Statistical tests

Theory

Sarah Kaspar

Biostatistical Basics 2021

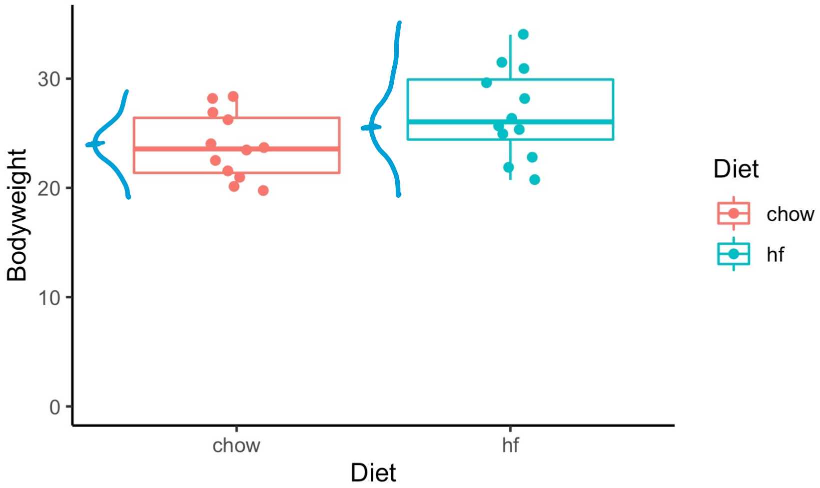
EMBL

Goals for this lecture:
- understand the common principles behind statistical tests
- learn how sampling distribution impacts your choice of test
- learn to spot common pitfalls

Content:
- binomial test
- t-test
- alternatives to t-test

Sources:
- These slides are based on a lecture on testing by Bernd Klaus and Wolfgang Huber (2018)

Example: Mice weights



Question: Is there a difference in weight between mice with control vs. high-fat diet?

Problem: The difference in means could be by chance:

- we only have a sample of mice for each diet
- there is variation in the weights

Knowing the rules for randomness / variation will tell us how likely it is to see this difference by chance.

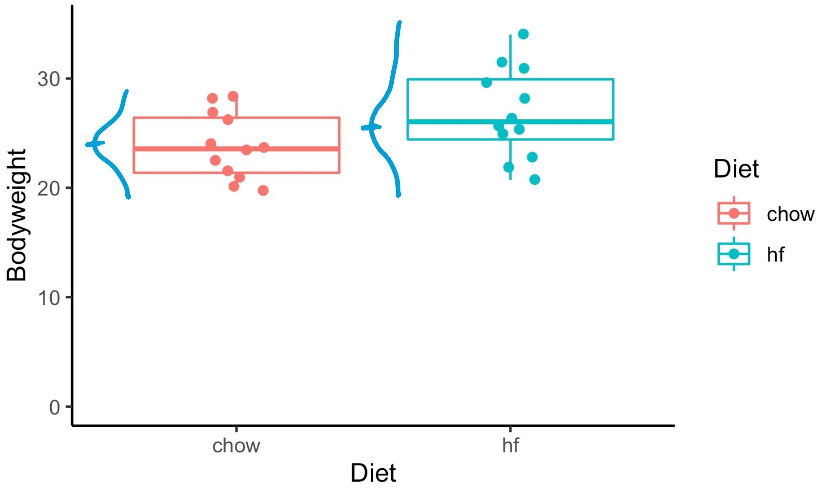
Statistical model:

$$weight = \ diet + \underline{residuals}$$

group means — *follow a statistical distribution*

data from Winzell and Ahrén (2004)

Null and alternative hypothesis



Null hypothesis (H0): There is no difference between the two diet groups.

Alternative hypothesis (H1): There is a difference between the two diet groups.

We reject the null hypothesis when – assuming it was true – it would be very unlikely to observe a difference as extreme as in our data just by chance.

<u>Alternative model:</u>

$$weight = \ diet + residuals$$

group means

<u>Null model:</u>

$$weight = grand\ mean + \ residuals$$

data from Winzell and Ahrén (2004)

Steps of hypothesis testing

EMBL

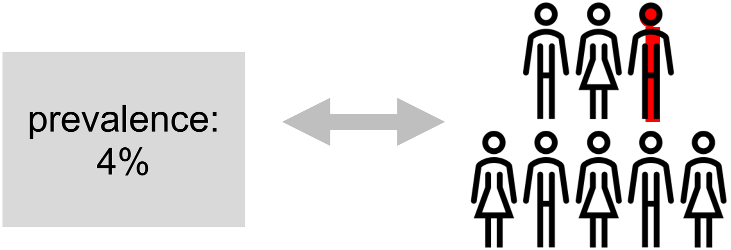
1. Set up a null model / null hypothesis
2. collect data
3. calculate the probability of the data in the null model
4. decide: Reject the null model, if the above probability is too small

Example: disease prevalence

EMBL

Scenario:

- Known prevalence: 4%
- 100 test persons with a precondition, 9 of them have the disease



Hypotheses:

- **H_0**: The prevalence in the test group is also 4% ("boring" outcome, we want to collect evidence against it)
- **H_A**: The prevalence in the test group differs from 4%.
- **Null model**: binomial distribution with n=100, p=0.04

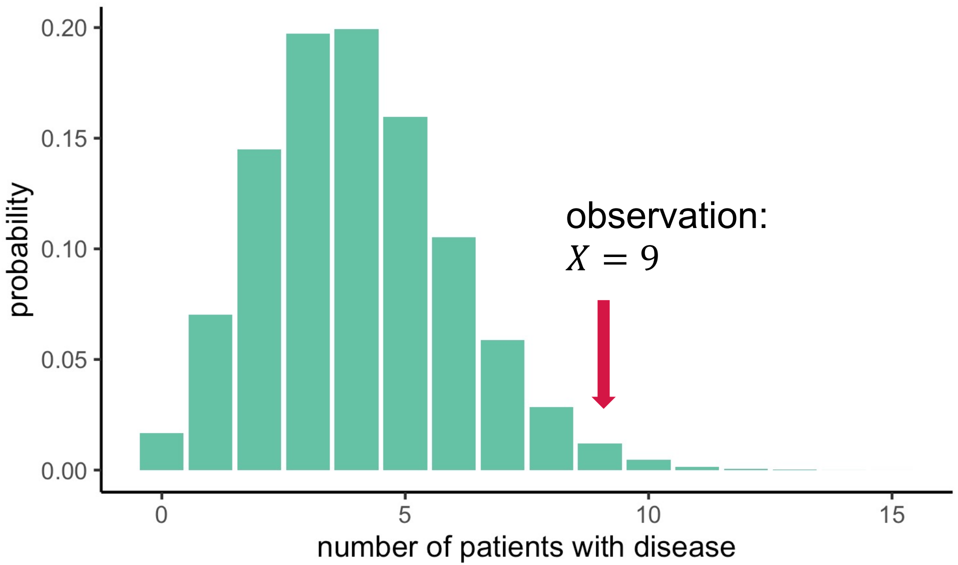
Example: disease prevalence

What is the probability of seeing an event at least as extreme as the observed one under H_0?

- The probability of observing 9 or more persons with disease is rather unlikely: $P(X \geq 9) = 0.019$

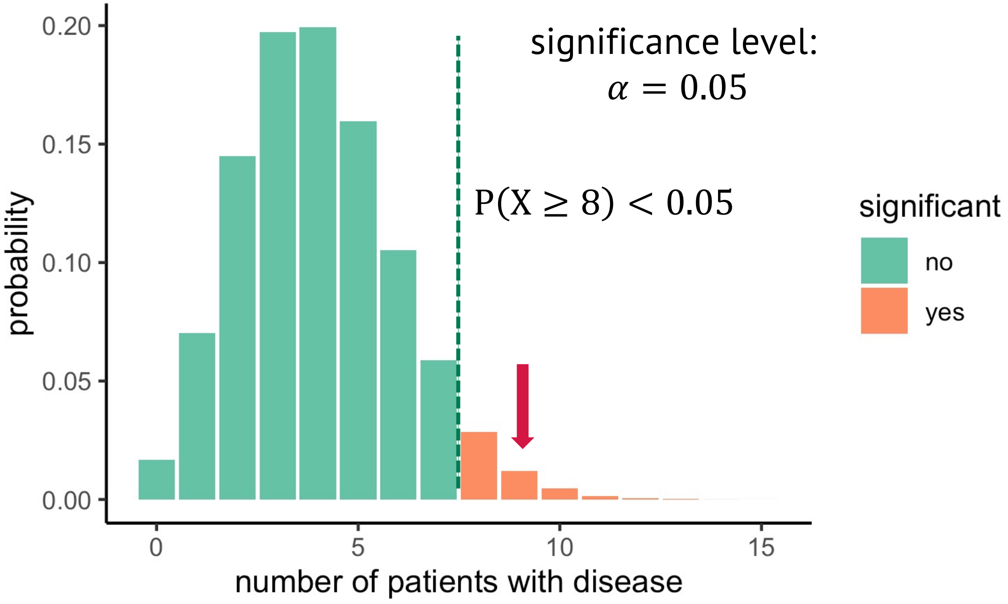
- The null hypothesis is likely false.

Null distribution



Example: disease prevalence

We usually call the result significant, if the pobability under H_0 is smaller than 5 %.



