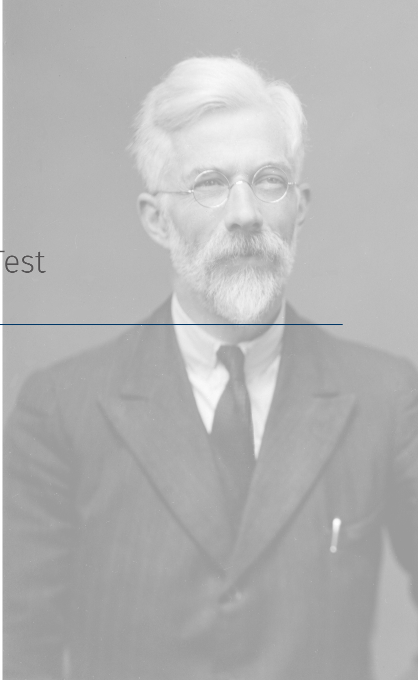
GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Data Science I

Lecture 10 – Experimental Design and Choosing a Test



Fabian Sinz

17. June 2024

Institute for Computer Science – Campus Institute for Data Science (CIDAS)

Things you should know from this lecture for the exam:

- What are the components of a statistical test?
- What is are type I (α) and type II (β) error rates?
- What is statistical power and which elements influence it how?
- What is a power analysis and how is it done?
- What is an effect size?
- How to chose the right statistical test from a book?
- What is p-hacking?
- What is multiple testing and how can it be corrected for?

Recap: Why do we use statistical tests?

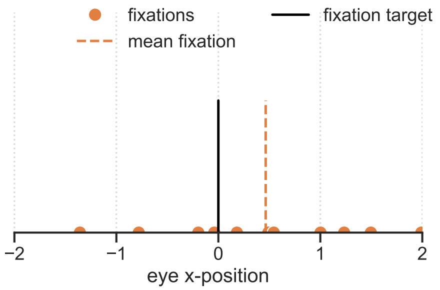
- We want to make a decision between the Null hypothesis H_0 and the alternative hypothesis H_A
- We want to quantify whether our results could have been happend by chance.
- We want to make a decision whether to accept our result/measurement as "real" based on a selected false positives level (α, type I error).
- We want to get others the chance to check whether they want to reject the Null hypothesis or not. That's why we report the p-value and not the decision.

Recap: Statistical Tests (introducing one sample t-test)

Eye Tracker

Assume you are writing a program for an eye tracker. You need to determine whether the user fixates on a target at $x = 0$. You get $n = 12$ fixation measures from the eye tracker. You

1. Need to make a decisions whether the user fixated correctly.
2. Are supposed to choose the threshold such that the program only rejects about 5% of actually correct fixations.



Recap: Statistical Tests (introducing one sample t-test)

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Eye Tracker

Assume you are writing a program for an eye tracker. You need to determine whether the user fixates on a target at $x = 0$. You get $n = 12$ fixation measures from the eye tracker. You

1. Need to make a decisions whether the user fixated correctly.
2. Are supposed to choose the threshold such that the program only rejects about 5% of actually correct fixations.

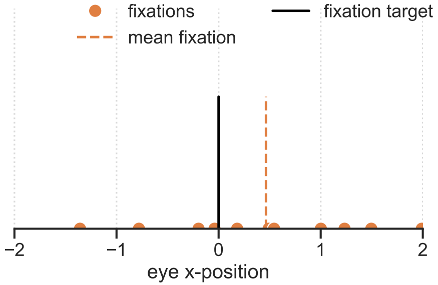
Scaffold of any test

1. Choose a **statistic**
2. Get a **null distribution** of the statistic (distribution of the statistic under the **null hypothesis**)
3. Use the null distribution to a **p-value**
4. Make a decision in favor or against the null hypothesis using the p-value and your acceptable level of type I errors.

William Sealy Gosset (1876 - 1937)

[https://en.wikipedia.org/wiki/William_Sealy_Gosset]

Choose a statistic



$$t = \frac{\hat{\mu} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

$\hat{\mu}$=measured mean, μ_0=target, $\frac{\hat{\sigma}}{\sqrt{n}}$=standard error

Recap: Statistical Tests (introducing one sample t-test)

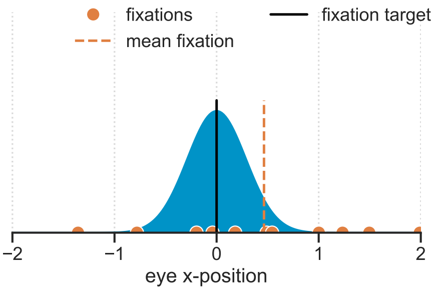
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Eye Tracker

Assume you are writing a program for an eye tracker. You need to determine whether the user fixates on a target at $x = 0$. You get $n = 12$ fixation measures from the eye tracker. You

1. Need to make a decisions whether the user fixated correctly.
2. Are supposed to choose the threshold such that the program only rejects about 5% of actually correct fixations.

Get a null distribution



$$t = \frac{\hat{\mu} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

$\hat{\mu}$=measured mean, μ_0=target, $\frac{\hat{\sigma}}{\sqrt{n}}$=standard error

Recap: Statistical Tests (introducing one sample t-test)

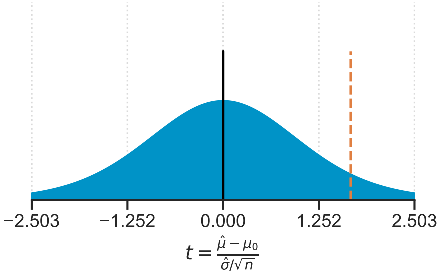
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Eye Tracker

Assume you are writing a program for an eye tracker. You need to determine whether the user fixates on a target at $x = 0$. You get $n = 12$ fixation measures from the eye tracker. You

1. Need to make a decisions whether the user fixated correctly.
2. Are supposed to choose the threshold such that the program only rejects about 5% of actually correct fixations.

Get a null distribution



$$t = \frac{\hat{\mu} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

$\hat{\mu}$=measured mean, μ_0=target, $\frac{\hat{\sigma}}{\sqrt{n}}$=standard error

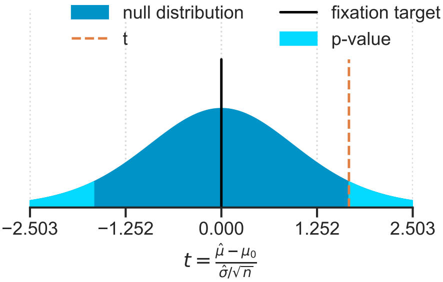
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Eye Tracker

Assume you are writing a program for an eye tracker. You need to determine whether the user fixates on a target at $x = 0$. You get $n = 12$ fixation measures from the eye tracker. You

1. Need to make a decisions whether the user fixated correctly.
2. Are supposed to choose the threshold such that the program only rejects about 5% of actually correct fixations.

Compute a p-value



$$p = P(T \geq t \mid H_0) = 0.065$$

t=measured statistic

T=any random drawn statistic from the null distribution

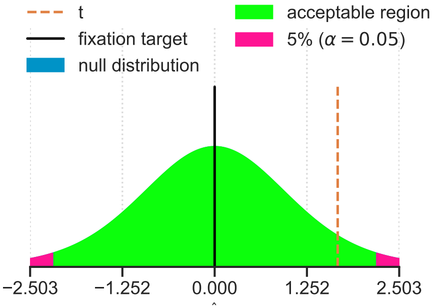
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Eye Tracker

Assume you are writing a program for an eye tracker. You need to determine whether the user fixates on a target at $x = 0$. You get $n = 12$ fixation measures from the eye tracker. You

1. Need to make a decisions whether the user fixated correctly.
2. Are supposed to choose the threshold such that the program only rejects about 5% of actually correct fixations.

Compare p-value with α-level



$$p \leq \alpha = 0.05?$$

Yes Reject H_0

No Do not reject H_0

Another one-sample test

Assume you carry out the following test to determine whether a coin is fair or not:

You throw the coin $n = 3$ times. If the result is either $3\times$ head or $3\times$ tail, you conclude that the coin is not fair.

Answer the following questions (for yourself first):

1. What is the meta-study?

Another one-sample test

Assume you carry out the following test to determine whether a coin is fair or not:

You throw the coin $n = 3$ times. If the result is either $3\times$ head or $3\times$ tail, you conclude that the coin is not fair.

Answer the following questions (for yourself first):

1. What is the meta-study? *Repeated experiments of 3 throws with this the coin.*

Another one-sample test

Assume you carry out the following test to determine whether a coin is fair or not:

You throw the coin $n = 3$ times. If the result is either $3\times$ head or $3\times$ tail, you conclude that the coin is not fair.

Answer the following questions (for yourself first):

1. What is the meta-study? *Repeated experiments of 3 throws with this the coin.*
2. What is the statistic used?

Another one-sample test

Assume you carry out the following test to determine whether a coin is fair or not:

You throw the coin $n = 3$ times. If the result is either $3\times$ head or $3\times$ tail, you conclude that the coin is not fair.

Answer the following questions (for yourself first):

1. What is the meta-study? *Repeated experiments of 3 throws with this the coin.*
2. What is the statistic used? *The number of heads (could also be tails).*

Another one-sample test

Assume you carry out the following test to determine whether a coin is fair or not:

You throw the coin $n = 3$ times. If the result is either $3\times$ head or $3\times$ tail, you conclude that the coin is not fair.

Answer the following questions (for yourself first):

1. What is the meta-study? *Repeated experiments of 3 throws with this the coin.*
2. What is the statistic used? *The number of heads (could also be tails).*
3. What is H_0?

Another one-sample test

Assume you carry out the following test to determine whether a coin is fair or not:

You throw the coin $n = 3$ times. If the result is either $3\times$ head or $3\times$ tail, you conclude that the coin is not fair.

Answer the following questions (for yourself first):

1. What is the meta-study? *Repeated experiments of 3 throws with this the coin.*
2. What is the statistic used? *The number of heads (could also be tails).*
3. What is H_0? *The coin is fair.*

Another one-sample test

Assume you carry out the following test to determine whether a coin is fair or not:

You throw the coin $n = 3$ times. If the result is either $3\times$ head or $3\times$ tail, you conclude that the coin is not fair.

