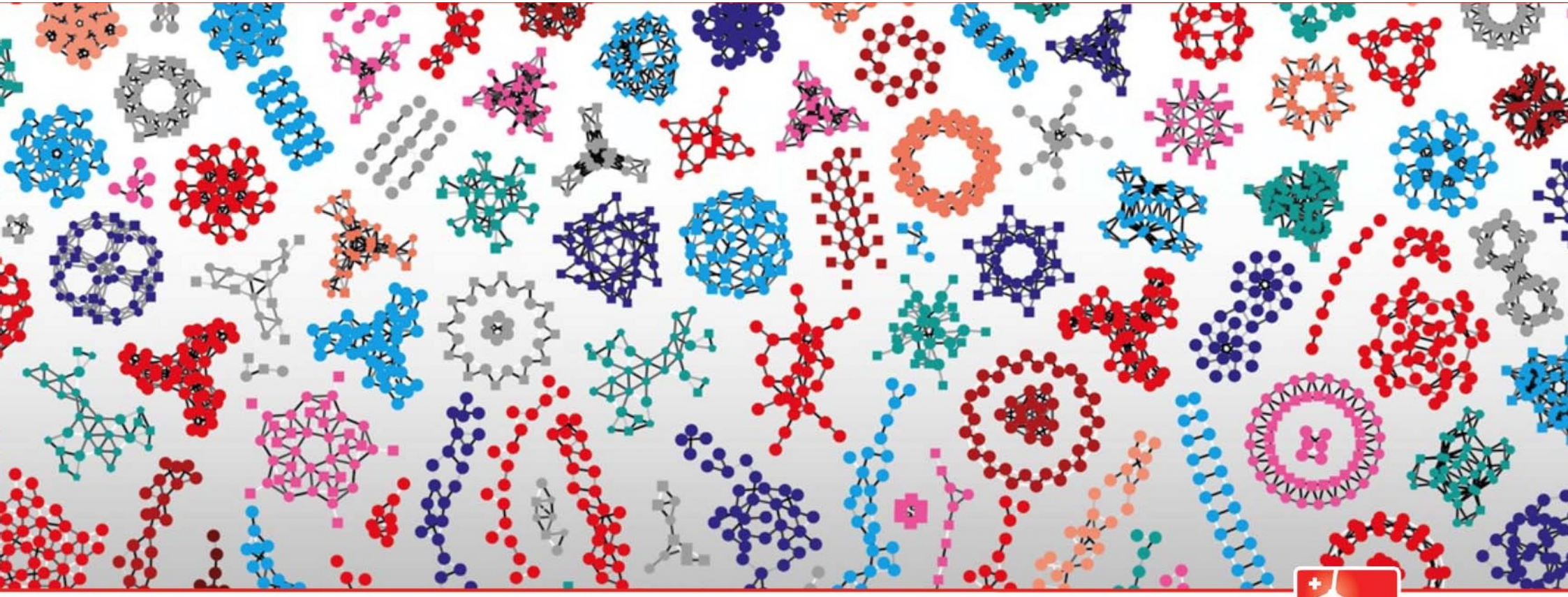


Introduction to Statistics

Joao Lourenço (joao.lourenco@sib.swiss) and Rachel Marcone (rachel.jeitziner@sib.swiss)

SIB Swiss Institute of Bioinformatics

6th-9th February 2023

A possible experiment

195 adults treated with treatment A ← 800 adults with flu → 605 adults treated with treatment B

Drug A works on 41 people out of a sample of 195. Drug B works on 351 people in a sample of 605. Are the two drugs comparable?

Introduction to hypothesis testing

Guideline for using statistics in biology

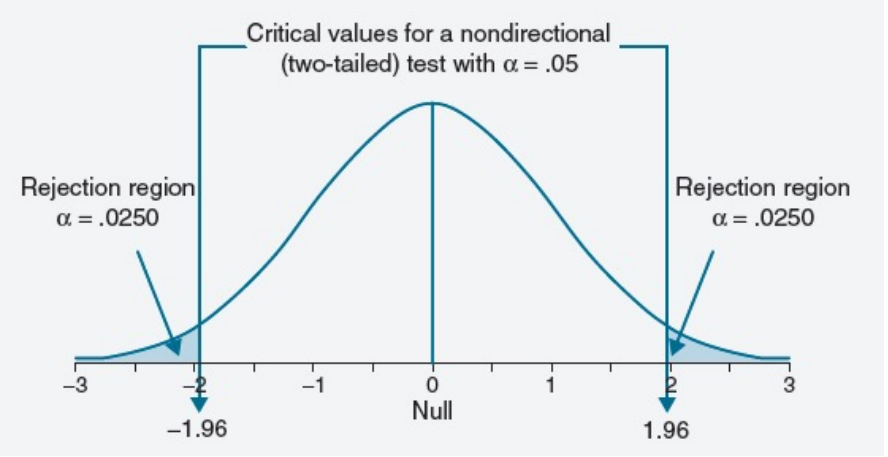
1. Specify the biological question of interest.
2. Put the question in the form of a **biological null hypothesis** and **alternate hypothesis**.
3. Put the question in the form of a **statistical null hypothesis** and **alternate hypothesis**.
4. Determine which **variables** are relevant to the question and what kind of variable each one is.
5. **Design an experiment** that controls or randomizes the **confounding variables**.
6. Based on the number of variables, the kinds of variables, the expected fit to the parametric assumptions, and the hypothesis to be tested, **choose the best statistical test to use**.
7. If possible, do a **power analysis** to determine a good **sample size** for the experiment.
8. Do the experiment.
9. **Examine the data** (explore variation and check if the assumptions of the statistical test you chose - primarily normality and homoscedasticity for tests of measurement variables - are met - if it doesn't, choose a more appropriate test).
10. **Apply the statistical test** you chose, and **interpret** the results.
11. **Communicate** your results **effectively**.

A possible experiment

195 adults treated with treatment A ⟸ 800 adults with flu ⟹ 351 adults treated with treatment B

Drug A works on 41 people out of a sample of 195. Drug B works on 351 people in a sample of 605. Are the two drugs comparable?

Comparison of 2 proportions ⟹ one option: Z test



•State the null hypothesis and alternate hypothesis.

H0: the proportions are the same.

H1: the proportions are different

.

•Choose an alpha level.

alpha = 0.05

$$z_t = \frac{\hat{p}_1 - \hat{p}_2}{\hat{\sigma}_D} = 8.99$$

- Find the critical value of z using the standard normal distribution.
- Calculate the z test statistic.
- Compare the test statistic to the critical z value and decide if you should support or reject the null hypothesis.

8.99 > 1.96, so we can reject the null hypothesis.

$$\hat{\sigma}_D = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

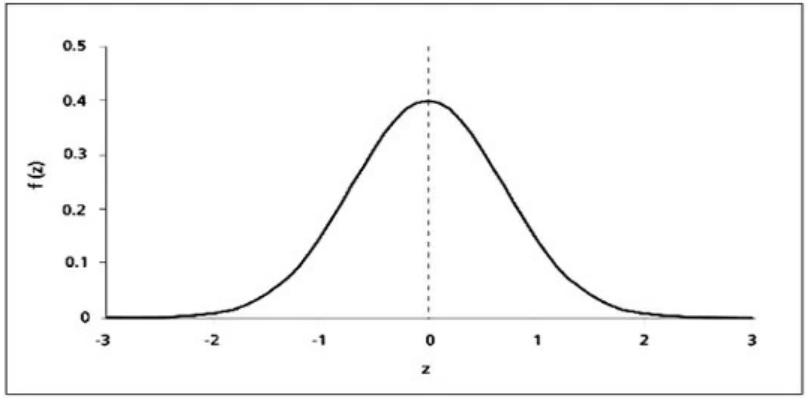
$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

How to judge whether a difference is significant ?

