Testing the difference between two means

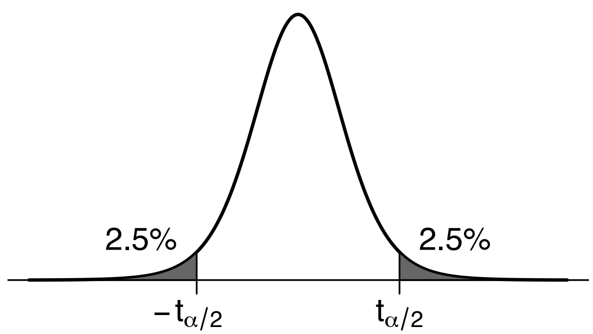
Strain A: $X_1, \ldots, X_n \sim$ iid Normal(μ_A, σ_A)

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

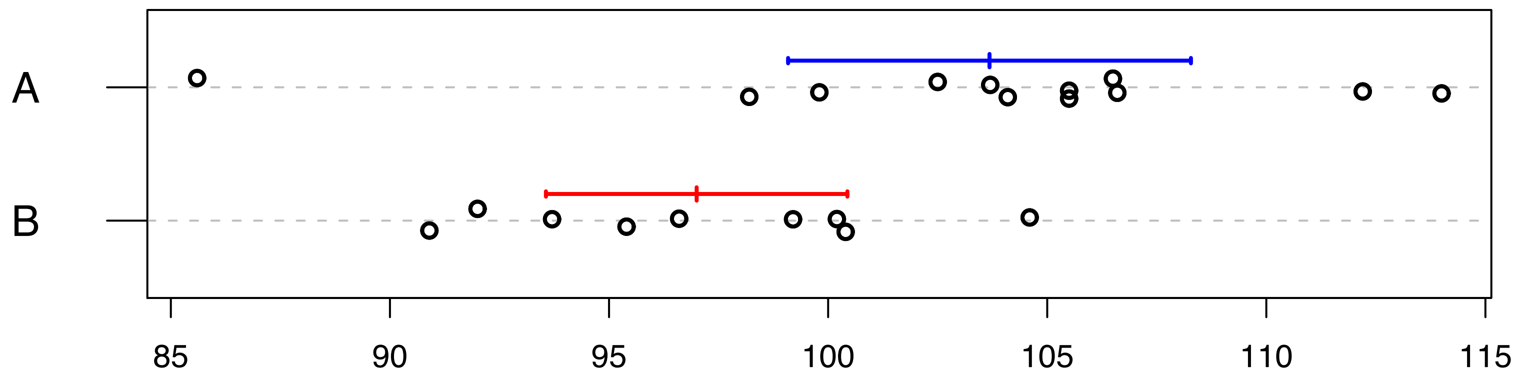
Reject H_0 if $|T| > t_{\alpha/2}$



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

$$\widehat{\text{SD}}(\overline{X} - \overline{Y}) = \sqrt{\frac{7.2^2}{12} + \frac{4.5^2}{9}} = 1.80$$

T = (103.7 – 97.0)/1.80 = 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

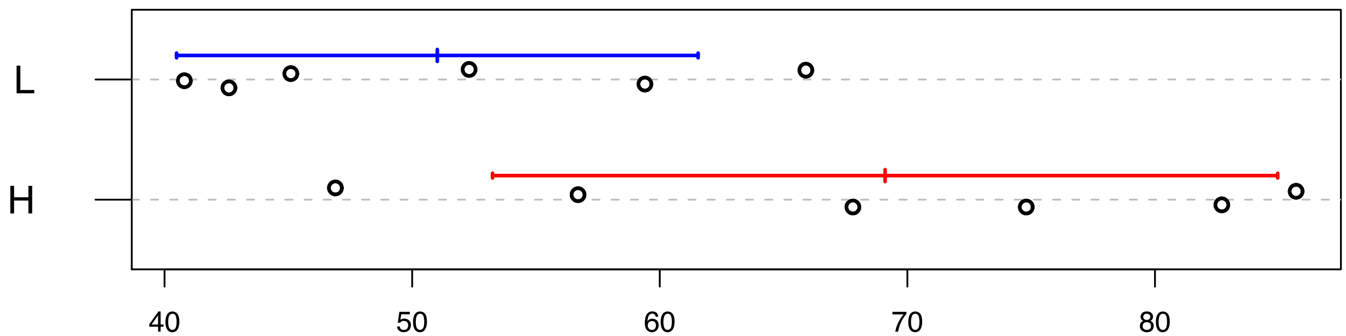
If α = 0.001, C = 3.90 $\Longrightarrow$ Fail to reject H_0

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 $X_1, \ldots, X_n \sim \text{iid Normal}(\mu_L, \sigma_L)$ low salt

$Y_1, \ldots, Y_m \sim \text{iid Normal}(\mu_H, \sigma_H)$ 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?