Answer the following questions (for yourself first):

1. What is the meta-study? *Repeated experiments of 3 throws with this the coin.*
2. What is the statistic used? *The number of heads (could also be tails).*
3. What is H_0? *The coin is fair.*
4. What is the Null distribution?

Another one-sample test

Assume you carry out the following test to determine whether a coin is fair or not:

You throw the coin $n = 3$ times. If the result is either $3\times$ head or $3\times$ tail, you conclude that the coin is not fair.

Answer the following questions (for yourself first):

1. What is the meta-study? *Repeated experiments of 3 throws with this the coin.*
2. What is the statistic used? *The number of heads (could also be tails).*
3. What is H_0? *The coin is fair.*
4. What is the Null distribution? *The distribution is binomial*

$$p(k \textit{ heads in } n \textit{ throws}) = \binom{n}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{n-k}$$

Another one-sample test

Assume you carry out the following test to determine whether a coin is fair or not:

You throw the coin $n = 3$ times. If the result is either $3\times$ head or $3\times$ tail, you conclude that the coin is not fair.

Answer the following questions (for yourself first):

1. What is the meta-study? *Repeated experiments of 3 throws with this the coin.*
2. What is the statistic used? *The number of heads (could also be tails).*
3. What is H_0? *The coin is fair.*
4. What is the Null distribution? *The distribution is binomial*

$$p(k \text{ heads in } n \text{ throws}) = \binom{n}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{n-k}$$

5. What is the Type I error of this test?

Another one-sample test

Assume you carry out the following test to determine whether a coin is fair or not:

You throw the coin $n = 3$ times. If the result is either $3\times$ head or $3\times$ tail, you conclude that the coin is not fair.

Answer the following questions (for yourself first):

1. What is the meta-study? *Repeated experiments of 3 throws with this the coin.*
2. What is the statistic used? *The number of heads (could also be tails).*
3. What is H_0? *The coin is fair.*
4. What is the Null distribution? *The distribution is binomial*

$$p(k \textit{ heads in } n \textit{ throws}) = \binom{n}{k}\left(\frac{1}{2}\right)^{k}\left(\frac{1}{2}\right)^{n-k}$$

5. What is the Type I error of this test? $p(HHH|H_0) + p(TTT|H_0) = \frac{2}{8}$

Statistical Power

What is statistical power?

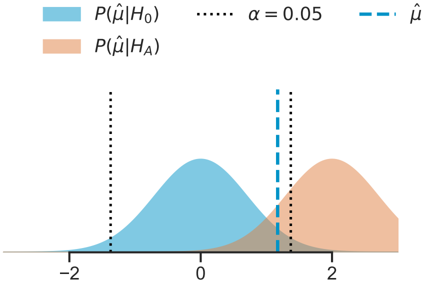
Assume we reject H_0 if $p \leq 0.05 = \alpha$.

	$p > 0.05$	$p \leq 0.05$
H_0 true		
H_A true		

This happens with probability

	$p > 0.05$	$p \leq 0.05$
H_0 true		
H_A true		

$\hat{\mu}$ comes from H_0



What is statistical power?

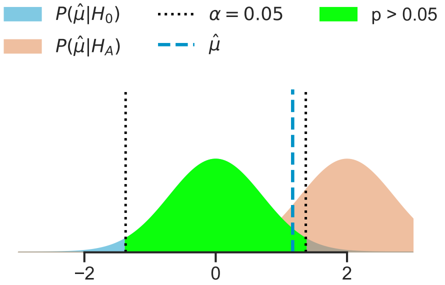
Assume we reject H_0 if $p \leq 0.05 = \alpha$.

	$p > 0.05$	$p \leq 0.05$
H_0 true	true negative	
H_A true		

This happens with probability

	$p > 0.05$	$p \leq 0.05$
H_0 true	$0.95 = 1 - \alpha$	
H_A true		

$\hat{\mu}$ comes from H_0



What is statistical power?

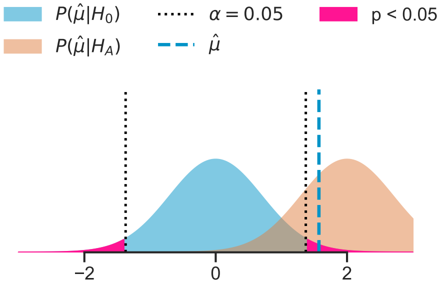
Assume we reject H_0 if $p \leq 0.05 = \alpha$.

	$p > 0.05$	$p \leq 0.05$
H_0 true	true negative	false positive
H_A true		

This happens with probability

	$p > 0.05$	$p \leq 0.05$
H_0 true	$0.95 = 1 - \alpha$	$0.05 = \alpha$
H_A true		

$\hat{\mu}$ comes from H_0



What is statistical power?

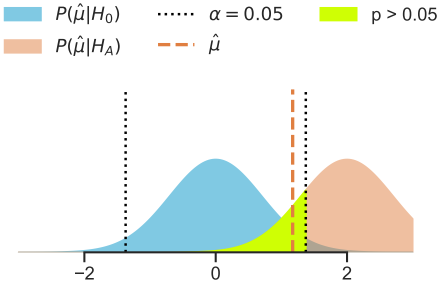
Assume we reject H_0 if $p \leq 0.05 = \alpha$.

	$p > 0.05$	$p \leq 0.05$
H_0 true	true negative	false positive
H_A true	false negative	

This happens with probability

	$p > 0.05$	$p \leq 0.05$
H_0 true	$0.95 = 1 - \alpha$	$0.05 = \alpha$
H_A true	β	

$\hat{\mu}$ comes from H_A



What is statistical power?

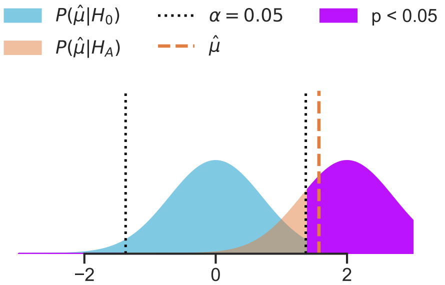
Assume we reject H_0 if $p \leq 0.05 = \alpha$.

	$p > 0.05$	$p \leq 0.05$
H_0 true	true negative	false positive
H_A true	false negative	true positive

This happens with probability

	$p > 0.05$	$p \leq 0.05$
H_0 true	$0.95 = 1 - \alpha$	$0.05 = \alpha$
H_A true	β	$1 - \beta$=power

$\hat{\mu}$ comes from H_A



What is statistical power?

Assume we reject H_0 if $p \leq 0.05 = \alpha$.

	$p > 0.05$	$p \leq 0.05$
H_0 true	true negative	false positive
H_A true	false negative	true positive

This happens with probability

	$p > 0.05$	$p \leq 0.05$
H_0 true	$0.95 = 1 - \alpha$	$0.05 = \alpha$
H_A true	β	$1 - \beta$=**power**

Statistical errors types

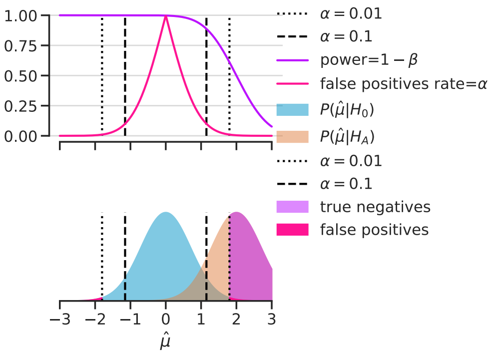
1. False positives are called **type I** errors. The error probability/rate is denoted with α.
2. False negatives are called **type II** errors. The error probability/rate is denoted with β.

Power

Power is the probability of accepting H_A (rejecting H_0) if H_A is true. It is $1 - \beta$. The higher the power the better.

How can we control power?

1. We can change the decision threshold: The higher α the higher $1-\beta$

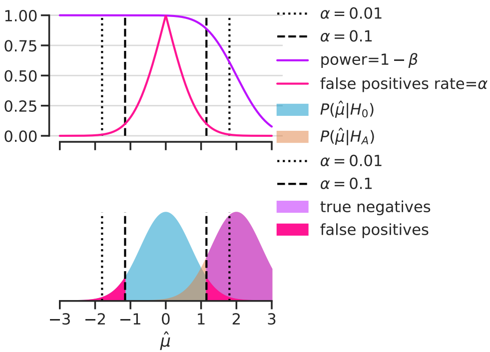


- Low false positives $\alpha = 0.01$
- Low power $1 - \beta \approx 0.61\%$

How can we control power?

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

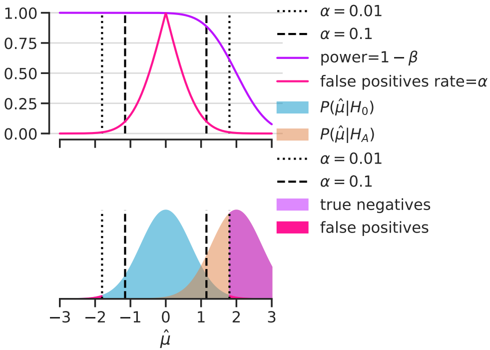
1. We can change the decision threshold: The higher α the higher $1 - \beta$



- High false positives $\alpha = 0.1$
- High power $1 - \beta \approx 0.89\%$

How can we control power?

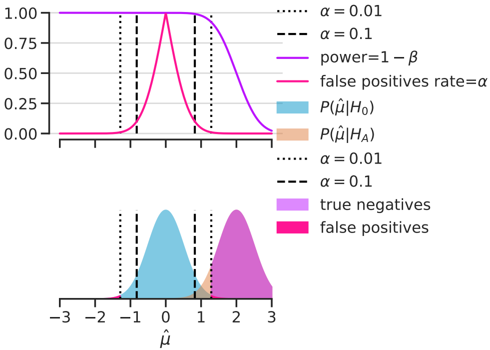
1. We can change the decision threshold: The higher α the higher $1-\beta$
2. You can increase n, the number of data points.



High n allows you to **decrease α** and **increase power $1-\beta$** at the same time.