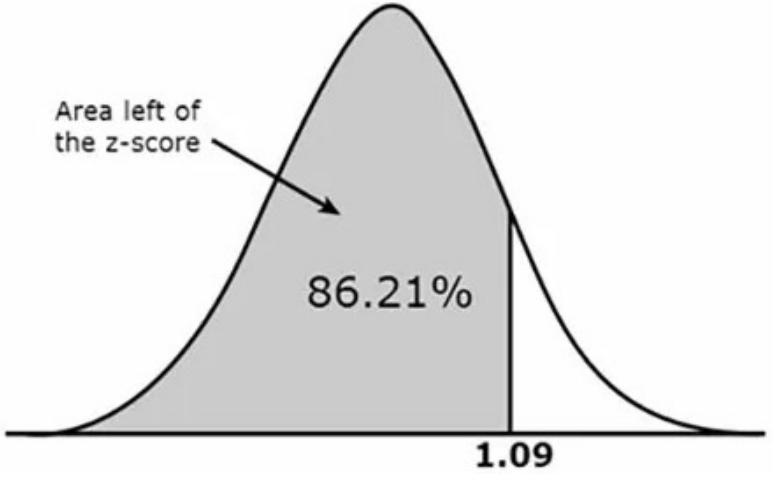
- The p-value is the probability of getting a result that is as or more extreme than the observed result, assuming that the null hypothesis is true.
- A p-value is **not** the probability that the null hypothesis is correct.
- A p-value is **not** the probability of making an error.

How to judge whether a difference is significant ?

- A predefined significance level (α) is defined, typically 0.05 or 0.01
- The value of the test statistic which correspond to the significance level is calculated or *often* read in tables.
- If the observed test statistic is above the threshold, we reject the null hypothesis.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

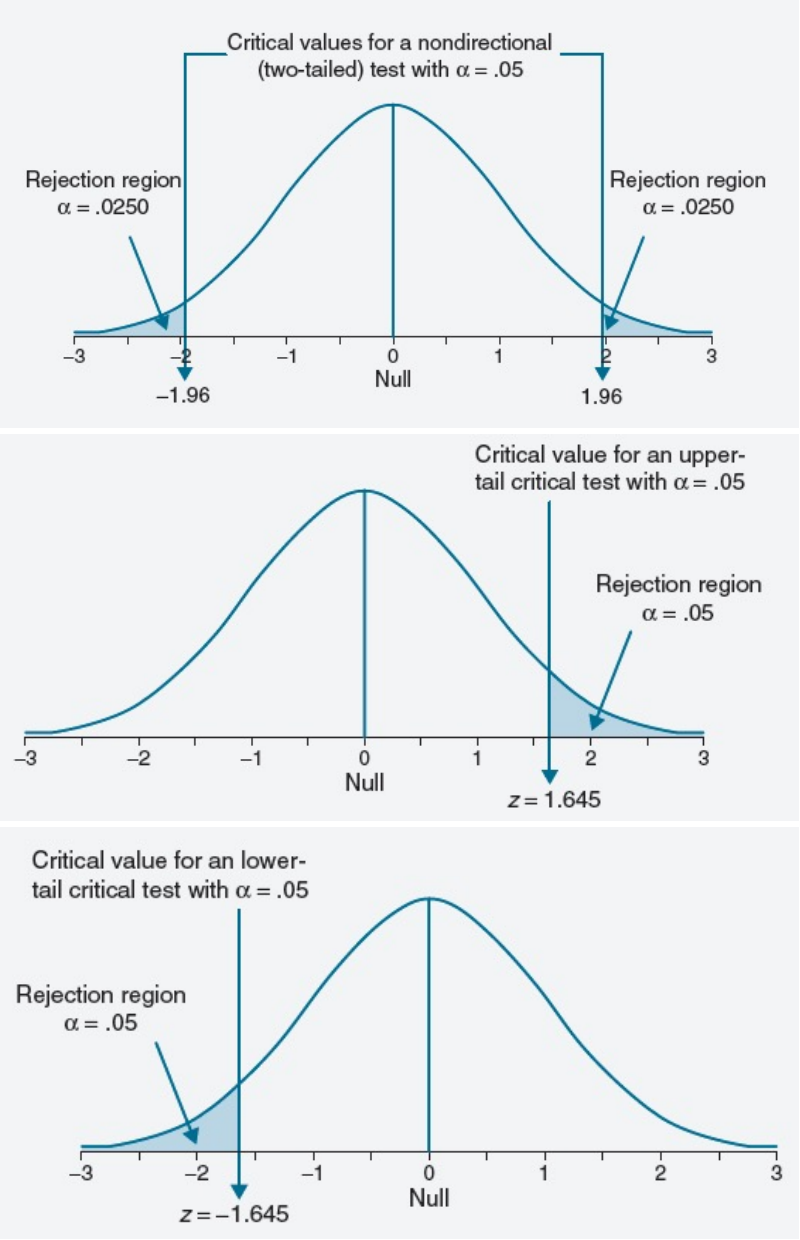


How to judge whether a difference is significant ?

- If "p<0.05", we don't know if it is 0.049 (barely significant) or 0.000000001 (extremely significant)
- Computers can now calculate exact p-values, which should be reported.
- “p<0.05” remains a magical threshold for many scientists

Two-sided test versus one-sided test

- Two-sided, nondirectional, two-tailed hypothesis tests (H1: ≠)
  - H0: the proportions are the same: p1 = p2
  - H1: the proportions are different: p1 ≠ p2
- One-sided, directional, upper-tail hypothesis tests (H1: >)
  - H0: the proportions are the same: p1 = p2
  - H1: p1 is larger than p2: p1 > p2
- One-sided, directional, lower-tail hypothesis tests (H1: >)
  - H0: the proportions are the same: p1 = p2
  - H1: p1 is smaller than p2: p1 < p2



Difference between two-samples and two-sided tests

- A **two-samples test** is a hypothesis test for answering questions about means for two different populations. Data are collected from two random samples of independent observations.

- A **two-sided test** (or two-tailed test) is a hypothesis test in which the values for rejecting the null hypothesis are in both tails of the probability distribution

- The choice between a one-sided test and a two-sided test is determined by the purpose of the investigation or prior information

Pitfalls in hypothesis testing

- Even if a result is 'statistically significant', it can still be due to chance.
- Conversely, if a result is not statistically significant, it may be only because you do not have enough data (**lack of power**)
- A test of significance does not say **how important the difference is**, or **what caused it** (Is H0 incorrect ? Was an assumption violated ? Were you unlucky ?)
- Using a significance level transforms a complicated, real-world problem, into a simple dichotomous question.

Statistical significance
is not the same
as practical importance.

Published: 14 December 2008

Six new loci associated with body mass index highlight a neuronal influence on body weight regulation

