – 97.0)/2.57 = 2.60.

k = . . . = 18.48, so C = 2.10. Thus we reject H_0 at $\alpha = 0.05$.

What to say

When rejecting H_0:

- The difference is statistically significant.
- The observed difference can not reasonably be explained by chance variation.

When failing to reject H_0:

- There is insufficient evidence to conclude that $\mu_A \neq \mu_B$.
- The difference is not statistically significant.
- The observed difference could reasonably be the result of chance variation.

What about a different significance level?

Recall T = 2.60 k = 18.48

If α = 0.10, C = 1.73 $\Longrightarrow$ Reject H_0

If α = 0.05, C = 2.10 $\Longrightarrow$ Reject H_0

If α = 0.01, C = 2.87 $\Longrightarrow$ Fail to reject H_0

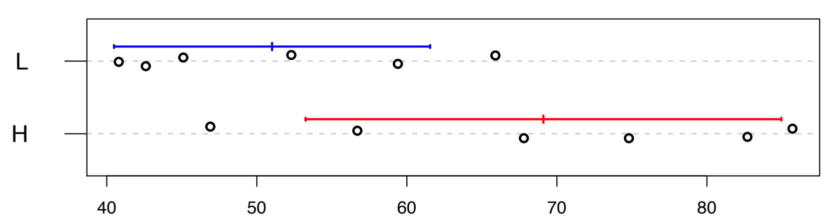
If α = 0.001, C = 3.90 $\Longrightarrow$ Fail to reject H_0

P-value: the smallest α for which you would still reject H_0 with the observed data.

With these data, P = `2*(1-pt(2.60,18.48))` = 0.018.

Another example

Suppose I measure the blood pressure of 6 mice on a low salt diet and 6 mice on a high salt diet. We wish to prove that the high salt diet causes an increase in blood pressure.



We imagine $X_1, \ldots, X_n \sim$ iid Normal(μ_L, σ_L) low salt

$Y_1, \ldots, Y_m \sim$ iid Normal(μ_H, σ_H) high salt

We want to test $H_0 : \mu_L = \mu_H$ versus $H_a : \mu_L < \mu_H$

$\longrightarrow$ Are the data compatible with H_0?

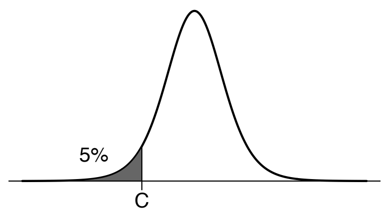
A one-tailed test

Test statistic: $T = \dfrac{\overline{X} - \overline{Y}}{\widehat{\text{SD}}(\overline{X} - \overline{Y})}$

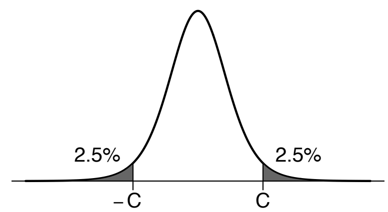
Since we seek to prove that μ_L is smaller than μ_H, only large negative values of the statistic are interesting.

Thus, our rejection region is $T < C$ for some critical value C.

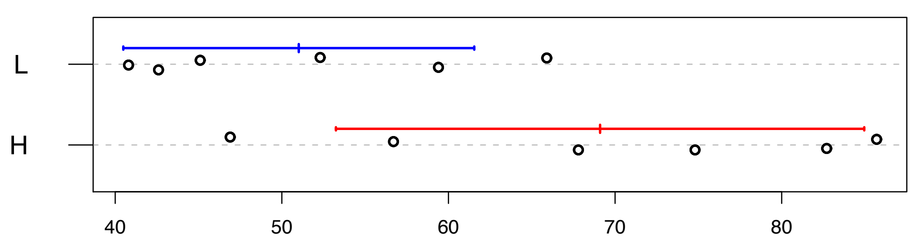
We choose C so that $\Pr(\ T < C \mid \mu_L = \mu_H\ ) = \alpha$.



vs



The example



Low salt: n = 6; sample mean = 51.0, sample SD = 10.0

High salt: n = 6; sample mean = 69.1, sample SD = 15.1

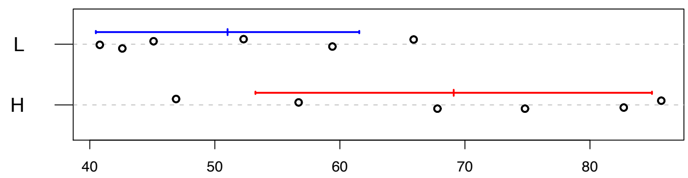
$\bar{x} - \bar{y} = -18.1$ $\widehat{\text{SD}}(\overline{X} - \overline{Y}) = 7.40$ $T = -18.1 / 7.40 = -2.44$

k = 8.69. If $\alpha = 0.05$, then C = –1.84.