How can we control power?

1. We can change the decision threshold: The higher α the higher $1-\beta$
2. You can increase n, the number of data points.

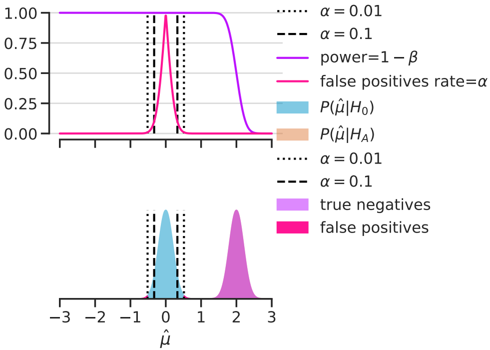


High n allows you to **decrease α** and **increase power $1-\beta$** at the same time.

How can we control power?

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

1. We can change the decision threshold: The higher α the higher $1-\beta$
2. You can increase n, the number of data points.



High n allows you to **decrease** α and **increase power** $1-\beta$ at the same time.

How can we control power?

1. We can change the decision threshold: The higher α the higher $1 - \beta$
2. You can increase n, the number of data points.
3. You should pick n to have sufficient power via an **a prior** power analysis.

- For example Lehr's rule of thumb says that for 80% power ($1 - \beta = 0.8$) at $\alpha = 0.05$ the number of data points should be

$$n = 16 \frac{s^2}{d^2}$$

  in a two-sample t-test where $d = \mu_{H_A} - \mu_{H_0}$ is the difference you want to detect and s^2 is an estimation of the population variance.
- You can find similar estimates in statistic books
- You could also simulate.
- Power analysis **needs assumptions about your statistic under H_A**

How can we control power?

1. We can change the decision threshold: The higher α the higher $1-\beta$
2. You can increase n, the number of data points.
3. You should pick n to have sufficient power via an **a prior** power analysis.
4. You should use a higher powered test.

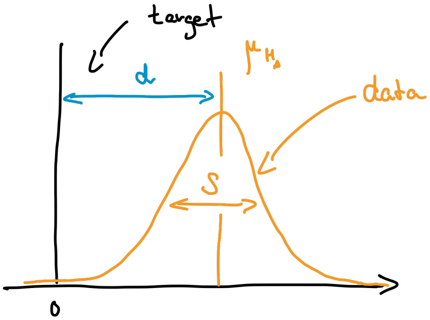
- **Rule of thumb:** The more "structure" (see later) a test can assume about your data, the more powerful.
- For instance: You could use an two-sample t-test on paired data (just forget that they come in pairs), but that would decrease the power.

Effect size

- In Lehr's rule we saw n estimated as a function of $\theta = \frac{d}{s}$

$$n = 16\frac{s^2}{d^2} = \frac{16}{\theta^2}$$

where $d = \mu_{H_A} - \mu_{H_0}$ is the difference you want to detect and s^2 is an estimation of the population variance.



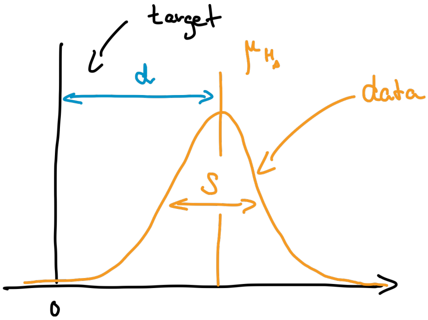
Effect size

- In Lehr's rule we saw n estimated as a function of $\theta = \frac{d}{s}$

$$n = 16\frac{s^2}{d^2} = \frac{16}{\theta^2}$$

where $d = \mu_{H_A} - \mu_{H_0}$ is the difference you want to detect and s^2 is an estimation of the population variance.

- θ is a measure of **effect size**.



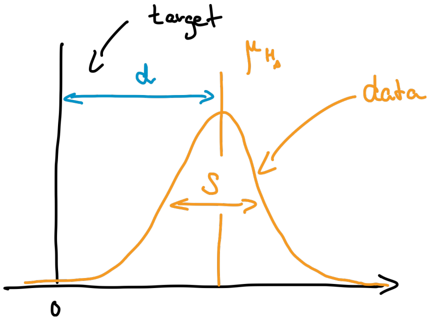
Effect size

- In Lehr's rule we saw n estimated as a function of $\theta = \frac{d}{s}$

$$n = 16\frac{s^2}{d^2} = \frac{16}{\theta^2}$$

where $d = \mu_{H_A} - \mu_{H_0}$ is the difference you want to detect and s^2 is an estimation of the population variance.

- θ is a measure of **effect size**.
- It measures the **size of the difference standardized by the variation in the data**.



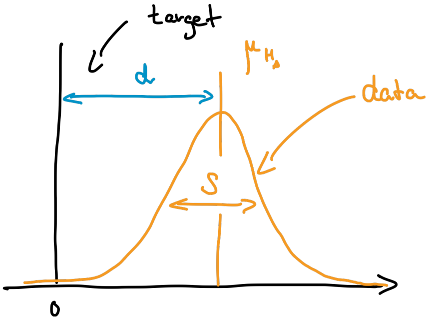
Effect size

- In Lehr's rule we saw n estimated as a function of $\theta = \frac{d}{s}$

$$n = 16\frac{s^2}{d^2} = \frac{16}{\theta^2}$$

where $d = \mu_{H_A} - \mu_{H_0}$ is the difference you want to detect and s^2 is an estimation of the population variance.

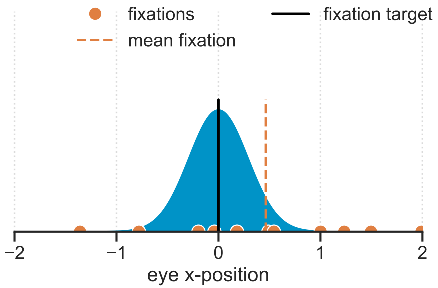
- θ is a measure of **effect size**.
- It measures the **size of the difference standardized by the variation in the data**.
- Effect sizes are important measures to report for the magnitude of an effect since tests will detect even tiny differences if n is large enough.



Eye Tracker

Assume you are writing a program for an eye tracker. You need to determine whether the user fixates on a target at $x = 0$. You get $n = 12$ fixation measures from the eye tracker. You

1. Need to make a decisions whether the user fixated correctly.
2. Are supposed to choose the threshold such that the program only rejects about 5% of actually correct fixations.



Eye tracker continued

Eye Tracker

Now your boss wants the decisions to be made faster and asks you if you could decrease n. Since you are a diligent data scientist, you check what power that test would have.

You assume $\alpha = 0.05$, a "typical" std of $s = 1$ for the fixations and a minimally detectable difference of $\mu_{H_A} - \mu_{H_0} = 1$.

You want to use a one sample t-test and you find a function to estimate the power in the library "'statsmodels".

Power analysis

```
from statsmodels.stats import power
pow = power.TTestIndPower()
power = pow.power(effect_size, n, alpha)
```

Checking the documentation you find that effect size here means difference in mean divided by the standard deviation

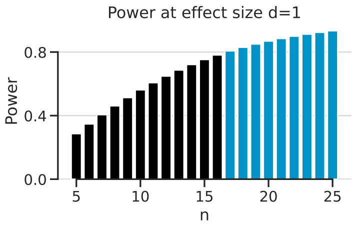
$$\theta = \frac{d}{s} = \frac{1}{1} = 1$$

Eye tracker continued

Eye Tracker

Now your boss wants the decisions to be made faster and asks you if you could decrease n. Since you are a diligent data scientist, you check what power that test would have.

You assume $\alpha = 0.05$, a "typical" std of $s = 1$ for the fixations and a minimally detectable difference of $\mu_{H_A} - \mu_{H_0} = 1$.



The power for $n = 12$ is only about 65%!

65% is not enough

What could you do?

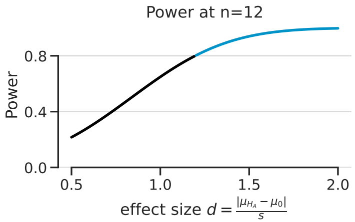
Eye tracker continued

Eye Tracker

Now your boss wants the decisions to be made faster and asks you if you could decrease n. Since you are a diligent data scientist, you check what power that test would have.

You assume $\alpha = 0.05$, a "typical" std of $s = 1$ for the fixations and a minimally detectable difference of $\mu_{H_A} - \mu_{H_0} = 1$.

You can in crease minimal detectable difference.



For $n = 12$ the minimal effect size you can detect with 80% power and $\alpha = 0.05$ is

$$d \approx 1.2$$

. For fewer n you need to be ok with an even bigger d.

Statistical power (German “Trennschärfe”)

- Is the probability to correctly reject H_0 if H_A is true.
- It is denoted by $1 - \beta$.
- For fixed n false positives (you want that low) and power (you want that high) are correlated (that’s bad).
- You can increase power by increasing n.
- Tests differ in power, you should use the one with better power.
- You should pick your n before a study to have sufficient power (power analysis, typical value 80%).

Statistical power (German “Trennschärfe”)

- Is the probability to correctly reject H_0 if H_A is true.
- It is denoted by $1 - \beta$.
- For fixed n false positives (you want that low) and power (you want that high) are correlated (that’s bad).
- You can increase power by increasing n.
- Tests differ in power, you should use the one with better power.
- You should pick your n before a study to have sufficient power (power analysis, typical value 80%).

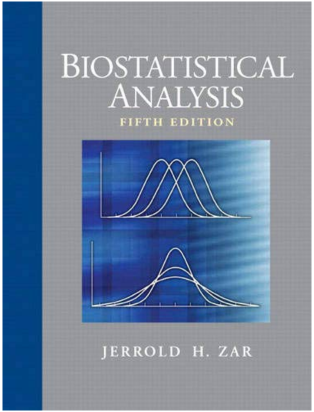
Effect size

- Effect size measures the size of a difference as a function of the “noise” in the data.
- It’s an important measure to report, since even small “meaningless” differences can be detected with tests if n is large enough.
- Power analysis is often done as a function of the effect size.

How to choose a statistical test?

