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.)

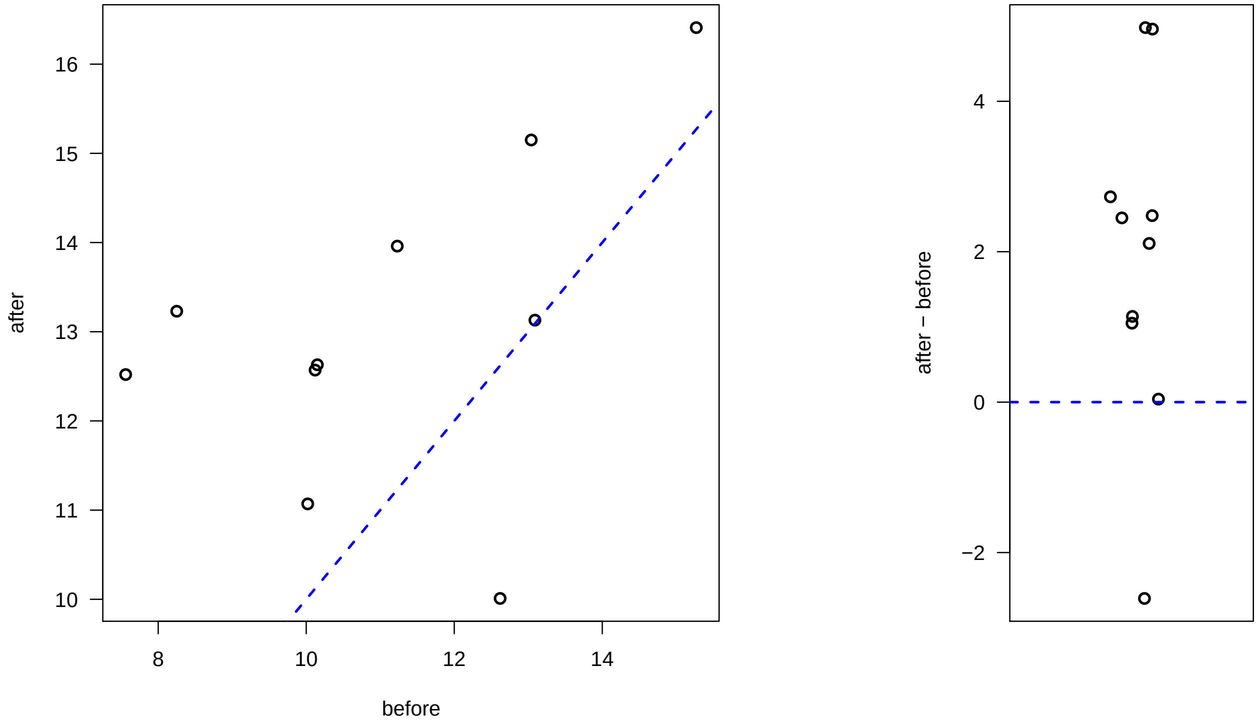
Let X_i denote the differences between the measurements (“after” minus “before”) for mouse i.

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

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: $\mu = 0$ **Alt. 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.

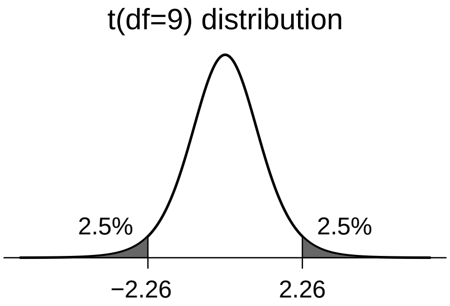
Test statistic: $T = (\bar{X} - 0)/(s/\sqrt{10})$

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 $\mu = 0$),

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

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:

$\bar{X} = 1.93$, $s = 2.24$, $n = 10$

$T = (1.93 - 0) / (2.24/\sqrt{10}) = 2.72$

Thus we reject H_0.

The goal

We seek to prove the alternative hypothesis.

We are happy if we reject H_0.

In the case that we reject H_0, we might say,

> "Either H_0 is false, or a rare event occurred."