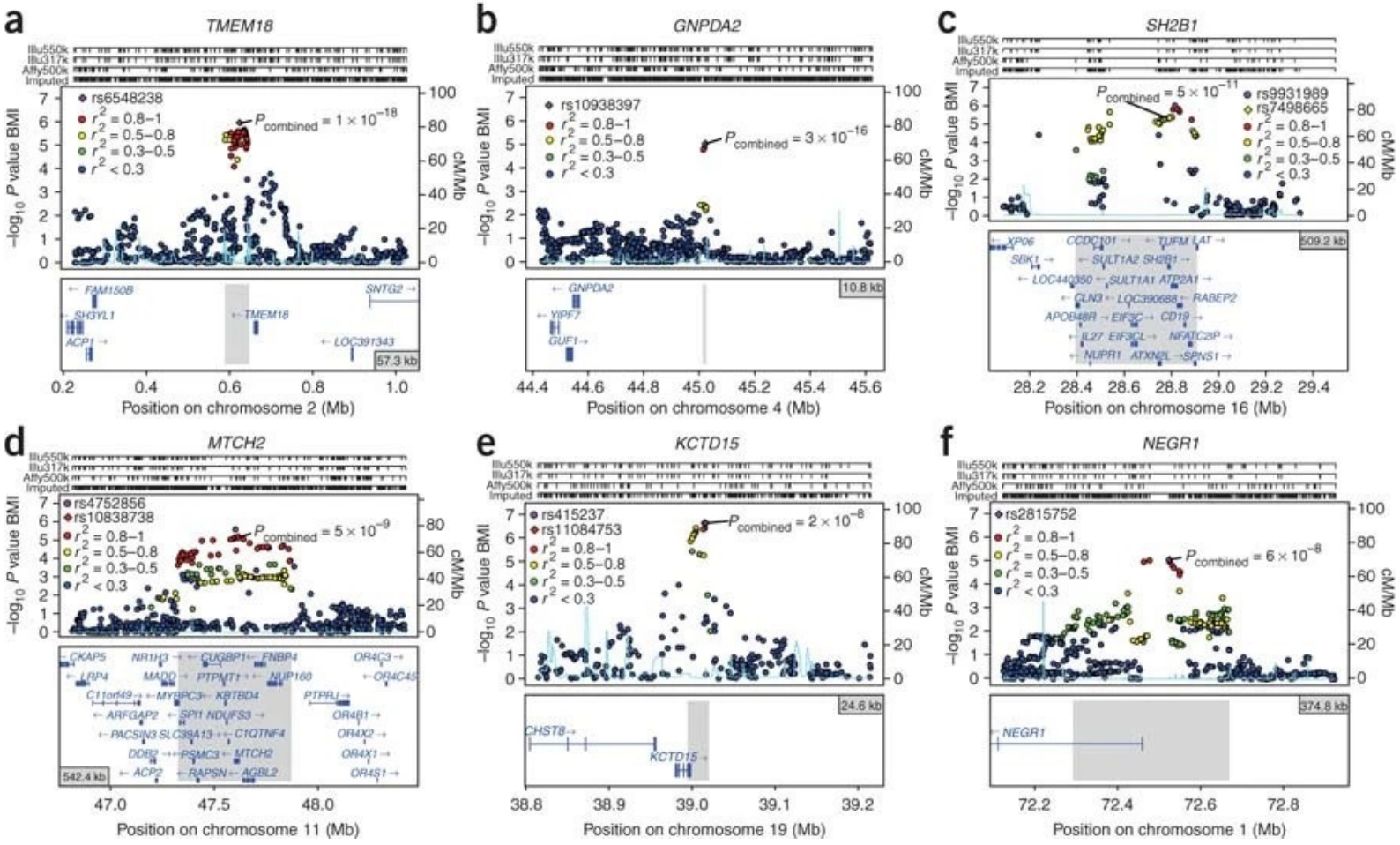
the GIANT Consortium

Nature Genetics **41**, 25–34(2009) | Cite this article

Abstract

Common variants at only two loci, *FTO* and *MC4R*, have been reproducibly associated with body mass index (BMI) in humans. To identify additional loci, we conducted meta-analysis of 15 genome-wide association studies for BMI ($n > 32{,}000$) and followed up top signals in 14 additional cohorts ($n > 59{,}000$). We strongly confirm *FTO* and *MC4R* and identify six additional loci ($P < 5 \times 10^{-8}$): *TMEM18*, *KCTD15*, *GNPDA2*, *SH2B1*, *MTCH2* and *NEGR1* (where a 45-kb deletion polymorphism is a candidate causal variant). Several of the likely causal genes are highly expressed or known to act in the central nervous system (CNS), emphasizing, as in rare monogenic forms of obesity, the role of the CNS in predisposition to obesity.

Statistical significance is not the same as practical importance



8 SNPs (6 discovered in the study) significantly associated with BMI.

They correspond to a change of **173–954 g** in weight per allele in adults who are 160–180 cm tall

Which test should I use ?

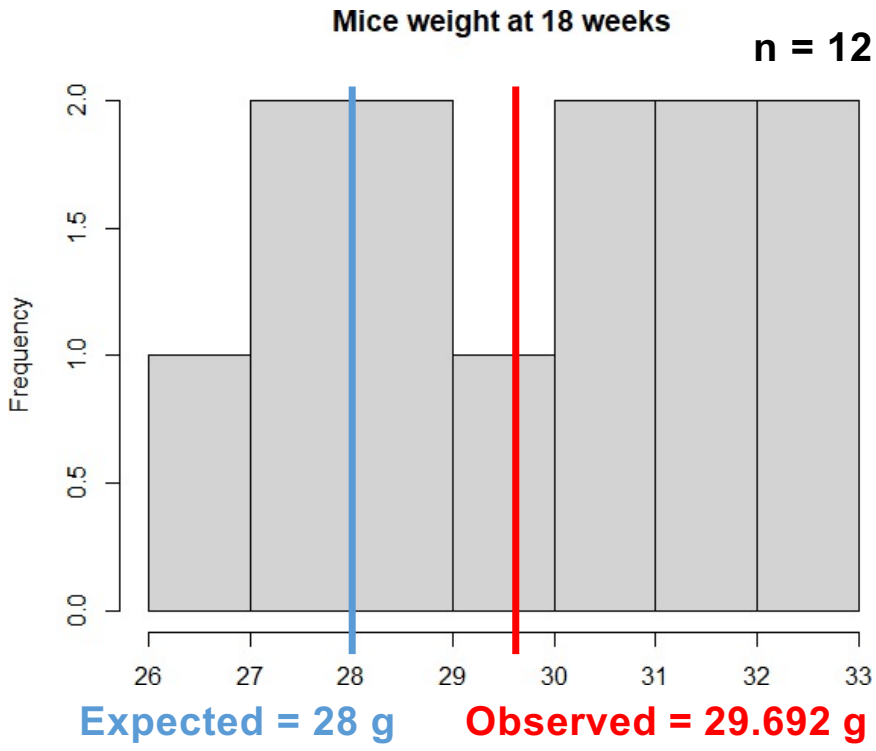
The most widely used tests try to answer questions about the *location* of the center of the data (e.g. mean or median).

Which test should I use ?

We have data about mice for which a gene was knocked out.

Question:

Is their weight different from the mean weight of the mice lab population (e.g. 28 g) ?



Which test should I use ?

H0: the mean of the mice weight in our sample is equal to 28
H1: the mean of the mice weight in our sample is not equal to 28

To perform this hypothesis test, we can use a one-sample t-test.

The most commonly used of all tests (80-90% of all papers ?)

Main assumptions:

- The data are **continuous**.
- The data are **independent**.
- The sample data have been **randomly** sampled from a population.
- **No significant outliers** in the data
- **Normality**: the data should be approximately normally distributed

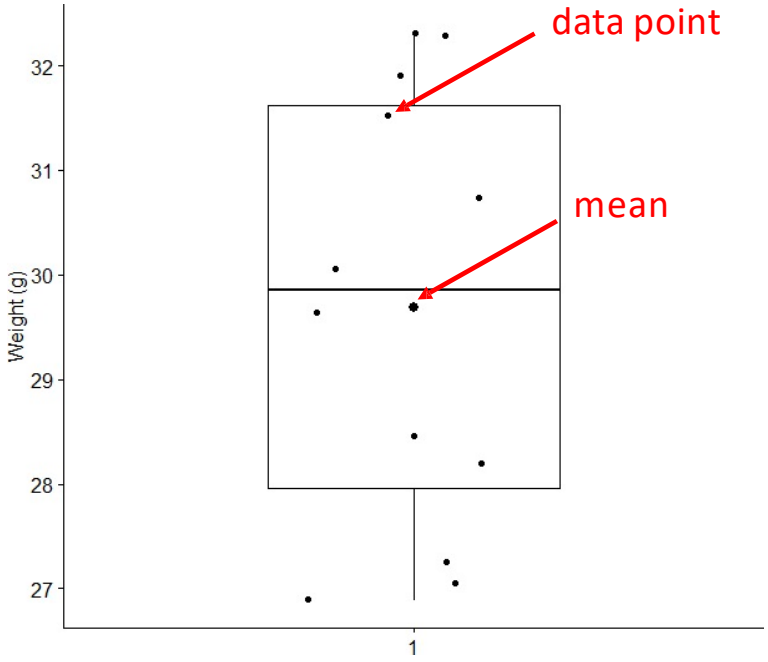
One-sample t-test

Main assumptions:

- **No significant outliers** in the data

```
>ggboxplot(weight$weight, width = 0.5, add =
c("mean","jitter"), ylab = "Weight (g)", xlab = F)
```

```
> identify_outliers(weight)
## [1] name       weight     is.outlier is.extreme
## <0 rows> (or 0-length row.names)
```