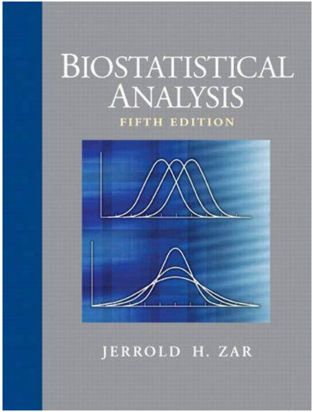
Look up tests

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN



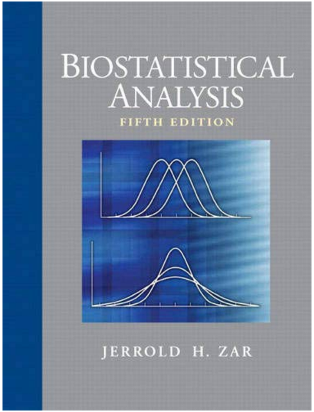
1. What is the data type?

Look up tests



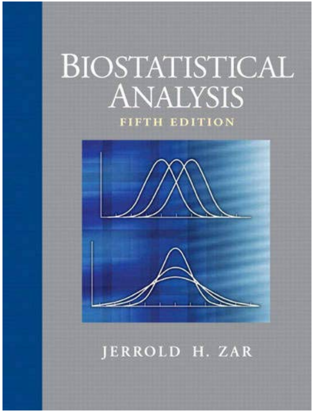
1. What is the data type?
2. Is the data normally distributed or not (for interval/ratio type; assumption of many tests)?

Look up tests



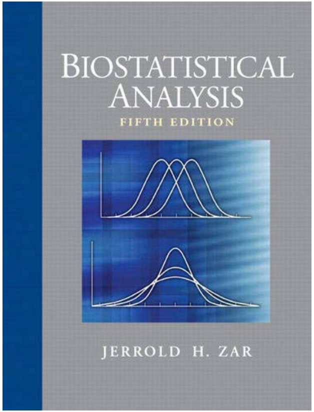
1. What is the data type?
2. Is the data normally distributed or not (for interval/ratio type; assumption of many tests)?
3. How many groups do I have?

Look up tests



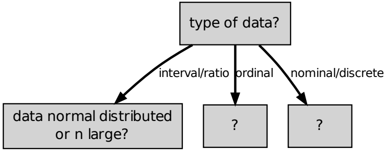
1. What is the data type?
2. Is the data normally distributed or not (for interval/ratio type; assumption of many tests)?
3. How many groups do I have?
4. Do I need to test for deviations in one or two directions (one-tailed/-sided or two-tailed/-sided tests)?

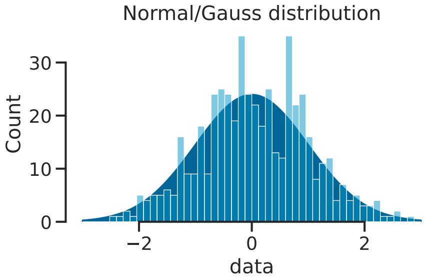
Look up tests



1. What is the data type?
2. Is the data normally distributed or not (for interval/ratio type; assumption of many tests)?
3. How many groups do I have?
4. Do I need to test for deviations in one or two directions (one-tailed/-sided or two-tailed/-sided tests)?
5. Is the data paired?

Question 1: What data type do I have and is it normal?

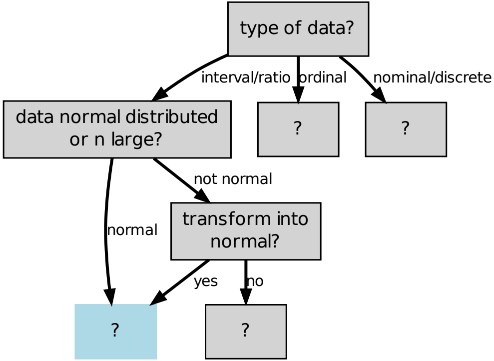


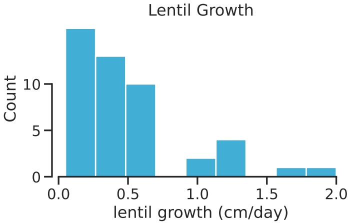


Sidenote

- Normality can be checked with a QQ-plot
- If n is large and the variance of the data distribution is finite, the central limit theorem guarantees normality for "summed statistics".

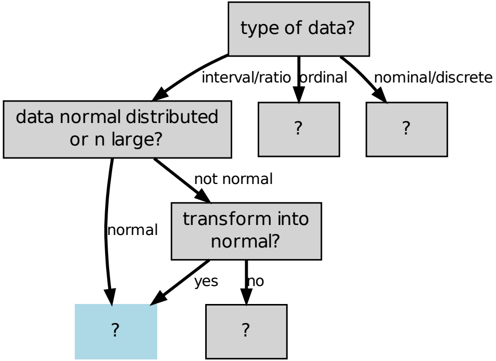
Is the data normally distributed?

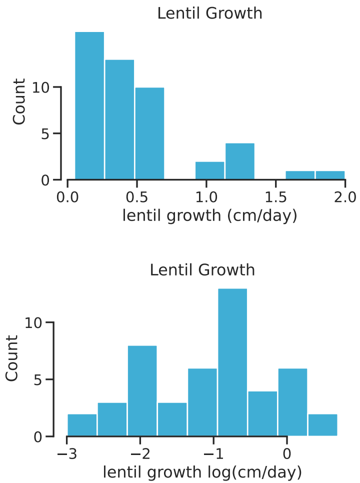




[Statistics for the Life Sciences (5th Ed.)]

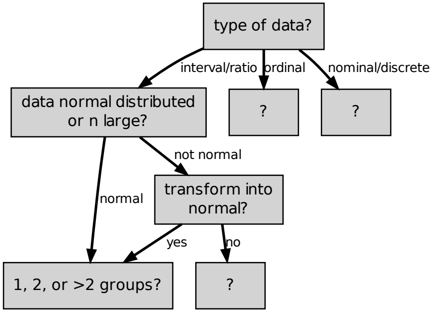
Is the data normally distributed?





[Statistics for the Life Sciences (5th Ed.)]

How many groups?



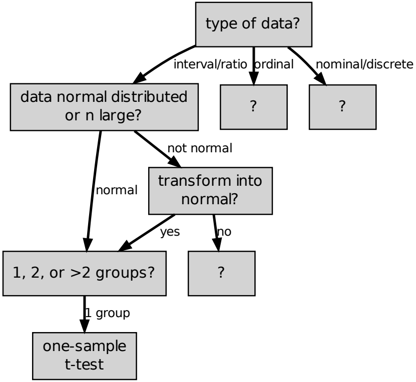
Menstrual Cycle

The data set contains the lengths of the menstrual cycles in a random sample of 15 women. Assume we want to the hypothesis that the mean length of human menstrual cycle is equal to a lunar month (29.5 days).

Use the one sample t-test from above

- The data contains **1 group**.

How many groups?



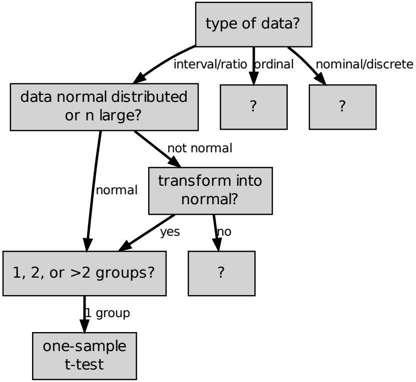
Menstrual Cycle

The data set contains the lengths of the menstrual cycles in a random sample of 15 women. Assume we want to the hypothesis that the mean length of human menstrual cycle is equal to a lunar month (29.5 days).

Use the one sample t-test from above

- The data contains **1 group**.
- What is H_0? What is H_A?

How many groups?



Menstrual Cycle

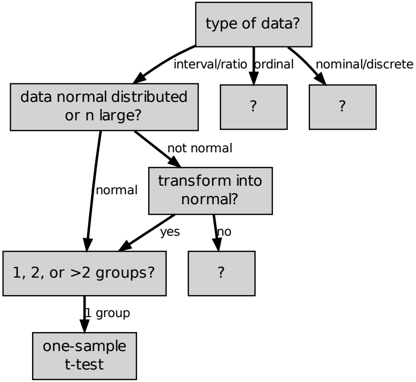
The data set contains the lengths of the menstrual cycles in a random sample of 15 women. Assume we want to the hypothesis that the mean length of human menstrual cycle is equal to a lunar month (29.5 days).

Use the one sample t-test from above

- The data contains **1 group**.
- What is H_0? What is H_A?

$H_0 : \hat{\mu} = 29.5$, $H_A : \hat{\mu} \neq 29.5$

How many groups?



Menstrual Cycle

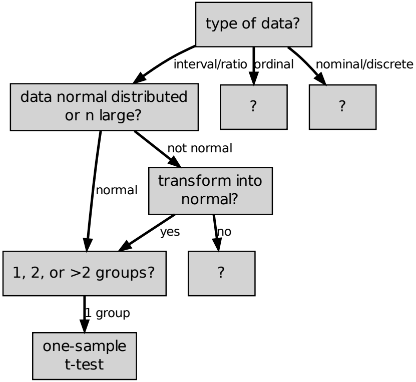
The data set contains the lengths of the menstrual cycles in a random sample of 15 women. Assume we want to the hypothesis that the mean length of human menstrual cycle is equal to a lunar month (29.5 days).

Use the one sample t-test from above

- The data contains **1 group**.
- What is H_0? What is H_A?

  $H_0 : \hat{\mu} = 29.5$, $H_A : \hat{\mu} \neq 29.5$
- What is the test statistic?

How many groups?



Menstrual Cycle

The data set contains the lengths of the menstrual cycles in a random sample of 15 women. Assume we want to the hypothesis that the mean length of human menstrual cycle is equal to a lunar month (29.5 days).

Use the one sample t-test from above

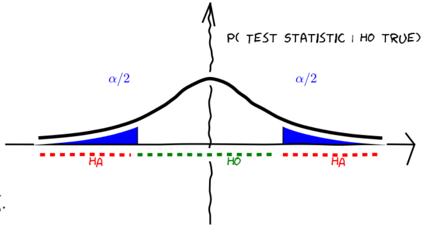
- The data contains **1 group**.
- What is H_0? What is H_A?