Since $T < C$, we reject H_0 and conclude that $\mu_L < \mu_H$.

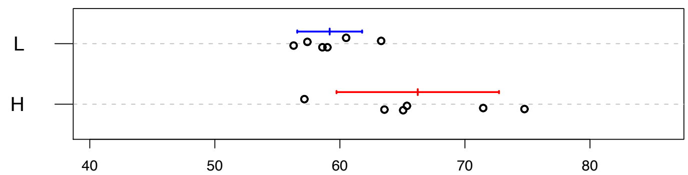
Note: P-value = `pt(-2.44, 8.69)` = 0.019.

Always give a confidence interval!



P = 0.019

95% CI: (–34.9, –1.2)



P = 0.019

95% CI: (–13.6, –0.5)

⟶ Make a statistician happy: draw a picture of the data.

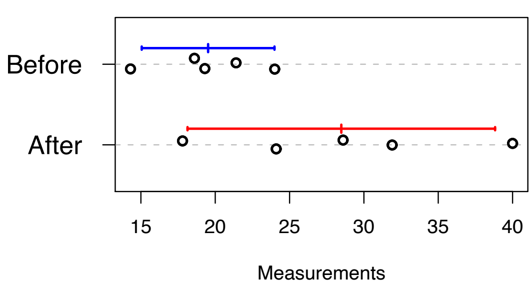
Example

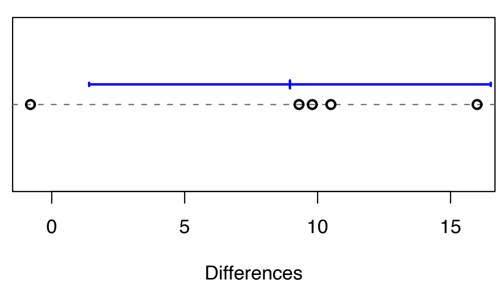
Suppose I do some pre/post measurements.

I make some measurement on each of 5 mice before and after some treatment.

Question: Does the treatment have any effect?

Mouse	1	2	3	4	5
Before	18.6	14.3	21.4	19.3	24.0
After	17.8	24.1	31.9	28.6	40.0





Pre/post example

In this sort of pre/post measurement example, study the differences as a single sample.

Why? The pre/post measurements are likely associated, and as a result one can more precisely learn about the effect of the treatment.

Mouse	1	2	3	4	5
Before	18.6	14.3	21.4	19.3	24.0
After	17.8	24.1	31.9	28.6	40.0
Difference	–0.8	9.8	10.5	9.3	16.0

n = 5; mean difference = 8.96; SD difference = 6.08.

95% CI for underlying mean difference = . . . = (1.4, 16.5)

P-value for test of $\mu_{\text{before}} = \mu_{\text{after}}$: 0.03.

Summary

- Tests of hypotheses $\rightarrow$ answering yes/no questions regarding population parameters.
- There are two kinds of errors:
  - Type I: Reject H_0 when it is true.
  - Type II: Fail to reject H_0 when it is false.
- We seek to reject the null hypothesis.
- If we fail to reject H_0, we do not "accept H_0".
- P-value $\rightarrow$ the probability, if H_0 is true, of obtaining data as extreme as was observed. Pr(data | no effect) rather than Pr(no effect | data).
- Power $\rightarrow$ the probability of rejecting H_0 when it is false.

Was the result important?

- Statistically significant is not the same as important.
- A difference is “statistically significant” if it cannot reasonably be ascribed to chance variation.
- With lots of data, small (and unimportant) differences can be statistically significant.
- With very little data, quite important differences will fail to be significant.
- Always look at the confidence interval as well as the P-value.

Does the difference prove the point?

- A test of significance does not check the design of the study.
- With observational studies or poorly controlled experiments, the proof of statistical significance may not prove what you want.
- Example: consider the tick/deer leg experiment. It may be that ticks are not attracted to deer-gland-substance but rather despise the scent of latex gloves and deer-gland-substance masks it.
- Example: In a study of gene expression, if cancer tissue samples were always processed first, while normal tissue samples were kept on ice, the observed differences might not have to do with normal/cancer as with iced/not iced.
- Don’t forget the science in the cloud of data and statistics.