Values above Q3 + 1.5xIQR or below Q1 - 1.5xIQR are considered as outliers.
Q1 and Q3 are the first and third quartile, respectively.
IQR is the interquartile range (IQR = Q3 - Q1).

```
> Q1 <- quantile(weight$weight, probs = 0.25)
> Q3 <- quantile(weight$weight, probs = 0.75)
> IQR <- Q3-Q1
```

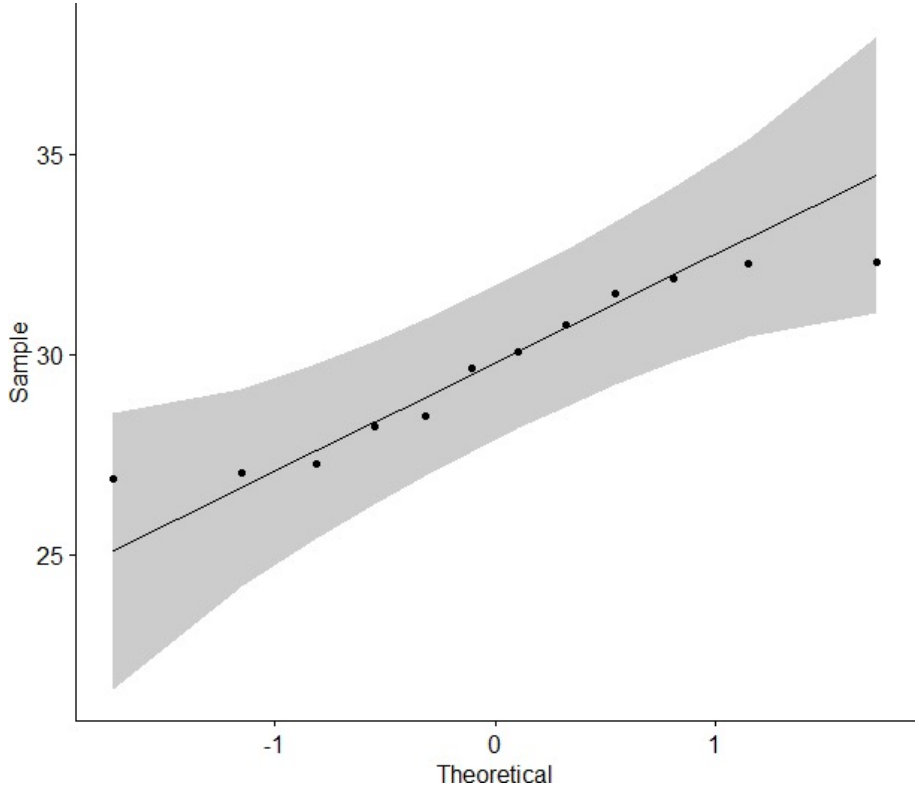
One-sample t-test

Main assumptions:

- **Normality**: the data should be approximately normally distributed

```
> shapiro_test(weight$weight)
# A tibble: 1 x 3
  variable       statistic p.value
  <chr>              <dbl>   <dbl>
1 weight$weight      0.902   0.166

> ggqqplot(weight, x = "weight")
```



One-sample t-test

Test-statistic (Student's t-statistic):

$$T = \frac{\bar{x} - \mu}{\sqrt{S^2/n}}$$

where

– $\bar{x}$ is the average of the observations (29.692g)
– μ is the mean weight of the mice lab population (28g)
– S is the (estimated) standard deviation (2.081g)
– n is the number of observations (12)

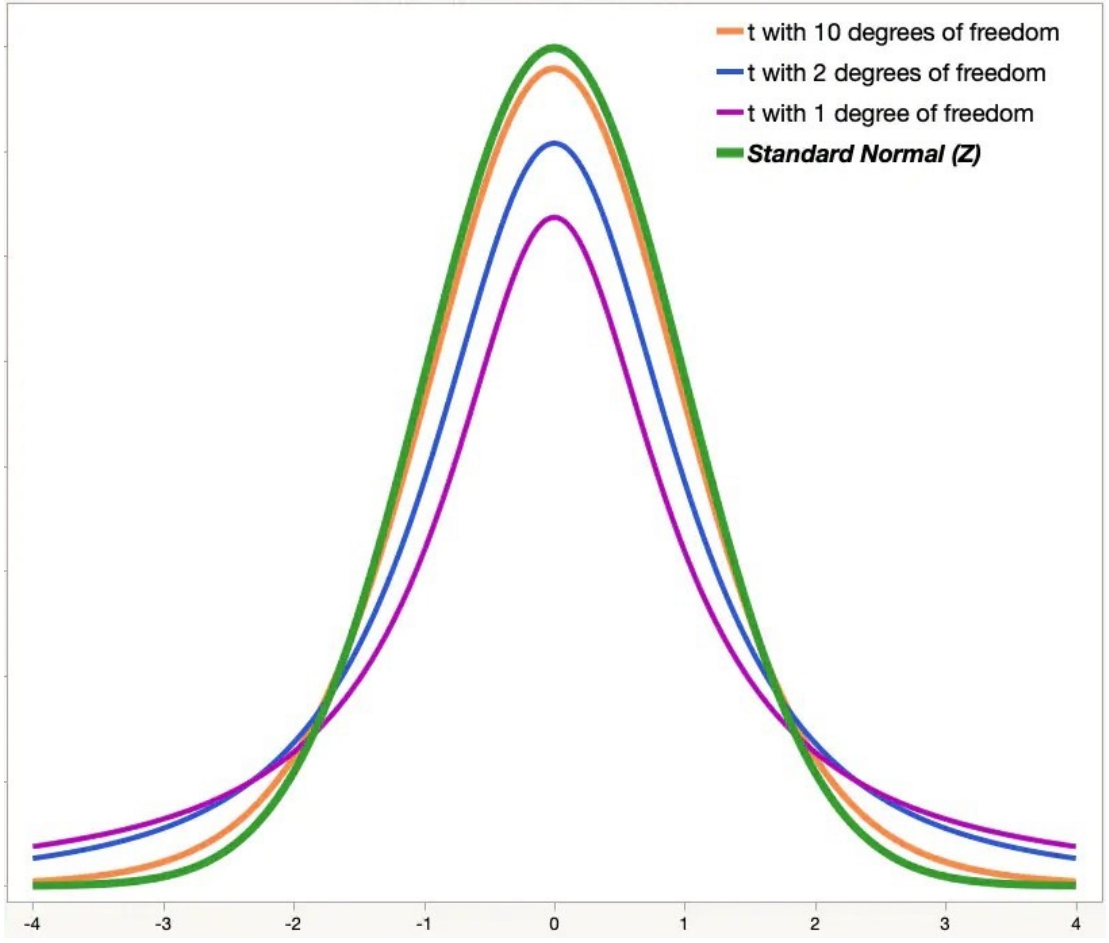
The t-distribution

The t-distribution describes the standardized distances of sample means to the population mean **when the population standard deviation is not known**, and the observations come from a normally distributed population.

The t-distribution is similar to a normal distribution.

- Like the normal distribution, the t-distribution has a **smooth shape**.
- Like the normal distribution, the t-distribution is **symmetric**.
- Like a standard normal distribution (or z-distribution), the t-distribution **has a mean of zero**.
- The t-distribution is defined by the **degrees of freedom**. These are related to the sample size.
- The t-distribution is most useful for **small sample sizes**, when the population standard deviation is not known, or both.
- As the sample size increases, the t-distribution becomes more similar to a normal distribution.