  $H_0 : \hat{\mu} = 29.5$, $H_A : \hat{\mu} \neq 29.5$
- What is the test statistic?

  $t = \frac{\hat{\mu} - 29.5}{\hat{\sigma}/\sqrt{n}}$

Do I need to test for deviations in both directions?

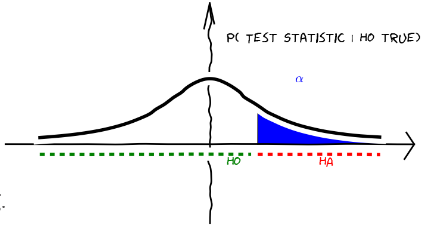
two tailed test



e.g.

- $H_0 : \mu = 0$
- $H_A : \mu \neq 0$

one tailed test



e.g.

- $H_0 : \mu = 0$
- $H_A : \mu > 0$
- $\hat{\mu} < 0$ must directly imply $\hat{\mu}$ came from $P(\hat{\mu}|H_0)$
- if that is not the case, using one-tailed is cheating

Is the data paired?

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Chirping

A scientist conducted a study of how often her pet parakeet chirps. She recorded the number of distinct chirps the parakeet made in a 30-minute period, sometimes when the room was silent and sometimes when music was playing. The data are shown in the following table. Test whether the bird changes its chirping behavior when music is playing.

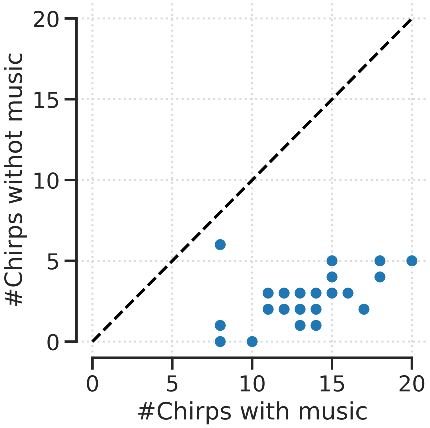
	Chirps in 30 minutes		
Day	With music	Without music	Difference
1	12	3	9
2	14	1	13
3	11	2	9
4	13	1	12
5	20	5	15
6	14	3	11
7	10	0	10
8	12	2	10
9	8	6	2
10	13	3	10
11	14	2	12
12	15	4	11
13	12	3	9
14	13	2	11
15	8	0	8
16	18	5	13
17	15	3	12
18	12	2	10
19	17	2	15
20	15	4	11
21	11	3	8

[Statistics for the Life Science (5th Ed.)]

Is the data paired?

Chirping

A scientist conducted a study of how often her pet parakeet chirps. She recorded the number of distinct chirps the parakeet made in a 30-minute period, sometimes when the room was silent and sometimes when music was playing. The data are shown in the following table. Test whether the bird changes its chirping behavior when music is playing.



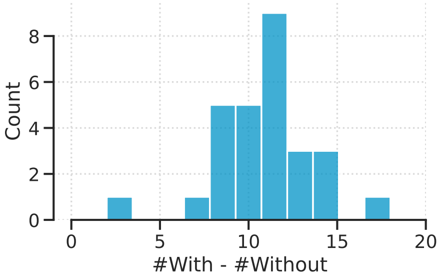
[Statistics for the Life Science (5th Ed.)]

Is the data paired?

Chirping

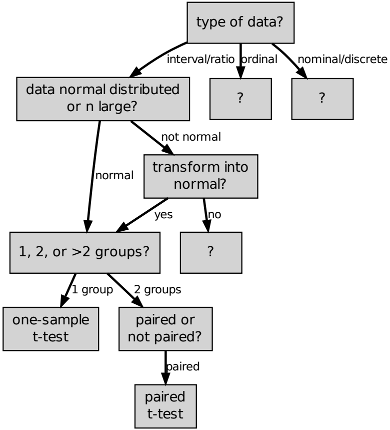
A scientist conducted a study of how often her pet parakeet chirps. She recorded the number of distinct chirps the parakeet made in a 30-minute period, sometimes when the room was silent and sometimes when music was playing. The data are shown in the following table. Test whether the bird changes its chirping behavior when music is playing.

Paired t-test
=one sample t-test against 0 on the difference

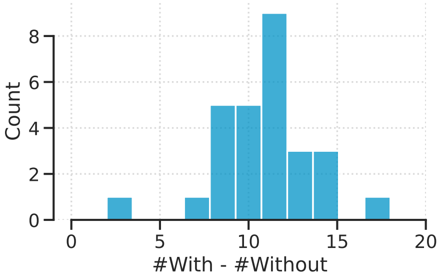


[Statistics for the Life Science (5th Ed.)]

Is the data paired?



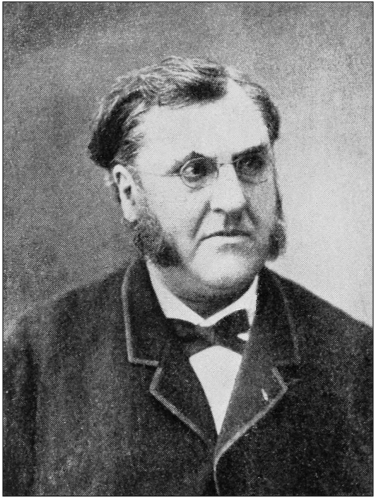
Paired t-test
=one sample t-test against 0 on the difference



[Statistics for the Life Science (5th Ed.)]

2 groups, but not paired (independent)

Paul Topinard (1830 - 1911)

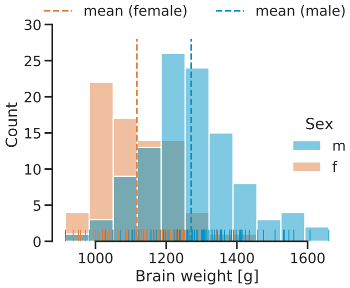


[https://en.wikipedia.org/wiki/Paul_Topinard]

Brain Weights (permutation test)

In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. The dataset contains brain weights of males and females. It consists of (i) two samples (male/female) which are (ii) not paired. We want to test whether the mean brain weights of males and females are different.

2 groups, but not paired (independent)



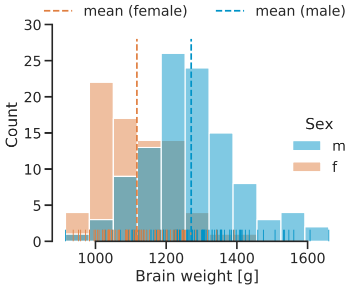
Brain Weights (permutation test)

In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. The dataset contains brain weights of males and females. It consists of (i) two samples (male/female) which are (ii) not paired. We want to test whether the mean brain weights of males and females are different.

[Statistics for the Life Science (5th Ed.)]

- What could we use as statistic?

2 groups, but not paired (independent)



Brain Weights (permutation test)

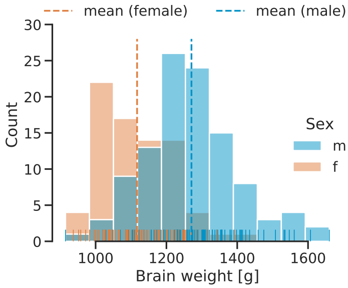
In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. The dataset contains brain weights of males and females. It consists of (i) two samples (male/female) which are (ii) not paired. We want to test whether the mean brain weights of males and females are different.

[Statistics for the Life Science (5th Ed.)]

- What could we use as statistic?

  the difference in the means

2 groups, but not paired (independent)



Brain Weights (permutation test)

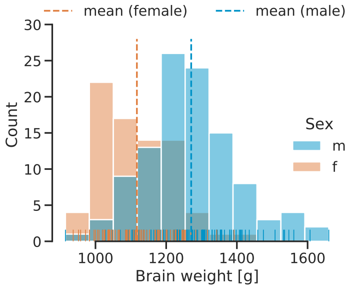
In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. The dataset contains brain weights of males and females. It consists of (i) two samples (male/female) which are (ii) not paired. We want to test whether the mean brain weights of males and females are different.

[Statistics for the Life Science (5th Ed.)]

- What could we use as statistic?

  the difference in the means
- What would be H_0?

2 groups, but not paired (independent)



Brain Weights (permutation test)

In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. The dataset contains brain weights of males and females. It consists of (i) two samples (male/female) which are (ii) not paired. We want to test whether the mean brain weights of males and females are different.

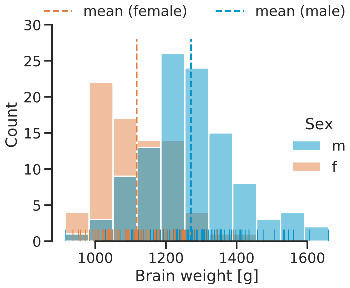
[Statistics for the Life Science (5th Ed.)]

- What could we use as statistic?

  the difference in the means
- What would be H_0?

  the difference is zero

2 groups, but not paired (independent)



Brain Weights (permutation test)

In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. The dataset contains brain weights of males and females. It consists of (i) two samples (male/female) which are (ii) not paired. We want to test whether the mean brain weights of males and females are different.

[Statistics for the Life Science (5th Ed.)]

- What could we use as statistic?

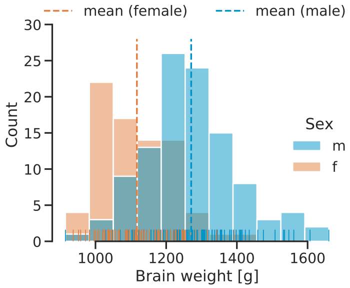
  the difference in the means

- What would be H_0?

  the difference is zero

- Think about a way to generate an estimate of the Null distribution with Python?

2 groups, but not paired (independent)



Brain Weights (permutation test)

In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. The dataset contains brain weights of males and females. It consists of (i) two samples (male/female) which are (ii) not paired. We want to test whether the mean brain weights of males and females are different.

[Statistics for the Life Science (5th Ed.)]