Another example

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

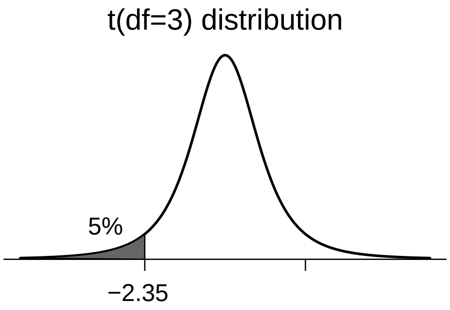
$X_1, X_2, \ldots, X_4 \sim$ iid normal(μ, σ).

Null hyp., H_0: $\mu \geq 6$ (unsafe)

Alt. hyp., H_a: $\mu < 6$ (safe)

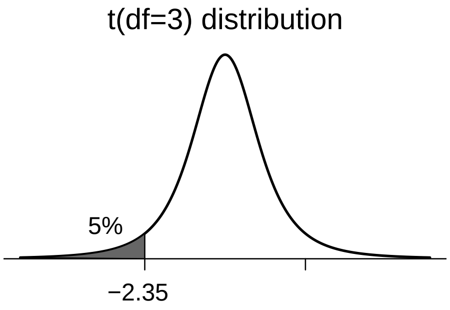
Test statistic: $T = \dfrac{\bar{X} - 6}{s/\sqrt{4}}$

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

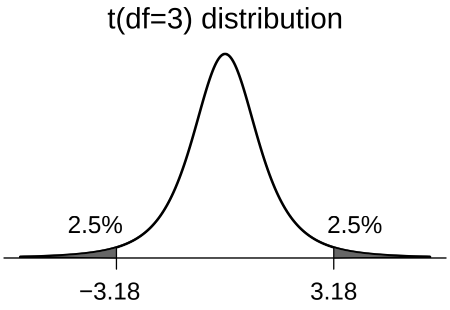


One-tailed vs two-tailed tests

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 $T > t^\star$.)



P-values

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

probability, if H_0 was true, of receiving data as extreme as what was observed.

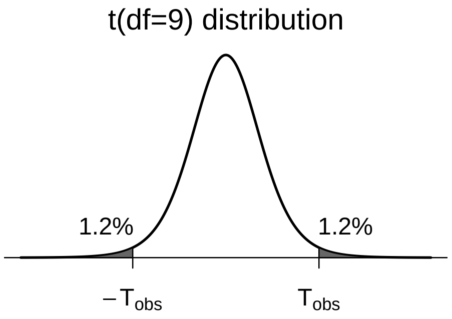
$X_1, \ldots, X_{10} \sim$ iid normal(μ, σ) H_0: $\mu = 0$; H_a: $\mu \neq 0$.

Observe: $\bar{X} = 1.93$; s = 2.24

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

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

= `2*pt(-2.72,9)`
= 2.4%.



Another example

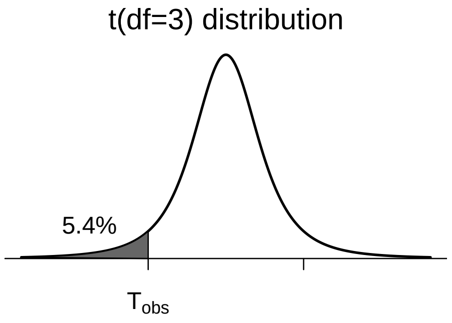
$X_1, \ldots, X_4 \sim$ normal(μ, σ) H_0: $\mu \geq 6$; $H_a : \mu < 6$.

Observe: $\bar{X} = 5.51$; s = 0.43

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

P-value = $\Pr(T < T_{obs} \mid \mu = 4)$

= `pt(-2.28, 3)` = 5.4%.



The P-value is (roughly) a measure of evidence against the null hypothesis.

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

Hypothesis tests and confidence intervals

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 iid normal(μ,σ).)

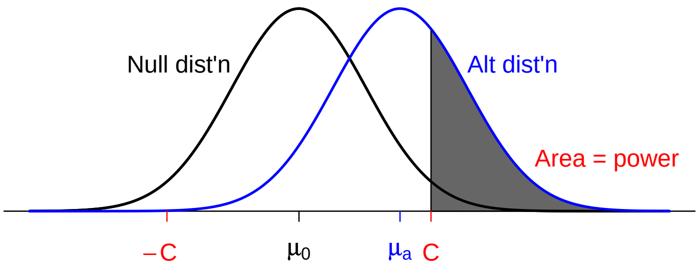
$\bar{X}$ = 9.98; s = 0.082; n = 6　　`qt(0.975,5)` = 2.57

95% CI for μ = 9.98 $\pm$ 2.57 $\cdot$ 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.”