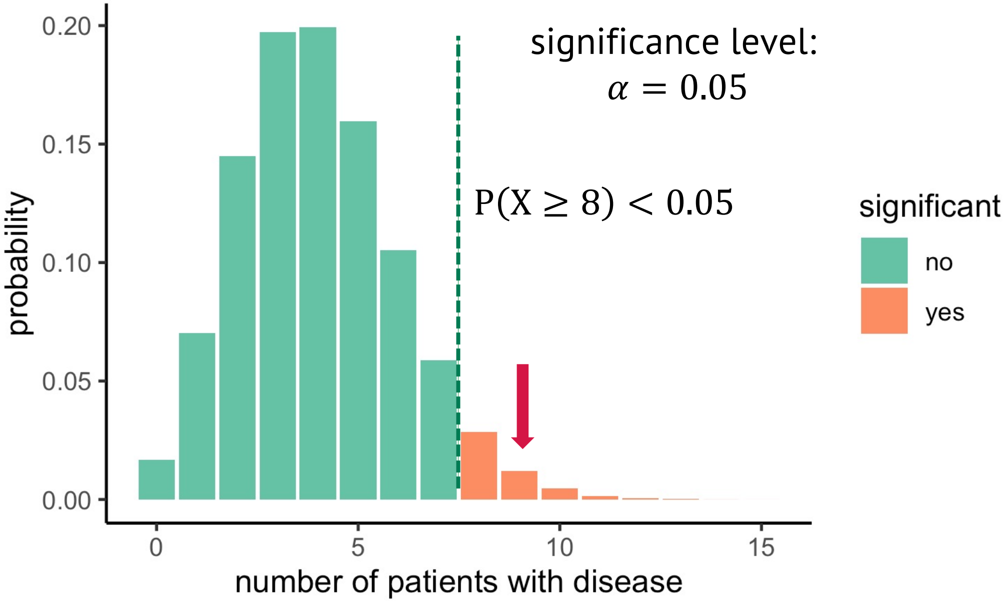
Question

What was wrong (conceptionally) about this test?

Example: disease prevalence

What we did was a one-sided test.

One-sided: look only in one direction:
$H_A: p > 0.04$ or
$H_A: p < 0.04$

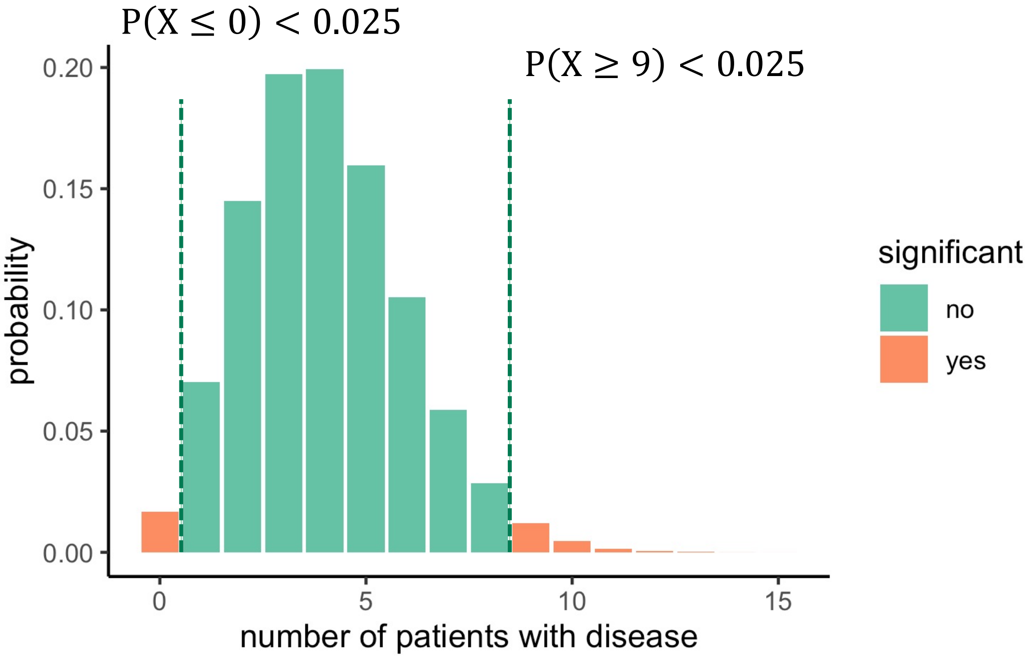


Example: disease prevalence

Which numbers of test persons are very unlikely / extreme, assuming H_0 is true?

Two-sided: look in both directions
H_A: p ≠ 0.04

- Observing less than one person with disease is very unlikely: $P(X = 0) = 0.017$
- observing more than 8 persons with disease is also very unlikely: $P(X > 8) = 0.019$



Excursion: Data snooping

What was wrong about the one-sided test?

- We decided on the direction to look at *after* collecting the data
- The significant level α is not true anymore!
- There is a 50% chance that your sample is higher or lower than the expected value → the true α is 0.1

Question: When is a one-sided test OK?