- What could we use as statistic?

  the difference in the means

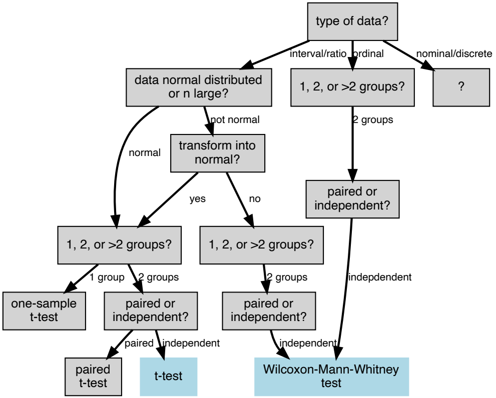
- What would be H_0?

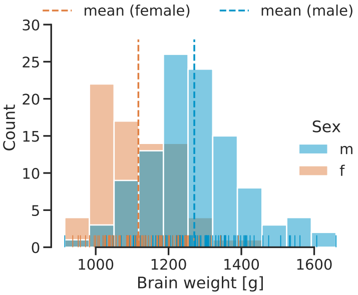
  the difference is zero

- Think about a way to generate an estimate of the Null distribution with Python?

  Permutation test: Shuffle the labels, compute difference in means, repeat ...

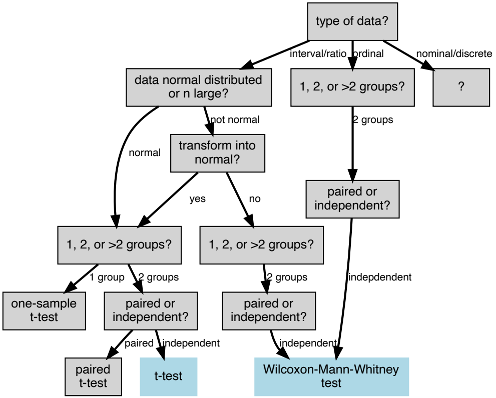
2 groups, but not paired (independent)

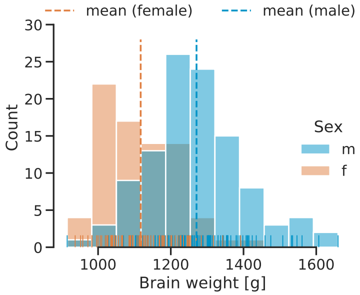




- There is **two-sample independent t-test** is the parametric test for this dataset.
- If normality does not hold, you can use the **Wilcoxon-Mann-Whitney test**

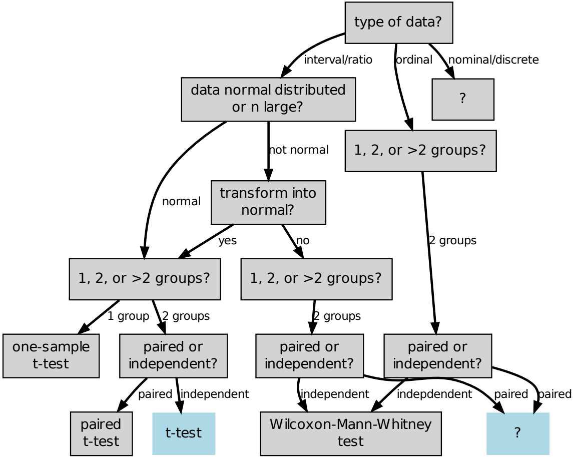
2 groups, but not paired (independent)



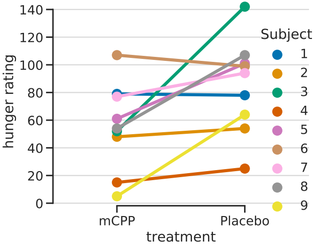


- There is **two-sample independent t-test** is the parametric test for this dataset.
- If normality does not hold, you can use the **Wilcoxon-Mann-Whitney test**

More paired data



GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

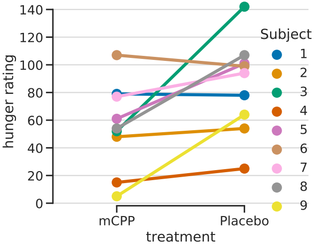


Hunger Rating

During a weight loss study each of nine subjects was given either the active drug m-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study, the subjects were asked to rate how hungry there were at the end of each 2-week period.

[Statistics for the Life Science (5th Ed.)]

Sign-test



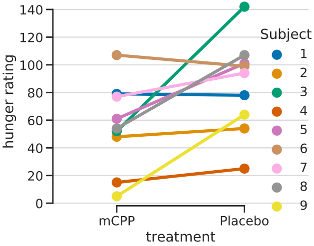
Hunger Rating

During a weight loss study each of nine subjects was given either the active drug m-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study, the subjects were asked to rate how hungry there were at the end of each 2-week period.

[Statistics for the Life Science (5th Ed.)]

- Why is it good that the data is paired and does not have two independent groups?

Sign-test



Hunger Rating

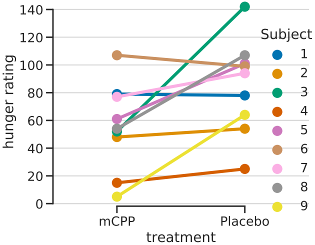
During a weight loss study each of nine subjects was given either the active drug m-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study, the subjects were asked to rate how hungry there were at the end of each 2-week period.

[Statistics for the Life Science (5th Ed.)]

- Why is it good that the data is paired and does not have two independent groups?

  Each person could have a different hunger "baseline".

Sign-test



Hunger Rating

During a weight loss study each of nine subjects was given either the active drug m-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study, the subjects were asked to rate how hungry there were at the end of each 2-week period.

[Statistics for the Life Science (5th Ed.)]

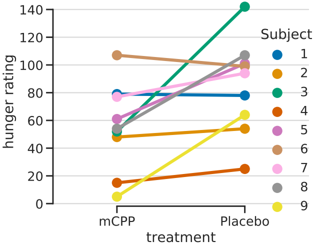
- Why is it good that the data is paired and does not have two independent groups?

  Each person could have a different hunger "baseline".

- What data types are involved?

Sign-test

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN



Hunger Rating

During a weight loss study each of nine subjects was given either the active drug m-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study, the subjects were asked to rate how hungry there were at the end of each 2-week period.

[Statistics for the Life Science (5th Ed.)]

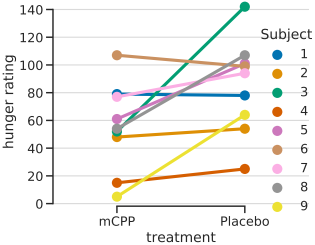
- Why is it good that the data is paired and does not have two independent groups?

  Each person could have a different hunger "baseline".

- What data types are involved?

  Ordinal (hunger rating), categorial (treatment, Subject)

Sign-test



Hunger Rating

During a weight loss study each of nine subjects was given either the active drug m-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study, the subjects were asked to rate how hungry there were at the end of each 2-week period.

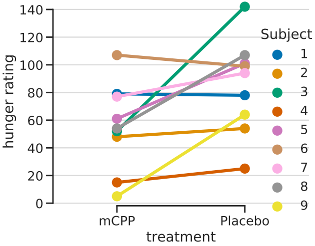
[Statistics for the Life Science (5th Ed.)]

- Why is it good that the data is paired and does not have two independent groups?

  Each person could have a different hunger "baseline".
- What data types are involved?

  Ordinal (hunger rating), categorial (treatment, Subject)
- What would be a good statistic to measure the difference between the treatments?

Sign-test



Hunger Rating

During a weight loss study each of nine subjects was given either the active drug m-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study, the subjects were asked to rate how hungry there were at the end of each 2-week period.

[Statistics for the Life Science (5th Ed.)]

- Why is it good that the data is paired and does not have two independent groups?

  Each person could have a different hunger “baseline”.

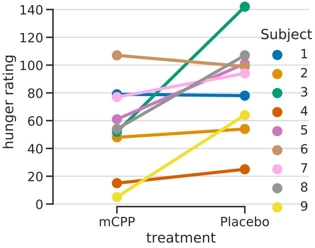
- What data types are involved?

  Ordinal (hunger rating), categorial (treatment, Subject)

- What would be a good statistic to measure the difference between the treatments?

  Count how many times “mCPP > Placebo”. Why is “mCPP-Placebo” not great?

Sign-test



Hunger Rating

During a weight loss study each of nine subjects was given either the active drug m-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study, the subjects were asked to rate how hungry there were at the end of each 2-week period.

[Statistics for the Life Science (5th Ed.)]

- Why is it good that the data is paired and does not have two independent groups?

  Each person could have a different hunger "baseline".

- What data types are involved?

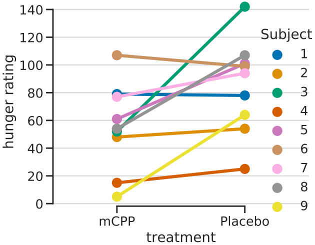
  Ordinal (hunger rating), categorial (treatment, Subject)

- What would be a good statistic to measure the difference between the treatments?

  Count how many times "mCPP > Placebo". Why is "mCPP-Placebo" not great?

- What is H_0?

Sign-test



Hunger Rating

During a weight loss study each of nine subjects was given either the active drug m-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study, the subjects were asked to rate how hungry there were at the end of each 2-week period.

[Statistics for the Life Science (5th Ed.)]

- Why is it good that the data is paired and does not have two independent groups?

  Each person could have a different hunger "baseline".

- What data types are involved?

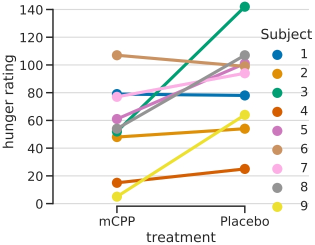
  Ordinal (hunger rating), categorial (treatment, Subject)

- What would be a good statistic to measure the difference between the treatments?

  Count how many times "mCPP > Placebo". Why is "mCPP-Placebo" not great?

- What is H_0?

  "$>$" occurs as often as "$\leq$"



Hunger Rating

During a weight loss study each of nine subjects was given either the active drug m-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study, the subjects were asked to rate how hungry there were at the end of each 2-week period.

[Statistics for the Life Science (5th Ed.)]