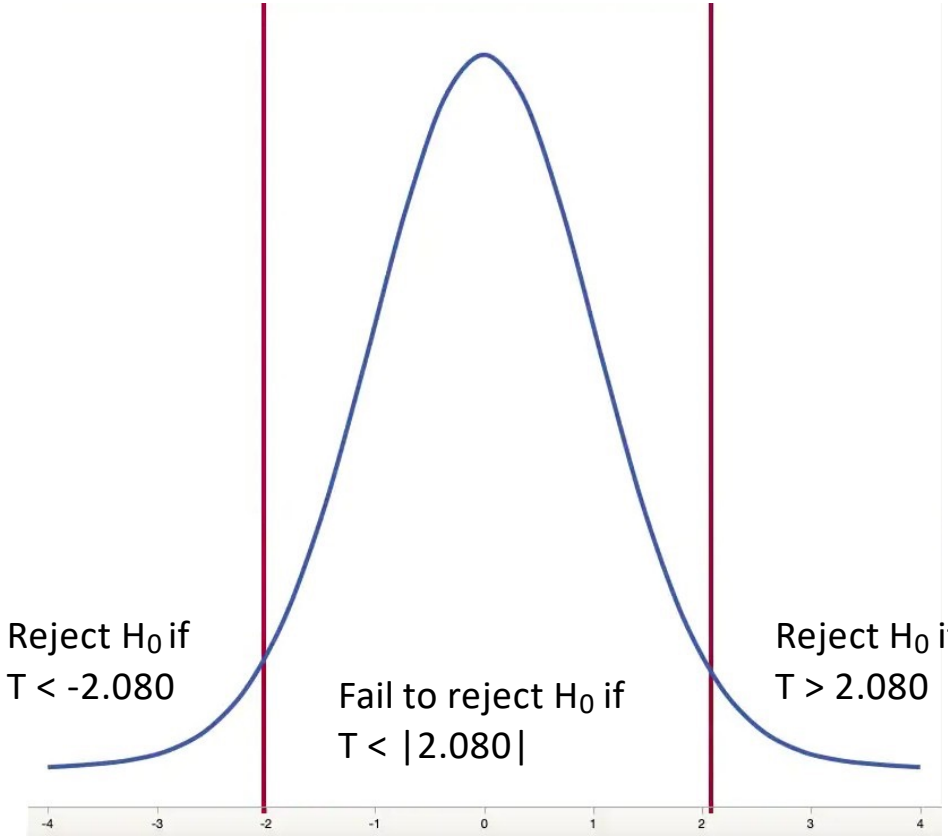
The t-distribution



The t-distribution

two-tailed test

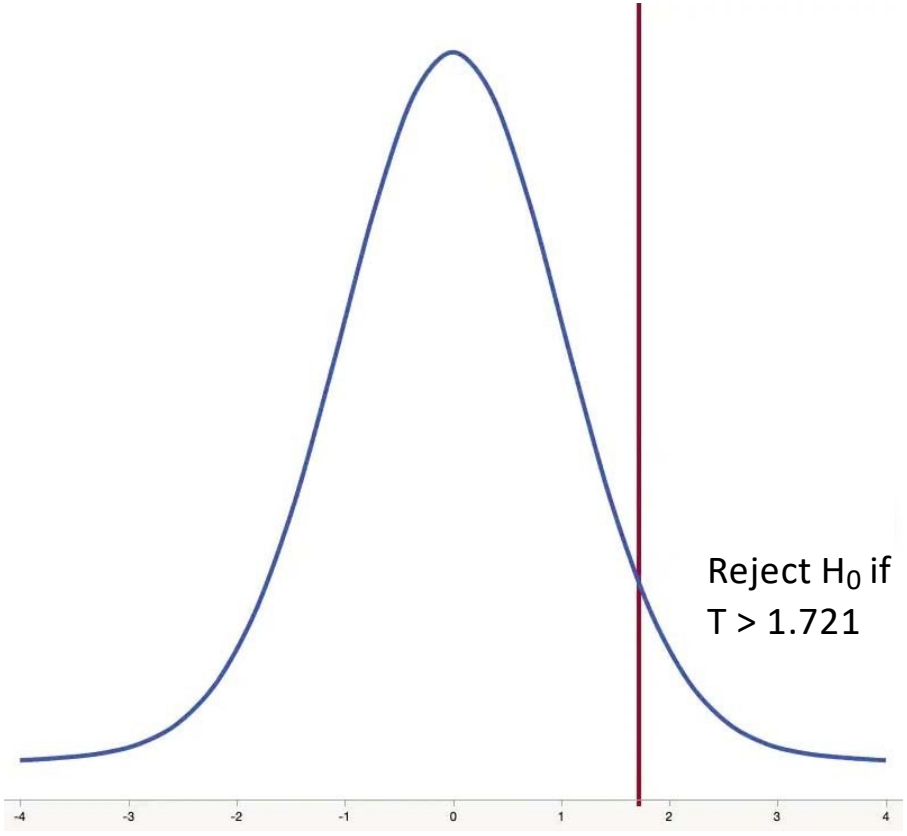
t-distribution with df = 21



$t_{n-1,1-\alpha/2}=t_{21,0.975}=2.080$

one-tailed test

t-distribution with df = 21



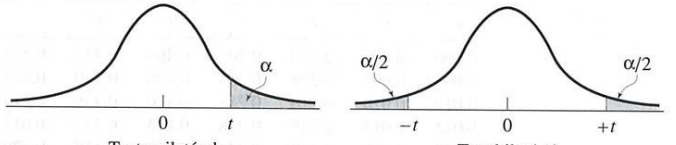
$t_{n-1,1-\alpha}=t_{21,0.95}=1.721$

One-sample t-test

```
> t.test(weight$weight, mu = 28)

                One Sample t-test

data:  weight$weight
t = 2.8162, df = 11, p-value = 0.01678
alternative hypothesis: true mean is not equal to 28
95 percent confidence interval:
 28.36953 31.01366
sample estimates:
mean of x
 29.69159
```



	Seuil de signification pour le test unilatéral								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	Seuil de signification pour le test bilatéral								
dl	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.620
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437

One-sample t-test

```
> t.test(weight$weight, mu = 28)

                One Sample t-test

data:  weight$weight
t = 2.8162, df = 11, p-value = 0.01678
alternative hypothesis: true mean is not equal to 28
95 percent confidence interval:
 28.36953 31.01366
sample estimates:
mean of x
 29.69159
```

$$T = \frac{\bar{x} - \mu}{\sqrt{S^2/n}}$$

$$df = n - 1$$

$$\Pr\{-2.201 < T < 2.201\} = 0.95$$

$$\Pr\left\{-2.201 < \frac{\bar{x} - \mu}{\sqrt{S^2/n}} < 2.201\right\} = 0.95$$

$$\Pr\left\{\bar{x} - 2.201\sqrt{S^2/n} < \mu < \bar{x} + 2.201\sqrt{S^2/n}\right\} = 0.95$$

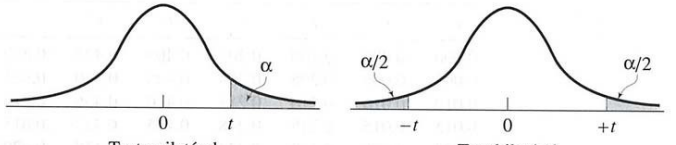
One-sample t-test

```
> t.test(weight$weight, mu = 28, alternative="greater")

                One Sample t-test

data:  weight$weight
t = 2.8162, df = 11, p-value = 0.008391
alternative hypothesis: true mean is greater than 28
95 percent confidence interval:
 28.61286      Inf
sample estimates:
mean of x
 29.69159
```