Errors in hypothesis testing

	Not rejected	rejected
H_0 true	true negative	false positive type I error
H_0 false	false negative type II error	true positive

we increase **type I error** by:

- multiple comparisons
- data snooping
- certain violations of assumptions (e.g. independence)

we try to avoid **type II error** by choosing methods with a high power

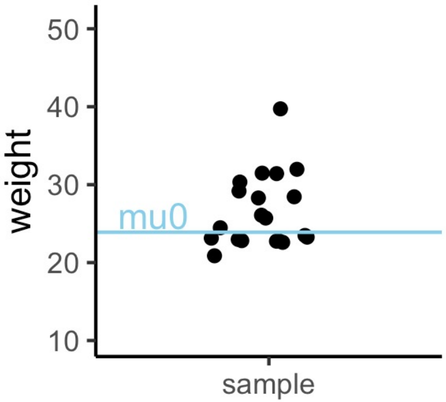
Great page:
https://en.wikipedia.org/wiki/Confusion_matrix

Mouse weights

Compare a sample mean to μ_0

Null hypothesis: The weight in the sample is μ_0.

Alternative hypothesis: The weight in the sample is different from μ_0.



data from Winzell and Ahrén (2004)

One-sample t-test

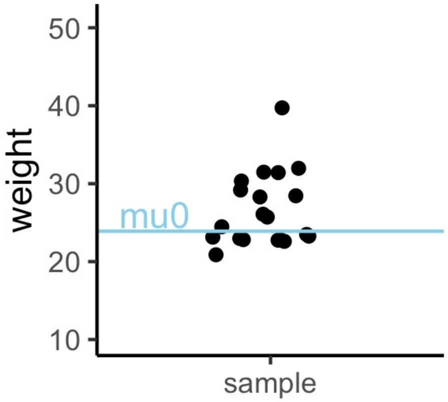
Compare a sample mean to μ_0

The t statistic:

difference between sample mean and μ_0

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

standard error of the mean



data from Winzell and Ahrén (2004)

Why is t a useful statistic?