- Why is it good that the data is paired and does not have two independent groups?

  Each person could have a different hunger "baseline".

- What data types are involved?

  Ordinal (hunger rating), categorial (treatment, Subject)

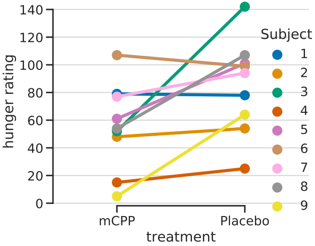
- What would be a good statistic to measure the difference between the treatments?

  Count how many times "mCPP > Placebo". Why is "mCPP-Placebo" not great?

- What is H_0?

  "$>$" occurs as often as "$\leq$"

- How could we generate a null distribution in Python?



Hunger Rating

During a weight loss study each of nine subjects was given either the active drug m-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study, the subjects were asked to rate how hungry there were at the end of each 2-week period.

[Statistics for the Life Science (5th Ed.)]

- Why is it good that the data is paired and does not have two independent groups?

  Each person could have a different hunger "baseline".

- What data types are involved?

  Ordinal (hunger rating), categorial (treatment, Subject)

- What would be a good statistic to measure the difference between the treatments?

  Count how many times "mCPP > Placebo". Why is "mCPP-Placebo" not great?

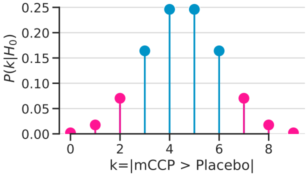
- What is H_0?

  "$>$" occurs as often as "$\leq$"

- How could we generate a null distribution in Python?

  Permutation test: Repeatedly swap the treatment labels per subject.

Sign-test



Hunger Rating

During a weight loss study each of nine subjects was given either the active drug m-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study, the subjects were asked to rate how hungry there were at the end of each 2-week period.

[Statistics for the Life Science (5th Ed.)]

- Why is it good that the data is paired and does not have two independent groups?

  Each person could have a different hunger "baseline".
- What data types are involved?

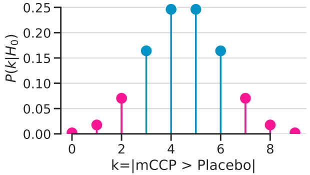
  Ordinal (hunger rating), categorial (treatment, Subject)
- What would be a good statistic to measure the difference between the treatments?

  Count how many times "mCPP > Placebo". Why is "mCPP-Placebo" not great?
- What is H_0?

  "$>$" occurs as often as "$\leq$"
- How could we generate a null distribution in Python?

  Permutation test: Repeatedly swap the treatment labels per subject.
- p-value: Proportion of simulated experiments show $7, 8, 9$ "$>$" or $0, 1, 2$ "$\leq$".

Sign-test



Hunger Rating

During a weight loss study each of nine subjects was given either the active drug m-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study, the subjects were asked to rate how hungry there were at the end of each 2-week period.

[Statistics for the Life Science (5th Ed.)]

- Why is it good that the data is paired and does not have two independent groups?

  Each person could have a different hunger "baseline".

- What data types are involved?

  Ordinal (hunger rating), categorial (treatment, Subject)

- What would be a good statistic to measure the difference between the treatments?

  Count how many times "mCPP > Placebo". Why is "mCPP-Placebo" not great?

- What is H_0?

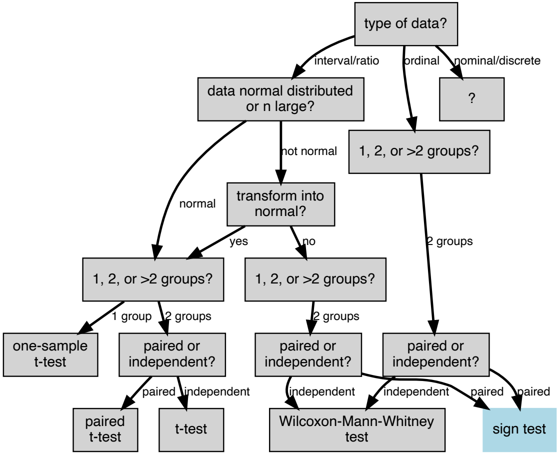
  "$>$" occurs as often as "$\leq$"

- How could we generate a null distribution in Python?

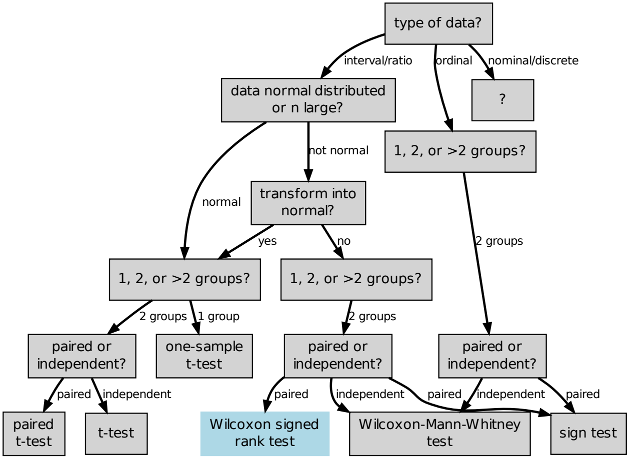
  Permutation test: Repeatedly swap the treatment labels per subject.

- p-value: Proportion of simulated experiments show 7, 8, 9 "$>$" or 0, 1, 2 "$\leq$".
- Analytical alternative: Binomial distribution (repeated coin flips) $\rightarrow$ sign test ($p \approx 0.1797$).

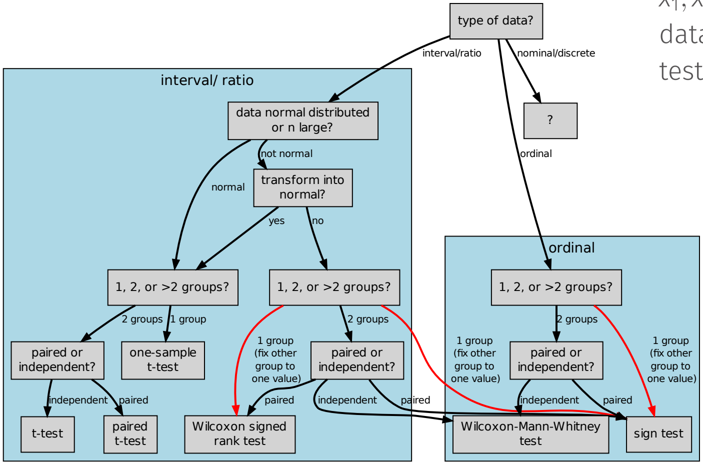
Sign test: For paired data that can be ordinal



Wilcoxon sign rank test: For paired interval data that is not normal

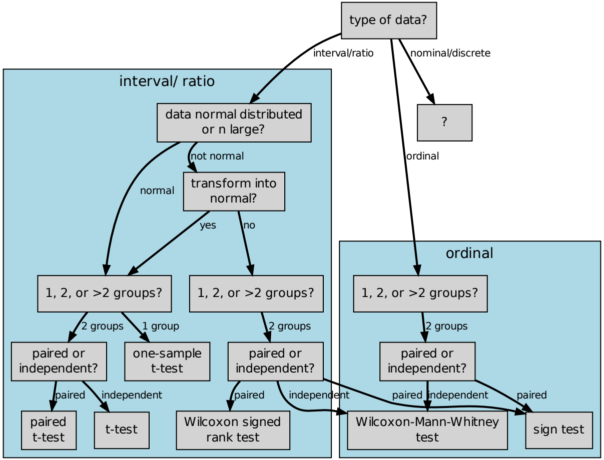


Convert paired test into one sample tests



$x_1, x_2, ..., x_n$ becomes "paired" data $(x_1, a), (x_2, a), ..., (x_n, a)$ to test against a.

Nominal/categorial/discrete data



Migraine Surgery

Patients who suffered from moderate to severe migraine headache took part in a double-blind clinical trial to assess an experimental surgery. A group of 75 patients were randomly assigned to receive either the real surgery on migraine trigger sites ($n = 49$) or a sham surgery ($n = 26$) in which an incision was made but no further procedure was performed. The surgeons hoped that patients would experience "a substantial reduction in migraine headaches" which we will label as "success." It the real surgery related to success?

[Statistics for the Life Science (5th Ed.)]

- The data is categorical. The table summarizes it.

χ^2-test: Testing for dependency between categorical variables

Table 10.2.1 Observed frequencies for migraine study

	Surgery		
	Real	Sham	Total
Success	41	15	56
No success	8	11	19
Total	49	26	75

Migraine Surgery

Patients who suffered from moderate to severe migraine headache took part in a double-blind clinical trial to assess an experimental surgery. A group of 75 patients were randomly assigned to receive either the real surgery on migraine trigger sites (n = 49) or a sham surgery (n = 26) in which an incision was made but no further procedure was performed. The surgeons hoped that patients would experience "a substantial reduction in migraine headaches" which we will label as "success." It the real surgery related to success?

[Statistics for the Life Science (5th Ed.)]

- The data is categorical. The table summarizes it.
- H_0 : success/no success is indepenent of sham/real surgery

χ^2-test: Testing for dependency between categorical variables

Table 10.2.1 Observed frequencies for migraine study

	Surgery: Real	Surgery: Sham	Total
Success	41	15	56
No success	8	11	19
Total	49	26	75

Table 10.2.2 Observed and expected frequencies for migraine study

	Surgery: Real	Surgery: Sham	Total
Success	41 (36.59)	15 (19.41)	56
No success	8 (12.41)	11 (6.59)	19
Total	49	26	75

Migraine Surgery

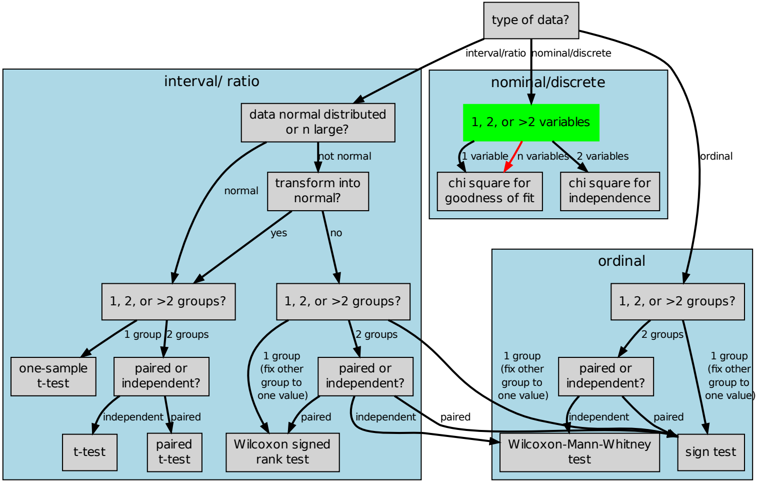
Patients who suffered from moderate to severe migraine headache took part in a double-blind clinical trial to assess an experimental surgery. A group of 75 patients were randomly assigned to receive either the real surgery on migraine trigger sites (n = 49) or a sham surgery (n = 26) in which an incision was made but no further procedure was performed. The surgeons hoped that patients would experience "a substantial reduction in migraine headaches" which we will label as "success." It the real surgery related to success?