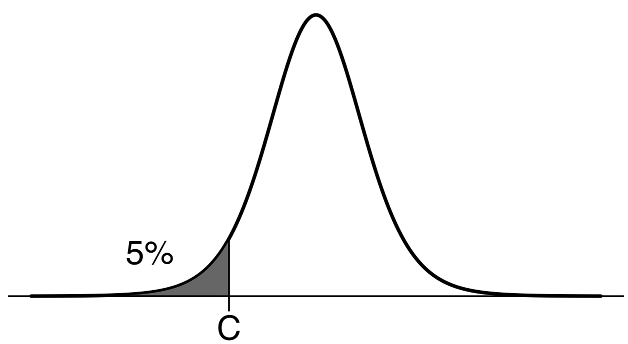
A one-tailed test

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

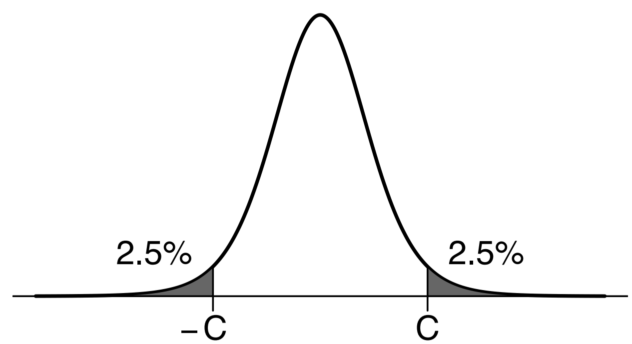
Since we seek to prove that μ_L is smaller than μ_H, only large negative values of the statistic are interesting.

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

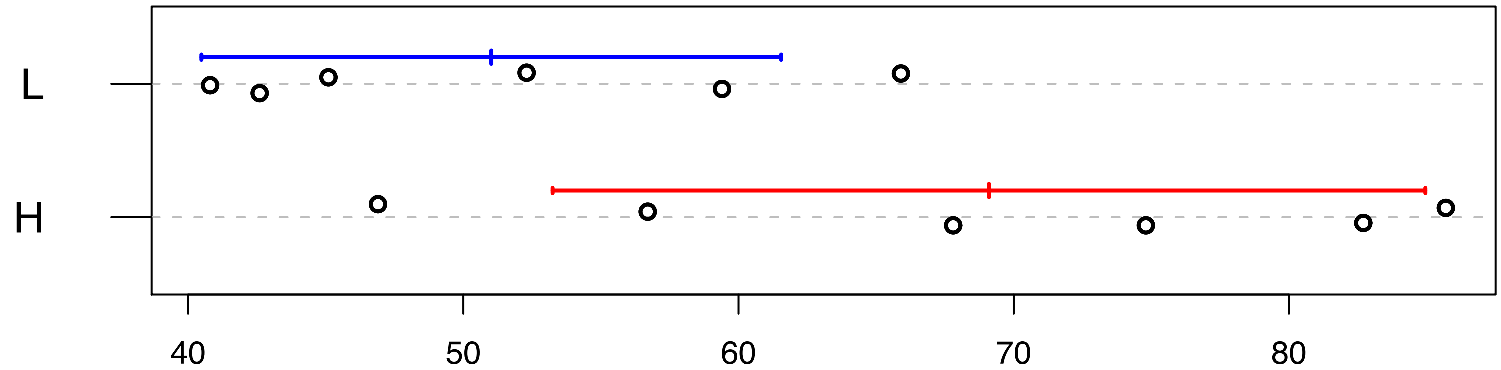
We choose C so that $\Pr(\, T < C \mid \mu_L = \mu_H \,) = \alpha$.