EMBL

difference between sample mean and μ_0

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

standard error of the mean

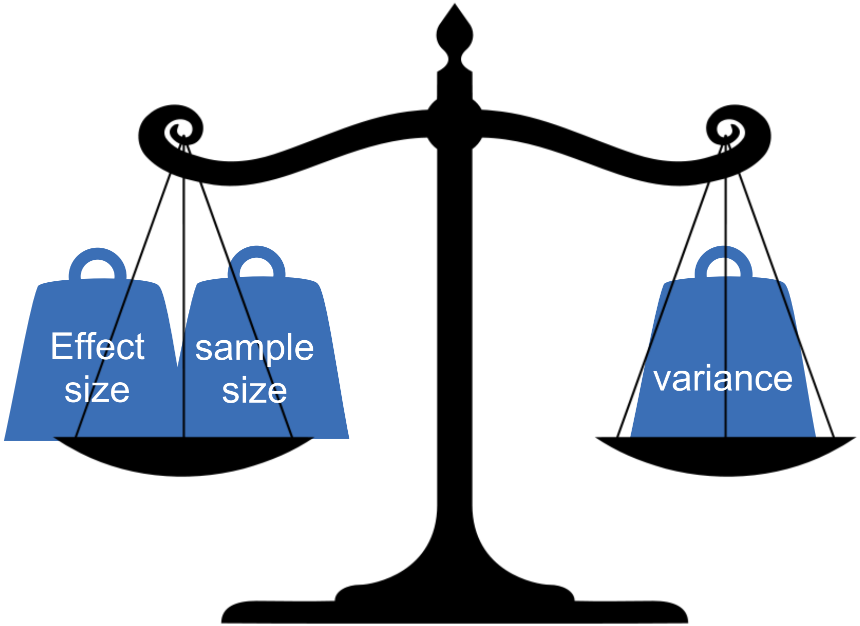


Image: https://de.m.wikipedia.org/wiki/Datei:Balanced_scale_of_Justice.svg

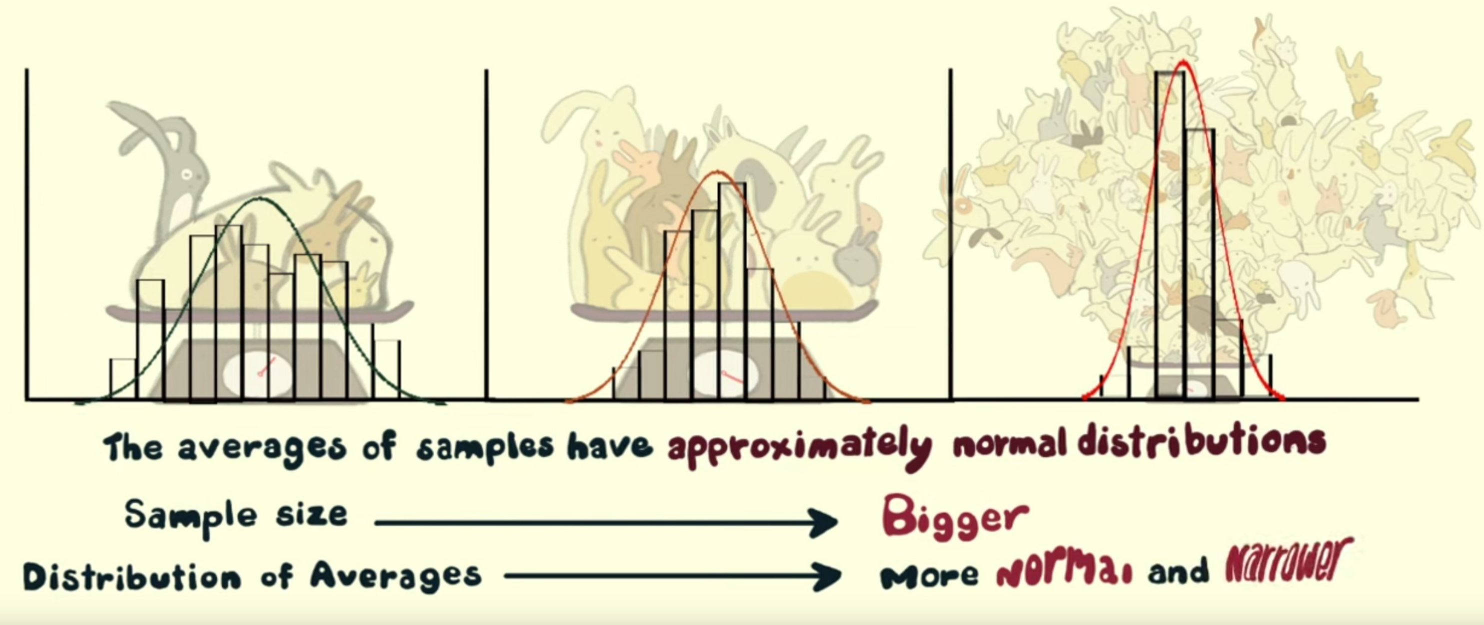
What is the null distribution of t?

In order to calculate a p-value, we have to find the null distribution of t.
→The distribution that t follows when the two groups are equal.

Two explanations

- Using the central limit theorem
- Through simulation (→ demonstration in R)

Central limit theorem



https://www.youtube.com/watch?v=jvoxEYmQHNM

Central limit theorem

EMBL

The sum of random variables tends towards a normal distribution with increasing N.

<u>For our example:</u>

The more mice we sample (N), the more the distribution of the sample average will look like a Gaussian distribution with

- mean = the true average weight of mice
- standard deviation = standard error of the mean

<u>Standard error of the mean:</u>
quantifies how well a sample estimates the mean

$$SE = \sigma/\sqrt{n}$$

more measurements lead to better approximation of the mean

One-sample t-test

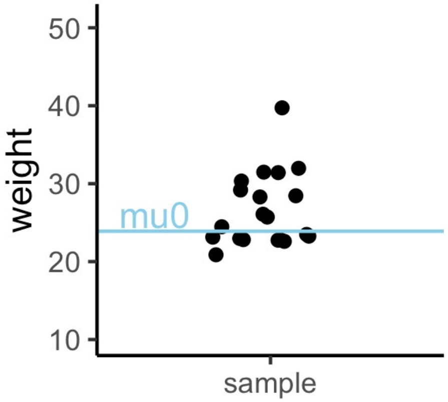
Compare a sample mean to μ_0

The t statistic:

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} = \frac{2.7}{1.04} = 2.57$$

Central limit theorem: If H_0 is true, then t follows a normal distribution with mean 0 and sd=1.

But: CLT is only true for large sample sizes!



data from Winzell and Ahrén (2004)

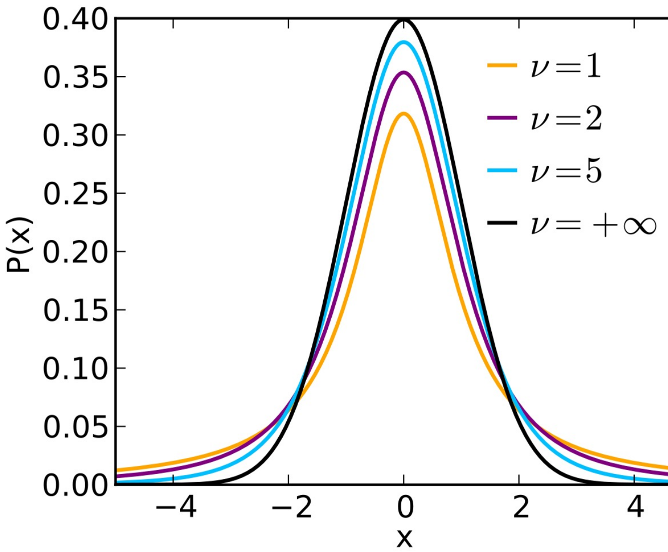
The t-distribution

Applicable to small sample sizes

difference between sample mean and μ_0

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

standard error of the mean



Degrees of freedom: the number of values in the calculation of t that are free to vary

If H_0 is correct, the t statistic follows a t distribution with $n-1$ degrees of freedom.

