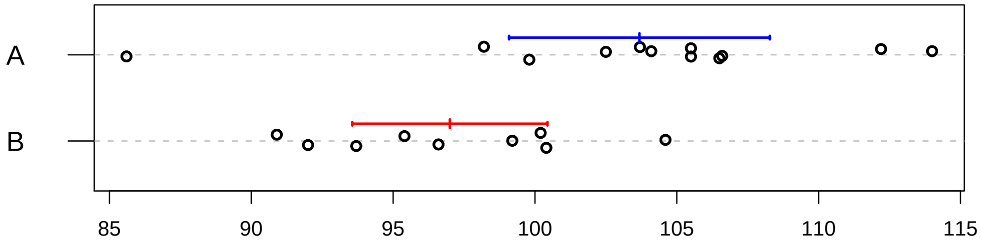
Hypothesis testing



Question: Do the two strains have the same mean?

We imagine $X_1, \ldots, X_n \sim$ iid normal(μ_A,σ_A)
$Y_1, \ldots, Y_m \sim$ iid normal(μ_B,σ_B)

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

Question: Are the data compatible with H_0?

The two errors

Type I ("false positive")

Conclude $\mu_A \neq \mu_B$ when they are actually equal.

Type II ("false negative")

Conclude $\mu_A = \mu_B$ when they are actually different.

We set things up so that the former is the worse error (which we wish to really avoid).

We avoid the latter by never really concluding $\mu_A = \mu_B$. Rather, we say, "We have insufficient evidence to conclude $\mu_A \neq \mu_B$."

Test statistic

In order to determine whether the data are compatible with H_0, we form a summary statistic, for which large values indicate evidence for a departure from the null hypothesis $\mu_A = \mu_B$.

The statistic to use depends on

(a) the types of parameters in question

(b) the form of the data

(c) our assumptions about the process generating the data

In the above example, we'd use $T = \dfrac{\bar{X} - \bar{Y}}{\widehat{\mathrm{SD}}(\bar{X} - \bar{Y})}$

Rejection rule: Reject H_0 if $|T| > C$, for some "critical value," C.

Significance level

We seek to avoid making a type I error (rejecting H_0 when it is true).

We choose our rejection rule so that $\Pr(\text{reject } H_0 \mid H_0 \text{ is true}) = \alpha$.

Generally, we use $\alpha = 0.05$.

But we could reasonably use the more stringent criterion $\alpha = 0.01$ or the less stringent one $\alpha = 0.10$.

I strongly advise *against* any hard-and-fast rule!

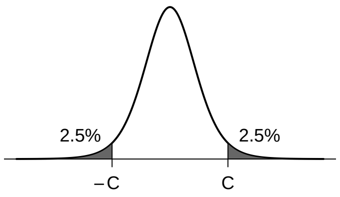
Null distribution

Crucial to the choice of the critical value (and thus for determining whether we may conclude $\mu_A \neq \mu_B$) is the null distribution of the test statistic.

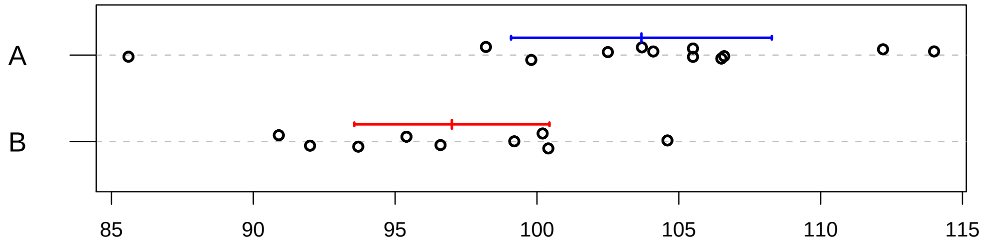
If H_0 is true (that is, if $\mu_A = \mu_B$), the above statistic, T, approximately follows a t distribution with k degrees of freedom

where k = (complicated formula suppressed).

The critical value for the test: C = the 97.5 percentile of this distribution, since then Pr(|T| $> C \mid \mu_A = \mu_B$) = 5%.



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

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

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 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 if we used 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 this data, P = `2 * (1 - pt(2.60,18.48))` = 0.018.