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

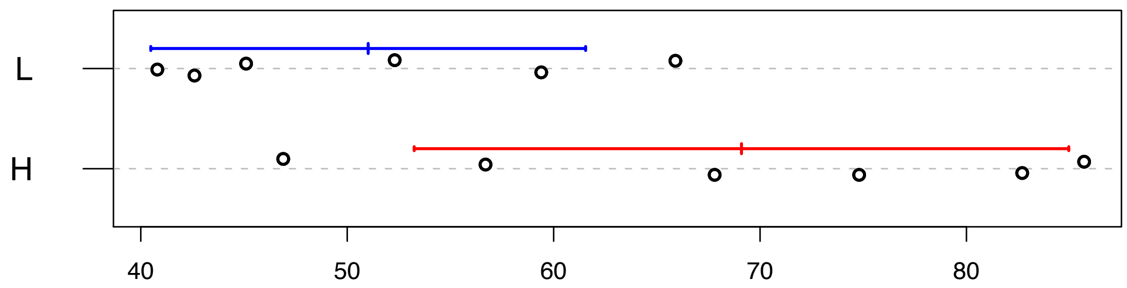
$\bar{x} - \bar{y} = -18.1$ $\widehat{\text{SD}}(\overline{X} - \overline{Y}) = 7.40$ T = –18.1 / 7.40 = –2.44

k = 8.69. If α = 0.05, then C = –1.84.

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

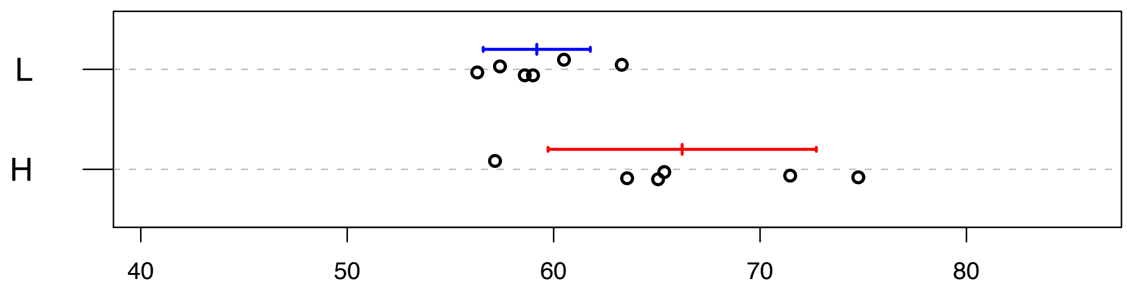
Note: P-value = `pt(-2.44, 8.69)` = 0.019.

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.

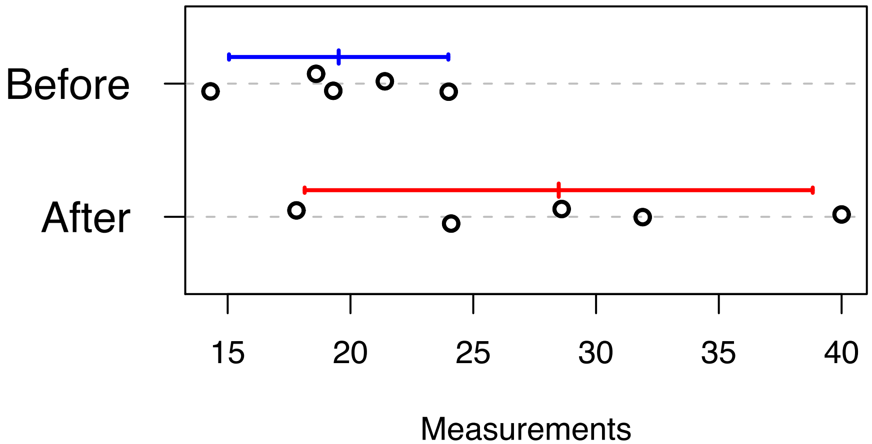
Example

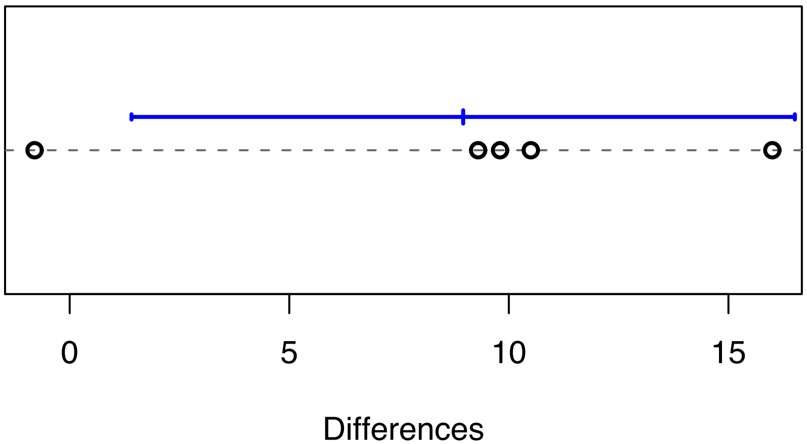
Suppose I do some pre/post measurements.

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.