Image from https://en.wikipedia.org

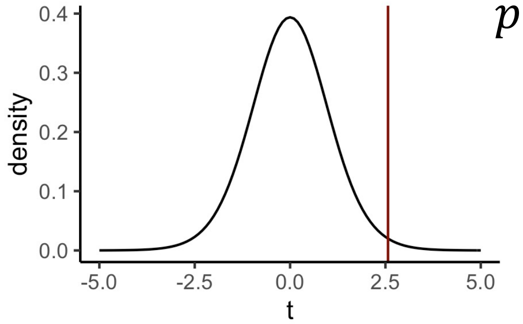
One-sample t-test

Compare a sample mean to μ_0

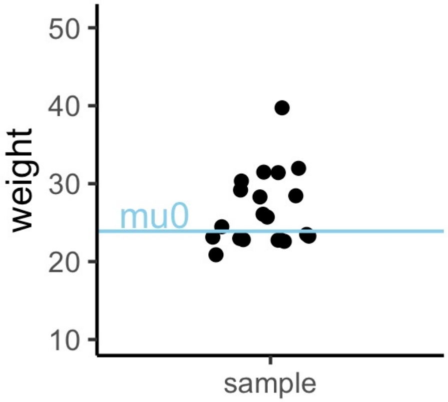
The t statistic:

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} = \frac{2.7}{1.04} = 2.57$$

P-value:



$$p = 0.02$$



data from Winzell and Ahrén (2004)

Two-sample t-test

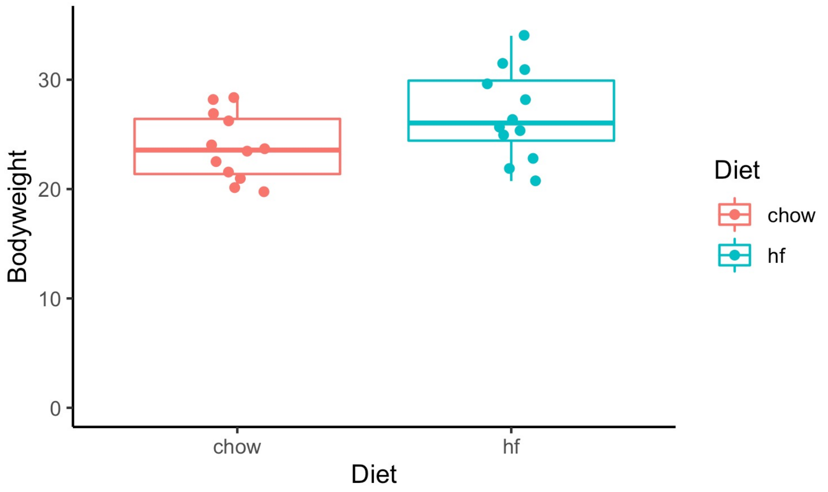
difference between the two sample means

$$t = \frac{\bar{x} - \bar{y}}{SE}$$

standard error

$$SE = \sqrt{\frac{\hat{\sigma}_x^2 + \hat{\sigma}_y^2}{n}}$$

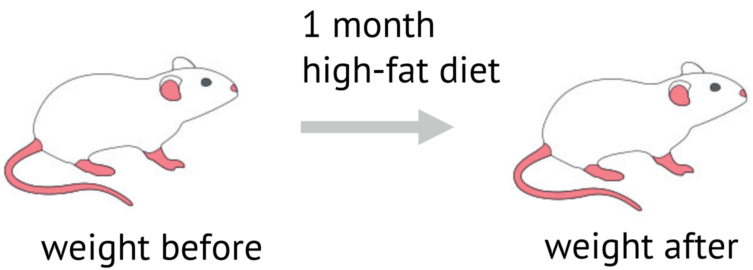
for equal variances and sample size

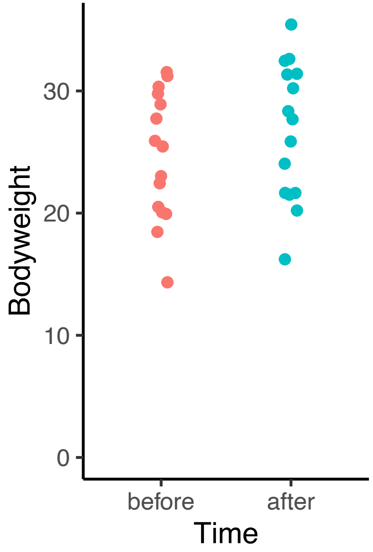


If H_0 is correct, the t statistic follows a t distribution with n_x+n_y-2 degrees of freedom.

Paired t-test

Example: The weight of 15 mice is measured before and after 1 month of high-fat diet.