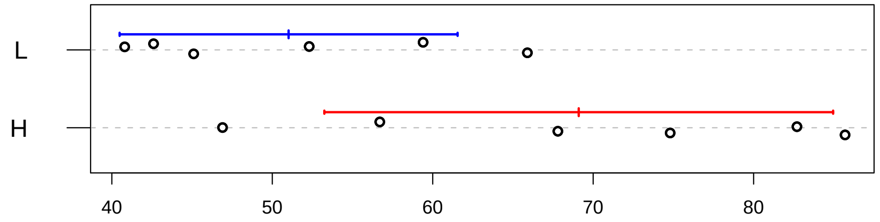
P-values

- P-values are a function of the data. (They are random, like data.)
- P-values measure the strength of evidence against H_0. (Take this with a grain of salt.)
- Small p-values indicate evidence against H_0.
- P = probability of getting this sort of extreme data, if the observed difference were just due to chance variation.
- **NOT** the probability that the observed difference is due to chance.
- Note that P=0.048 is essentially the same as P=0.053.

Another example

Suppose I measure the blood pressure of 6 mice on a low salt diet and 6 mice on a high salt diet.

I 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]

$H_0 : \mu_L = \mu_H$ $H_a : \mu_L < \mu_H$

Question: Are the data compatible with H_0?

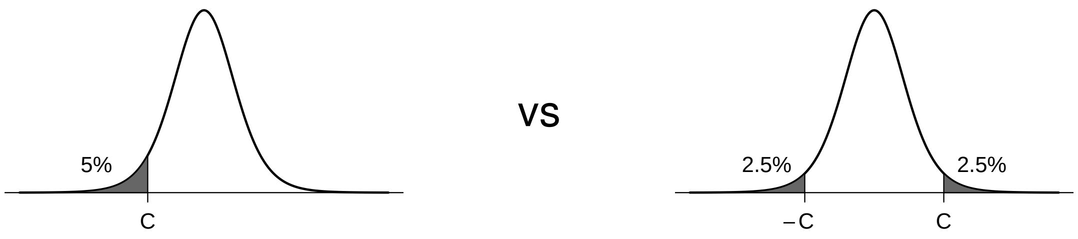
A one-tailed test

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

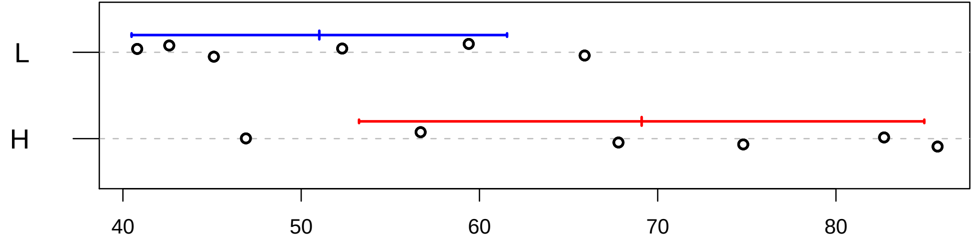
Since we seek to prove that $\mu_H > \mu_L$, only large negative values of the statistic are interesting.

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

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



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

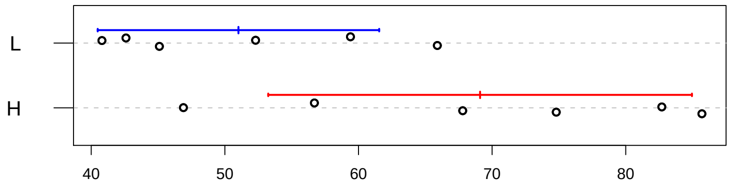
$\bar{X} - \bar{Y} = -18.1$ $\widehat{\mathrm{SD}}(\bar{X} - \bar{Y}) = 7.40$ $T = -18.1 / 7.40 = -2.44$

k = 8.69. If α = 0.05, C = –1.84.

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

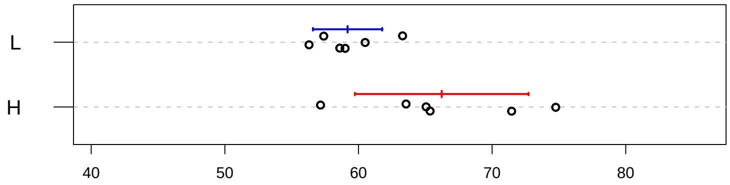
Note: P-value = `pt(-2.44, 8.69)` = 0.019.