$$\Pr\{\ \mu > \bar{x} - 1.796\sqrt{S^2/n}\ \} = 0.95$$



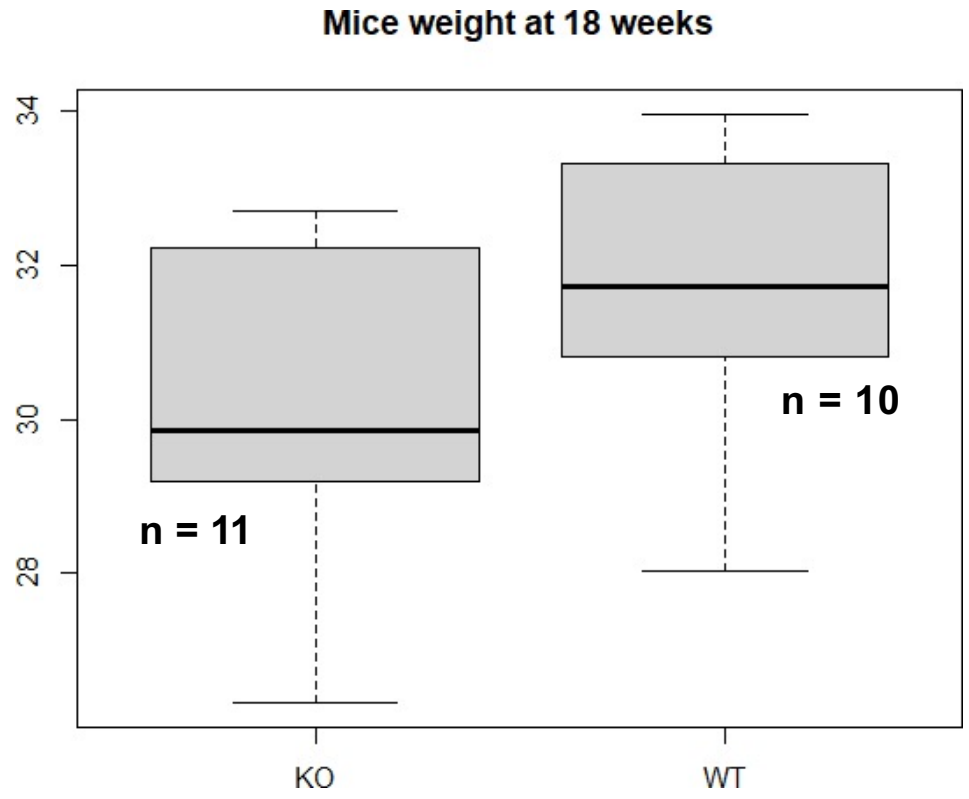
	Seuil de signification pour le test unilatéral								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	Seuil de signification pour le test bilatéral								
dl	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.620
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437

Which test should I use ?

We have data about mice for which a gene was knocked out (KO), as well as control mice (WT)

Question:

Is there a significant difference between the average weight of these two groups ?



Which test should I use ?

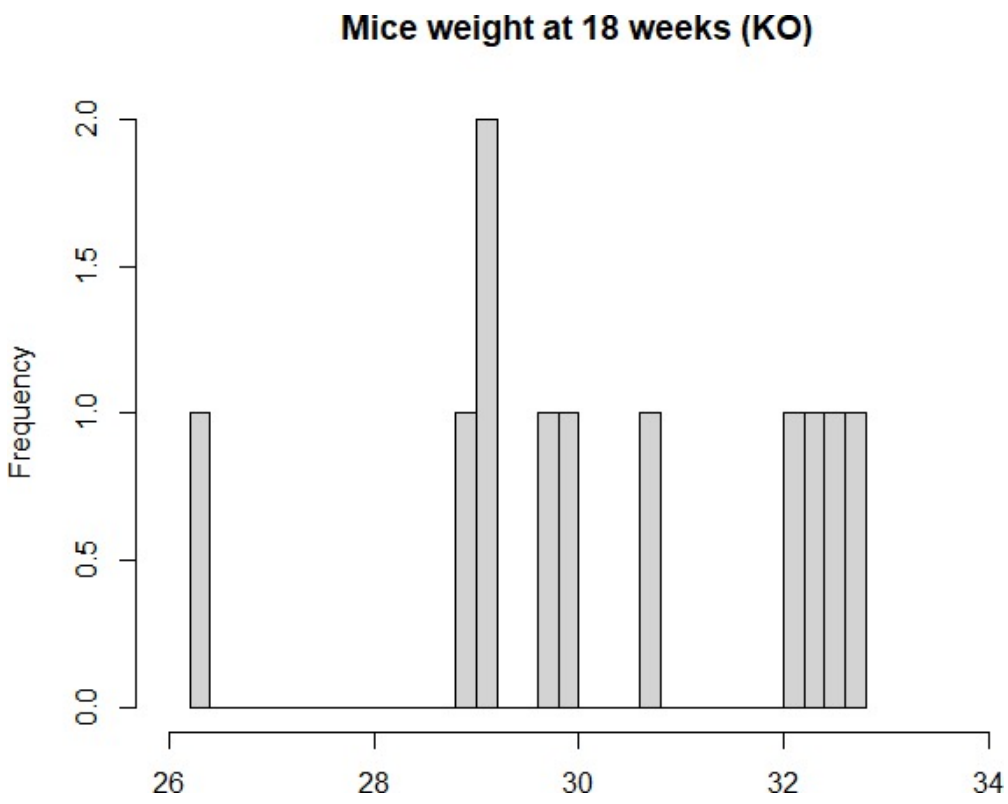
H0: the mean of the two groups is the same
H1: the mean of the two groups is different

To perform this hypothesis test, we can use a two-sample t-test.

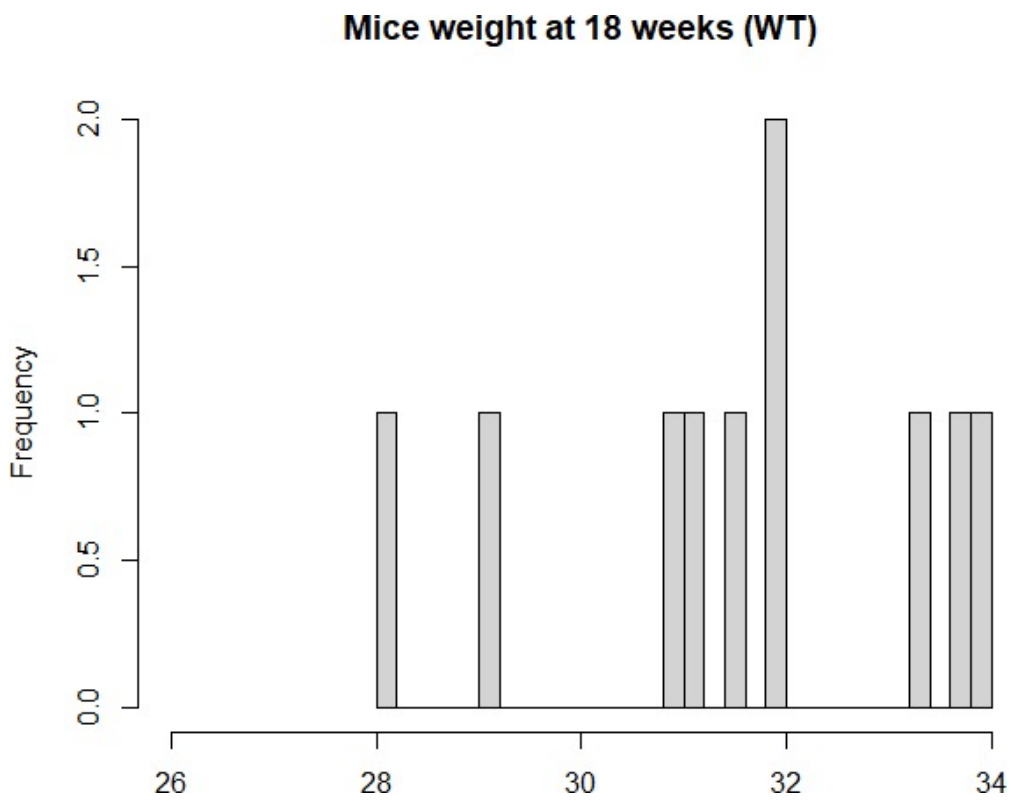
Main assumptions:

- Data values must be **independent**.
- Data in each group must be obtained via a **random sample** from the population.
- Data in each group are **normally** distributed.
- Data values are **continuous**.
- The variances for the two independent groups are **equal**.