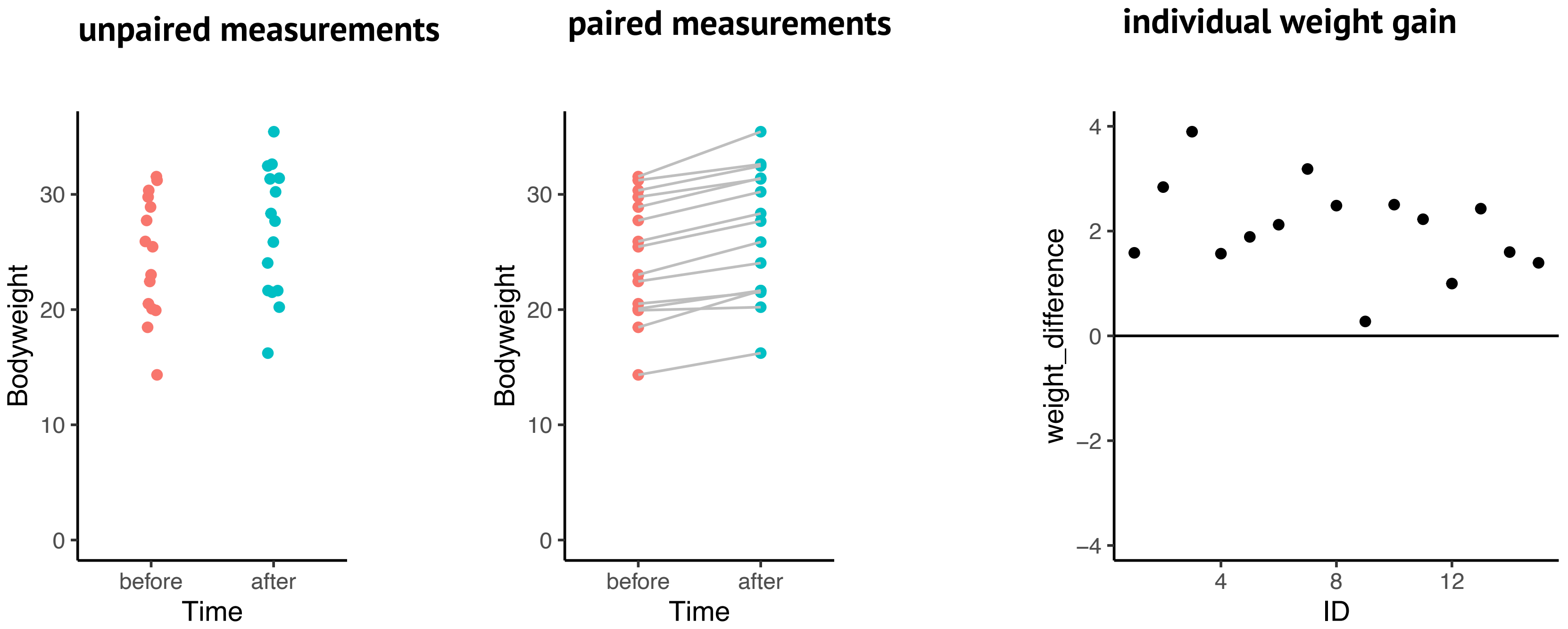


high variance between individual mice weights

Unpaired t-test:
$p = 0.31$
estimated difference: 2.06

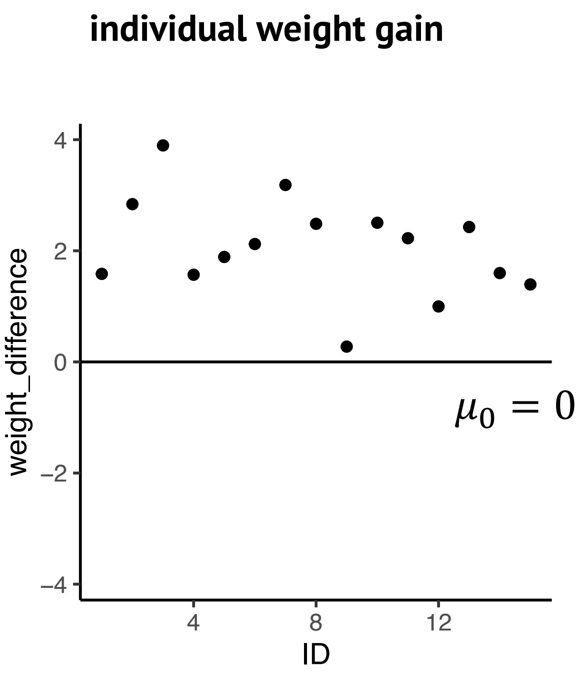
Simulated data

Paired t-test



Simulated data

Paired t-test



One-sample t-test

H_0: the mean weight gain/loss is equal to zero.

estimated difference: 2.06
p-value: 3×10^{-7}

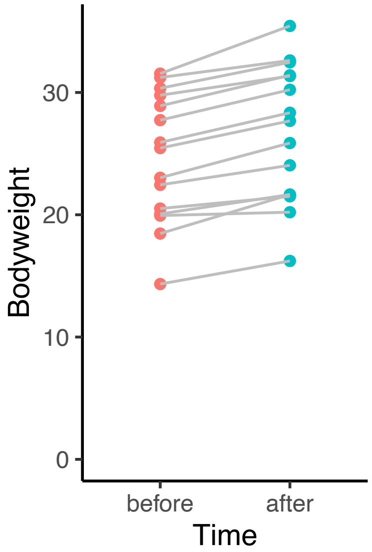
Pairing increases power

The paired t-test has an increased power compared to the two-sample t-test

Sources of randomness:

- individual responses to the treatment
- mice have different weights to start with

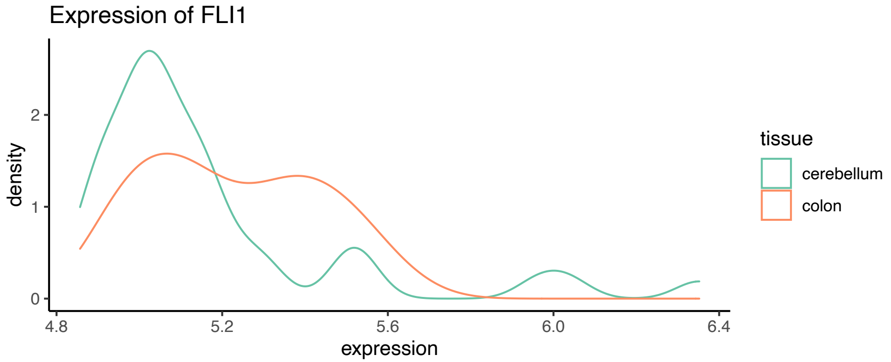
controlled for in paired design

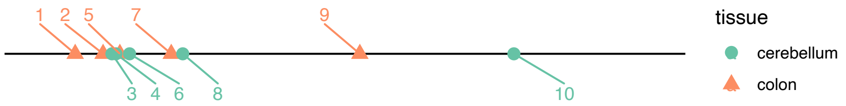


Question

Why did the authors of the real study decide NOT to set up a paired experiment?

Wilcoxon test





<u>test statistic</u>: $U_x = \sum_X rank - \frac{n_x(n_x-2)}{2}$

$U = \min(U_x, U_y)$

rank sum in case X has all the lower ranks

This test is used for non-Gaussian distributions.

Null hypothesis:
The two distributions X and Y are equal.
$P(X > Y) = P(Y > X)$

Statistic:
The value of U gets small in case the rank sums differ between the groups.

Be aware:

- distances don't matter!
- t-test usually has higher power than the Wilcoxon test.

p-value: 0.04

Data from dslabs package

Summary: testing workflow

EMBL

1. Set up a hypothesis H_0 that you want to reject.

2. Find a test statistic that should be sensitive to deviations from H_0 .

3. Find the null distribution of the test statistic – the distribution that it follows under the null hypothesis.

4. Compute the actual value of the test statistic.

5. Compute the p–value: The probability of seeing a value as least as extreme as the computed value in the null distribution.

6. Decide (based on significance level) whether to reject the null hypothesis.

In practice

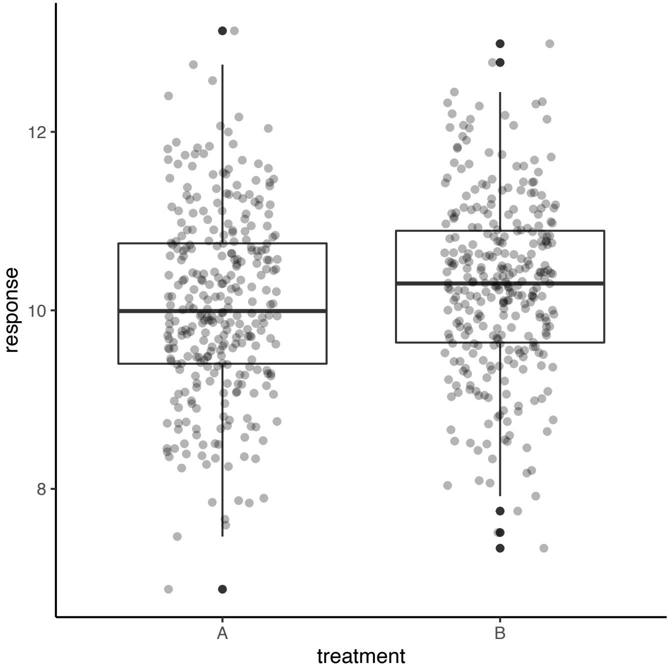
1. Look at your data!
2. Decide on a distribution that your data follow.
3. Possibly transform your data to match a suitable distribution (suitable: a convenient test is available for this distribution).
4. Find a test that answers your question and is suitable for the distribution (or generally: the properties) of your data.
5. Perform the test. Report the p-value **and** the effect size.

Interpreting p-values

- The p-value is the probability that the observed data could happen, under the condition that the null hypothesis is true.
- It is *not* the probability that the null hypothesis is true.
- Absence of evidence is not evidence of absence.
- Significance levels are arbitrary.
- Siginficant effect does not imply *relevant* effect.

Question

How do you interpret this outcome?



T-test:
p=0.01

References

Winzell, M. S., & Ahrén, B. (2004). The high-fat diet-fed mouse: A model for studying mechanisms and treatment of impaired glucose tolerance and type 2 diabetes. *Diabetes*, *53*(SUPPL. 3). https://doi.org/10.2337/diabetes.53.suppl_3.S215