Always give a confidence interval!



P = 0.019

95% CI: (–34.9, –1.2)

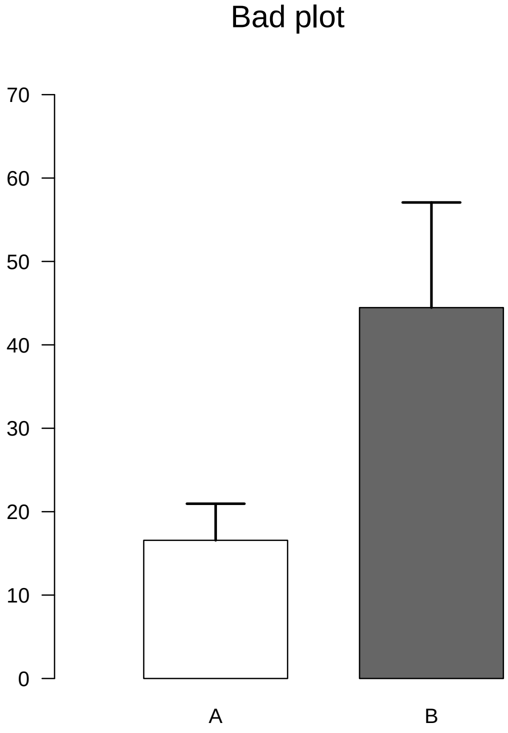


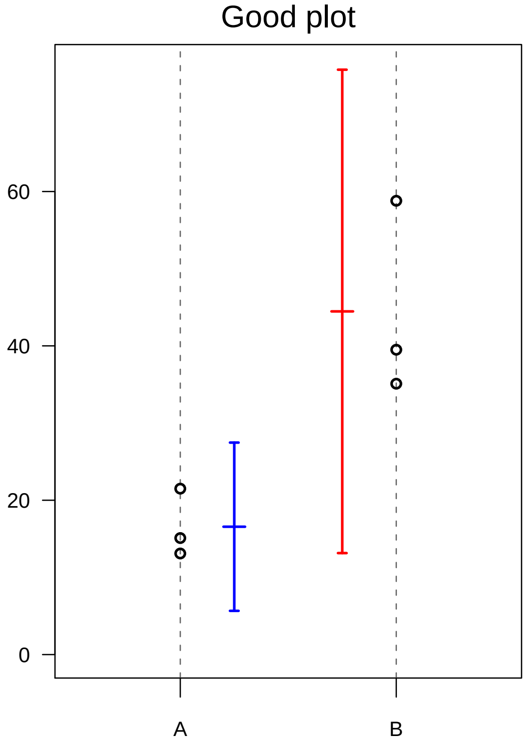
P = 0.019

95% CI: (–13.6, –0.5)

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

Good plot, bad plot


Bad plot


Good plot

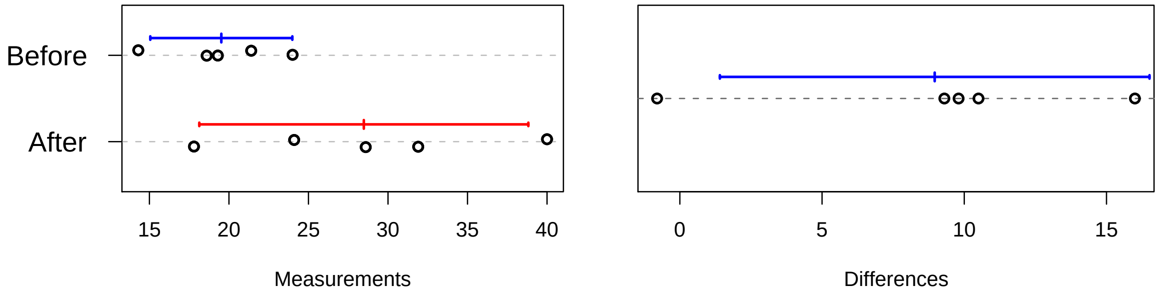
Example

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.