Two-sample t-test



n = 11
Mean = 30.324
Standard deviation = 1.986



n = 10
Mean = 31.542
Standard deviation = 1.928

One-sample t-test

Test-statistic (Student's t-statistic):

$$T = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{S_{\mathrm{p}}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

– $\overline{x_1}$ is the average of the observations for WT mice (30.324g)

– $\overline{x_2}$ is the average of the observations for KO mice (31.542g)

– S_{p}^2 is the (estimated) pooled variance

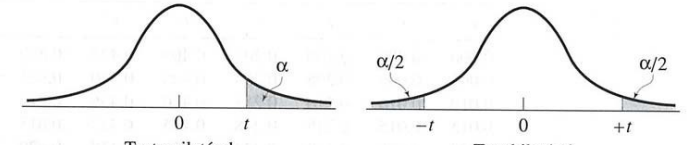
$$S_{\mathrm{p}}^2 - \frac{((n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Two-sample t-test

```
> t.test(KO_WT$weight ~ KO_WT$genotype)

                    Welch Two Sample t-test

data:  KO_WT$weight by KO_WT$genotype
t = -1.4261, df = 18.905, p-value = 0.1702
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.0078465  0.5705536
sample estimates:
mean in group KO mean in group WT
       30.32366        31.54231
```



	Seuil de signification pour le test unilatéral								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	Seuil de signification pour le test bilatéral								
dl	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.620
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437

Two-sample t-test

```
> t.test(KO_WT$weight ~ KO_WT$genotype)

                    Welch Two Sample t-test

data:  KO_WT$weight by KO_WT$genotype
t = -1.4261, df = 18.905, p-value = 0.1702
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.0078465  0.5705536
sample estimates:
mean in group KO mean in group WT
       30.32366        31.54231
```

$$T = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1 - 1)} + \frac{S_2^4}{n_2^2(n_2 - 1)}}$$

Two-sample t-test

```
> t.test(KO_WT$weight ~ KO_WT$genotype, var.equal = T)

        Two Sample t-test

data:  KO_WT$weight by KO_WT$genotype
t = -1.4239, df = 19, p-value = 0.1707
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.0099018  0.5726089
sample estimates:
mean in group KO mean in group WT
        30.32366         31.54231
```

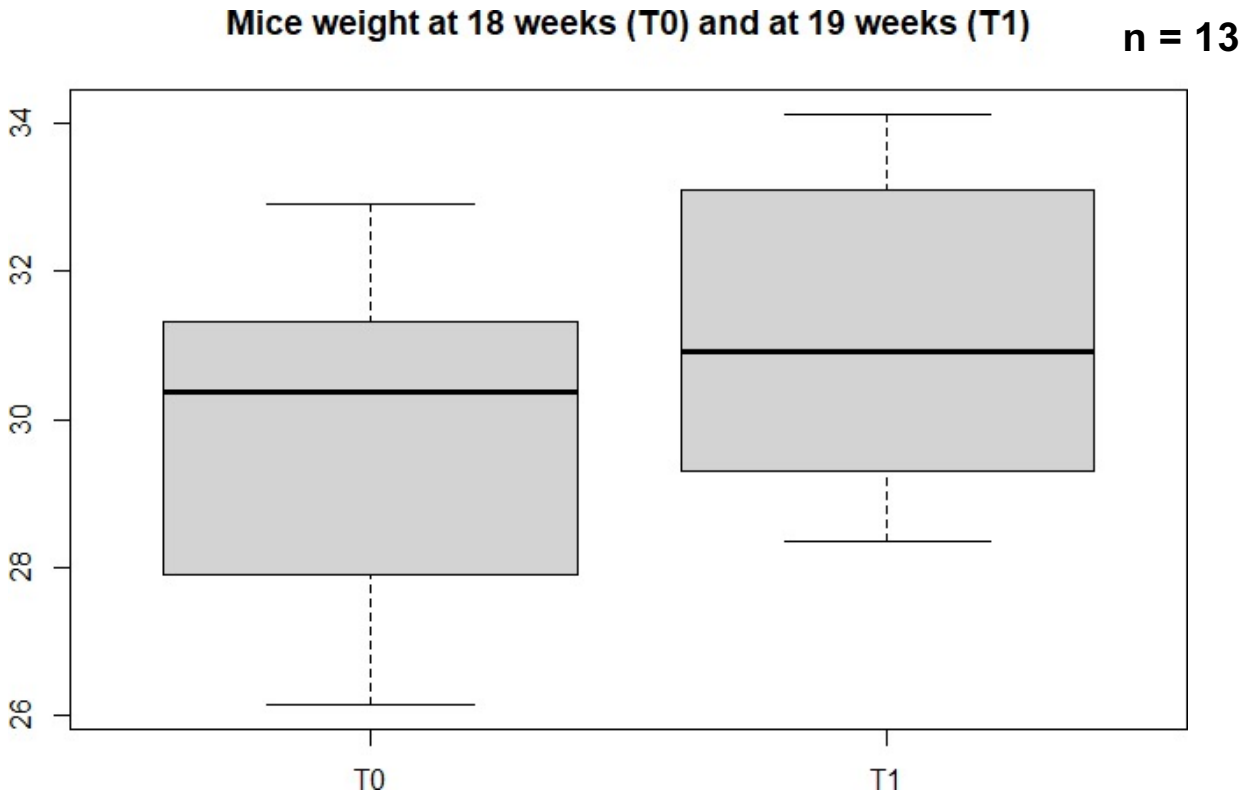
$$T = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad df = n_1 + n_2 - 2$$

Which test should I use ?

We have data about mice at two different time points (T_0 and T_1)

Question:

Is there a significant difference between the average weight of these mice at these two time points ?



Which test should I use ?

H0: the mean of the differences is zero
H1: the mean of the differences is not zero

To perform this hypothesis test, we can use a paired t-test. Main

assumptions:

- Subjects must be **independent**. Measurements for one subject do not affect measurements for any other subject.
- Each of the **paired measurements** must be obtained from the **same** subject.
- The measured differences are **normally** distributed.

Paired t-test

- In the two-sample t-test, we compared two samples of unrelated data points
- If the data between the two samples is paired, that is, each point x_i in the first sample correspond to a point y_i in the second sample, we can do a paired t-test
- Equivalent to testing if the difference between the pairs is significantly different from zero.
- More powerful than the two-sample t-test because we provide more information (the pairing) to the test

Two-sample t-test

```
> t.test(T0_T1$weight_T0,T0_T1$weight_T1)

        Welch Two Sample t-test

data:  T0_T1$weight_T0 and T0_T1$weight_T1
t = -2.3758, df = 23.97, p-value = 0.02585
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.8996244 -0.2738217
sample estimates:
mean of x mean of y
 29.89671  31.98343
```

$$T = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1 - 1)} + \frac{S_2^4}{n_2^2(n_2 - 1)}}$$

Paired t-test

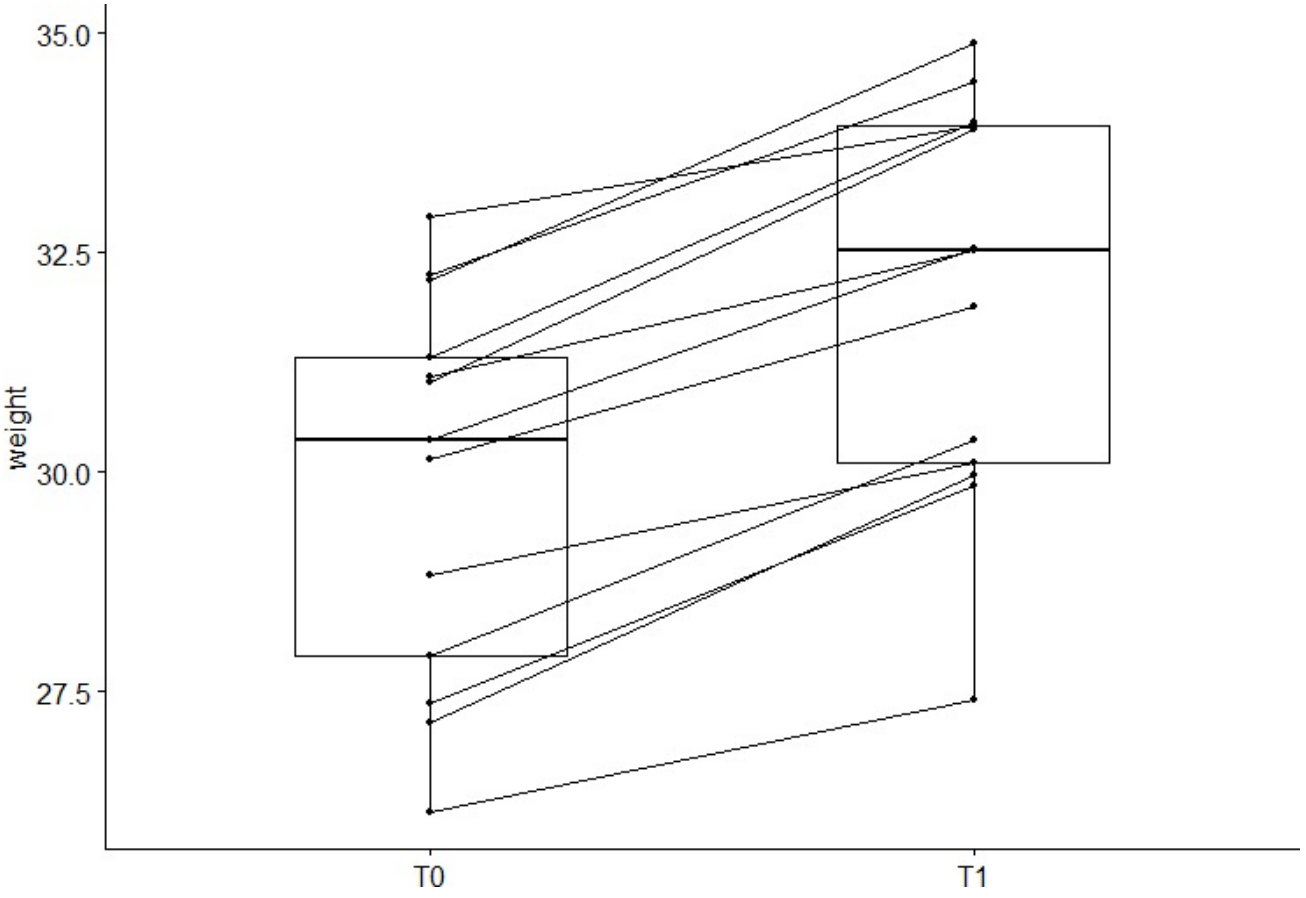
```
> t.test(T0_T1$weight_T0,T0_T1$weight_T1, paired = T)

        Paired t-test

data:  T0_T1$weight_T0 and T0_T1$weight_T1
t = -11.537, df = 12, p-value = 7.491e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.480824 -1.692622
sample estimates:
mean of the differences
              -2.086723
```