Why? We have very low power to detect similar alternatives. We may have low power to detect anything but extreme differences.

We control the rate of type I errors (“false positives”) at 5% (or whatever), but we have little or no control over the rate of type II errors.

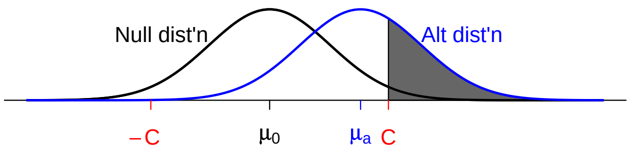
The effect of sample size

Let $X_1, \ldots, X_n$ be iid normal(μ, σ).

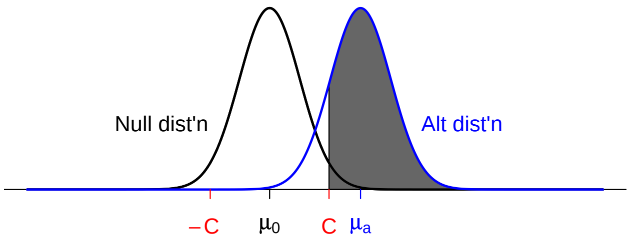
We wish to test $H_0: \mu = \mu_0$ vs $H_a: \mu \neq \mu_0$.

Imagine $\mu = \mu_a$.

n = 4



n = 16



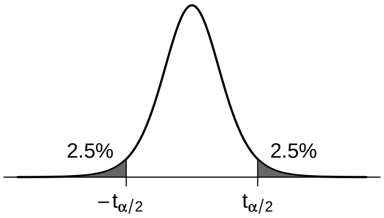
Testing the difference between two means

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: $T = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_A^2}{n} + \frac{s_B^2}{m}}}$



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 the last lecture)

Example

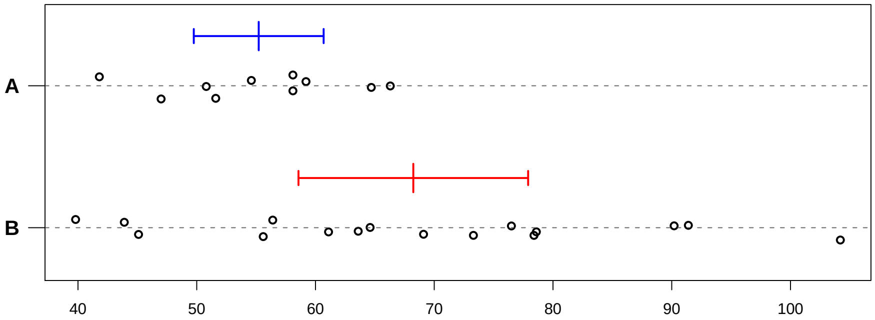
Strain A: n=10; $\bar{X}$=55.2; s_A=7.64

Strain B: m=16; $\bar{Y}$=68.2; s_B=18.14

$\bar{X} - \bar{Y}$ = –13.0 $\qquad \widehat{SD}(\bar{X} - \bar{Y}) = \sqrt{7.64^2/10 + 18.1^2/16} = 5.14$

T = –13.0 / 5.14 = –2.53 $\qquad$ k = . . . = 21.8

P-value = `2*pt(-2.53,21.8)` = 1.9%.



Cite CIs as well as P-values

Example 1: 95% CI for $\mu_A - \mu_B$ = (–23.7, –2.4)

P-value for test of $\mu_A = \mu_B$ = 1.9%.

Example 2: 95% CI for $\mu_A - \mu_B$ = (–1.84, –0.16)

P-value for test of $\mu_A = \mu_B$ = 2.2%.

The P-value is just one number, and only says so much.

The confidence interval contains much more information.

Summary

- Tests of hypotheses = answering yes/no questions regarding population parameters
- 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 don't "accept H_0."
- P-value = 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 = probability of rejecting H_0 when it is false.
- Always look at the confidence interval as well as the P-value