$$T = \frac{\bar{X} - \bar{Y}}{S_D / \sqrt{n}} \qquad S_D = \sqrt{\frac{\sum (x_i - y_i)^2 - \frac{(\sum (x_i - y_i))^2}{n}}{n-1}} \qquad df = n - 1$$

The right data visualization for paired data



Multiple testing

Type I and type II errors

Decision / «Truth»	H0 true	H1 true
Do not reject H0	Correct decision 1-α	Incorrect decision Type II error β
Reject H0	Incorrect decision Type I error α	Correct decision 1-β

$$\alpha = P(Type\ I\ error) \qquad \beta = P(Type\ II\ error)$$

Why Multiple Testing Matters

If we perform *m* hypothesis tests, what is the probability of at least 1 false positive?

$$P(Type\ I\ error) = \alpha$$

$$P(not\ making\ Type\ I\ error) = 1 - \alpha$$

$$P(not\ making\ Type\ I\ error\ in\ m\ tests) = (1 - \alpha)^m$$

$$P(making\ at\ least\ 1\ Type\ I\ error\ in\ m\ tests) = 1 - (1 - \alpha)^m$$

Probability of false positives increases with number of tests

Number of tests	Probability that **at least one event** is significant just by chance
1	0.050
2	0.097
3	0.142
4	0.185
5	0.226
10	0.401
20	0.641
50	0.923
100	0.994

$$P(at\ least\ one\ significant\ result) = 1 - (1 - 0.05)^{\text{number of tests}}$$

Counting errors

Assume we are testing $H^1, H^2, \ldots, H^m$

m_0 = # of true hypotheses R = # of rejected hypotheses

	Null True	**Alternative True**	**Total**
Not Called Significant	***U***	***T***	***m - R***
Called Significant	***V***	***S***	***R***
	m_0	$m-m_0$	***m***

V = # Type I errors [false positives]

What Does Correcting for Multiple Testing Mean?

- Adjusting p-values for the number of hypothesis tests performed means controlling the Type I error rate
- Very active area of statistics - many different methods have been described

- Different Approaches To Control Type I Errors:
  - Family-wise error rate (FEWR): the probability of at least one type I error

$$FEWR = P(V \geq 1) \leq \alpha$$

  - False discovery rate (FDR) is the expected proportion of Type I errors among the rejected hypotheses

$$FDR = E\left(\frac{V}{R}\right) \leq \alpha$$

Bonferroni correction controls FWER

- Significance threshold = α/m
- Bonferroni correction tends to be too conservative

$$P(at\ least\ one\ significant\ result) = 1 - (1 - \frac{0.05}{20})^{20} = 0.0488$$

- It assumes that all tests are independent of each other. In practical applications, that is often not the case. Depending on the correlation structure of the tests, the Bonferroni correction could be extremely conservative, leading to a high rate of false negatives.

Holm's method controls FWER

- To control FWER at level α=0.05:
  1. Order the unadjusted p-values: $p_1 \leq p_2 \leq .. \leq p_{\mathrm{m}}$
  2. The step-down Holm adjusted p-values are

$$\tilde{p}_{\mathrm{j}} = \min[(m - j + 1)p_{\mathrm{j}}, 1]$$

  3. The point here is that we don't multiply every p_{j} by the same factor *m*

$$if\ m = 1000: \tilde{p}_1 = 1000.p_1, \tilde{p}_2 = 999.p_2, .., \tilde{p}_{\mathrm{m}} = 1.p_{\mathrm{m}}$$

FWER or FDR ?

- FWER is appropriate when you want to guard against ANY false positives
- However, in many cases (particularly in genomics) we can live with a certain number of false positives
- In these cases, the more relevant quantity to control is the false discovery rate (FDR)

Benjamini Hochberg controls FDR

- To control FDR at level δ=0.05:
  1. Order the unadjusted p-values: $p_1 \leq p_2 \leq .. \leq p_m$
  2. Find the test with the highest rank, j, for which the p value, p_j, is less than equal to $\frac{j}{m} x\delta$
  3. Declare the tests of rank 1, 2, .., j as significant

Controlling the FDR at δ = 0.05

Rank (j)	P-value	(j/m)× δ	Reject H_0 ?
1	0.0008	0.005	1
2	0.009	0.010	1
3	0.165	0.015	0
4	0.205	0.020	0
5	0.396	0.025	0
6	0.450	0.030	0
7	0.641	0.035	0
8	0.781	0.040	0
9	0.900	0.045	0
10	0.993	0.050	0

Multiple testing correction in R: p.adjust

p.adjust {stats} R Documentation

Adjust P-values for Multiple Comparisons

Description

Given a set of p-values, returns p-values adjusted using one of several methods.

Usage

```
p.adjust(p, method = p.adjust.methods, n = length(p))

p.adjust.methods
# c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY",
#   "fdr", "none")
```

Arguments

`p`
: numeric vector of p-values (possibly with `NA`s). Any other `R` object is coerced by `as.numeric`.

`method`
: correction method, a `character` string. Can be abbreviated.

`n`
: number of comparisons, must be at least `length(p)`; only set this (to non-default) when you know what you are doing!

Multiple testing correction

	X	baseMean	log2FoldChange	lfcSE	stat	pvalue
1	ENSMUSG00000000001	1200.3945707	-0.0148535315	0.09208117	-0.161309114	0.8718499450
2	ENSMUSG00000000028	26.6663265	-0.0411264150	0.40057975	-0.102667235	0.9182270789
3	ENSMUSG00000000031	21.4444727	0.0426268105	0.51792967	0.082302314	0.9344063142
4	ENSMUSG00000000037	52.3910190	-0.4151892308	0.30340015	-1.368454265	0.1711699284
5	ENSMUSG00000000049	3.4930947	-0.0136930701	1.10102747	-0.012436629	0.9900772616
6	ENSMUSG00000000056	835.3274881	0.1064909330	0.08664733	1.229015714	0.2190659145
7	ENSMUSG00000000058	446.2751056	0.1683537754	0.11789673	1.427976594	0.1532985955
8	ENSMUSG00000000078	412.0205179	-0.1306807947	0.11359624	-1.150397147	0.2499803346
9	ENSMUSG00000000085	774.8375178	-0.0217666588	0.09468269	-0.229890586	0.8181767921
10	ENSMUSG00000000088	1449.9781814	0.1003037335	0.08810335	1.138478125	0.2549208881