[Statistics for the Life Science (5th Ed.)]

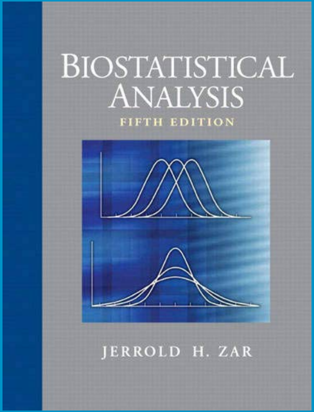
- The data is categorical. The table summarizes it.
- H_0 : success/no success is indepenent of sham/real surgery
- If that's the case, each entry in the table should be a product of propotions, e.g.

$$n_{\text{sham, success}} = n \cdot p_{\text{sham}} \cdot p_{\text{success}} = 75 \cdot \frac{26}{75} \cdot \frac{56}{75} = 19.41$$

There are more χ^2-tests



The following questions help you find a statistical test



1. What do I want to test (mean equal to a value? equality of two means? ...)?
2. What is the data type (interval/ratio, ordinal, categorical, ...)?
3. Is the data normally distributed or not?
4. How many groups do I have (1, 2, many)?
5. Is the data paired?
6. Do I need to test for deviations in one or two directions (one-tailed/-sided or two-tailed/-sided tests)?

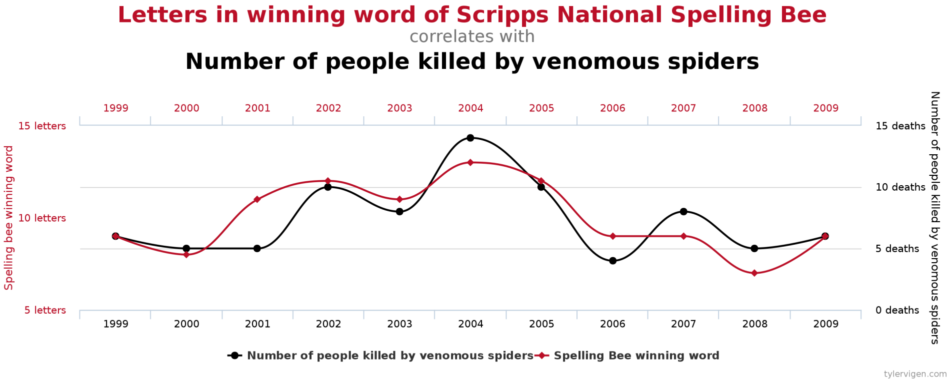
- The above diagram is only a small fraction of all possible tests.
- The pattern always stays the same: choose statistic, get null distribution, use it to quantify (p-value) whether what you observed could have happened by chance

Recommendations

- There is not a general recipe, not a general way of looking at data or doing data analysis (otherwise data scientists would be unemployed and a computer would do their job).
- Use your intelligence (and the book by Zar for instance) to choose the right one.
- Ask if you don't know what to take (e.g. `stats.stackexchange.com`).
- Play around in Python with toy example to get a feeling for a particular method/test/idea ...
- You can always use permutation tests or bootstrapping to verify you intuition.

p-Hacking

Should we be worried?



[https://en.wikipedia.org/wiki/Data_dredging]

A brave data scientist?

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Determination

A friend of yours works at a company and convinced that (s)he has found a new food supplement product that let's people loose weight quickly. Tests with participants do yields optimal results in the beginning, but (s)he is convinced of her/his idea and keeps trying. Finally, after 20 attempts, the test shows the expected weight loss with $p < 0.05$. Excited (s)he want to go to her/his boss to present to new invention to her.

What do you recommend you friend?

- You recommend to not go to the boss, because (s)he has been fishing for significant results and they are thus meaningless.

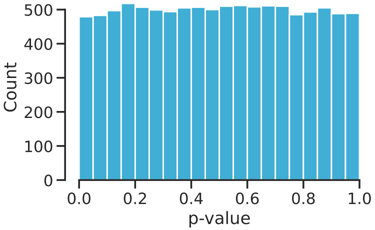
A brave data scientist?

Determination

A friend of yours works at a company and convinced that (s)he has found a new food supplement product that let's people loose weight quickly. Tests with participants do yields optimal results in the beginning, but (s)he is convinced of her/his idea and keeps trying. Finally, after 20 attempts, the test shows the expected weight loss with $p < 0.05$. Excited (s)he want to go to her/his boss to present to new invention to her.

What do you recommend you friend?

- You recommend to not go to the boss, because (s)he has been fishing for significant results and they are thus meaningless.
- If H_0 is true (there is no effect), the distribution p-value is uniform between 0 and 1.



A brave data scientist?

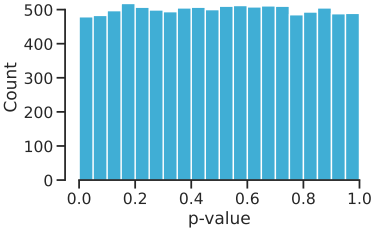
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Determination

A friend of yours works at a company and convinced that (s)he has found a new food supplement product that let's people loose weight quickly. Tests with participants do yields optimal results in the beginning, but (s)he is convinced of her/his idea and keeps trying. Finally, after 20 attempts, the test shows the expected weight loss with $p < 0.05$. Excited (s)he want to go to her/his boss to present to new invention to her.

What do you recommend you friend?

- You recommend to not go to the boss, because (s)he has been fishing for significant results and they are thus meaningless.
- If H_0 is true (there is no effect), the distribution p-value is uniform between 0 and 1.



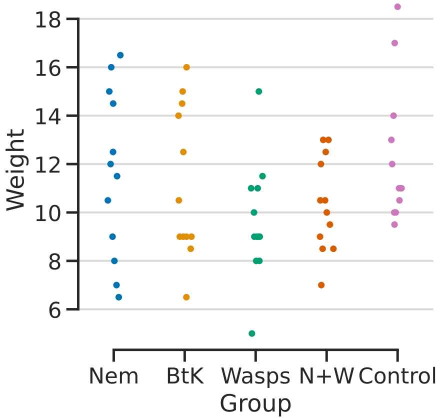
- Thus, $\alpha \cdot 100\%$ of p-values are expected to be $\leq \alpha$.

Sweet Corn

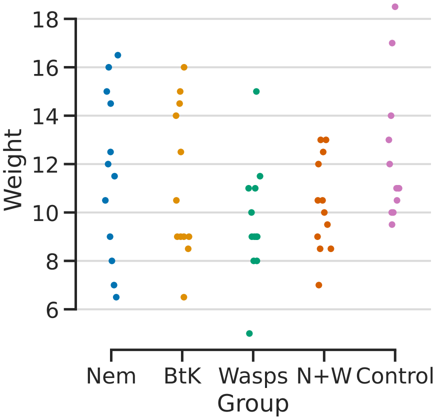
When growing sweet corn, can organic methods be used successfully to control harmful insects and limit their effect on the corn. In a study of this question researchers compared the weights of ears of corn under five conditions in an experiment in which sweet corn was grown using organic methods.

Are all the means equal?

[Statistics for the Life Sciences (5th Ed.)]

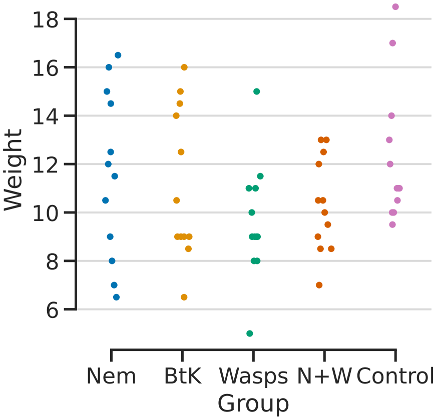


- $H_0 : \mu_1 = \mu_2 = \ldots = \mu_5$



Multiple Testing

- $H_0 : \mu_1 = \mu_2 = \ldots = \mu_5$
- Can we test all possible pairs and reject H_0 if one pair doesn't match?



Multiple Testing

- $H_0 : \mu_1 = \mu_2 = \ldots = \mu_5$
- Can we test all possible pairs and reject H_0 if one pair doesn't match?
- Assume they are equal (H_0 is true $\Rightarrow$ p is uniformly distributed between 0 and 1). Then all comparisons between 5 means would yield 10 p-values. Let's see what happens when we simulate that.

[0.38, 0.28, 1.00, 0.54, 0.74, 0.34, 0.10, 0.39, 0.47, 0.78]
[0.58, 0.51, 0.94, 0.90, 0.71, 0.61, 0.48, 0.52, 0.93, 0.24]
[0.97, 0.99, 0.38, 0.59, 0.93, 0.08, 0.43, 0.04, 0.01, 0.72]
[0.50, 0.97, 0.45, 0.22, 0.50, 0.32, 0.07, 0.76, 0.94, 0.38]
[0.20, 0.60, 0.03, 0.86, 0.92, 0.99, 0.97, 0.66, 0.32, 0.01]
[0.89, 0.19, 0.34, 0.69, 0.73, 0.08, 0.12, 0.59, 0.43, 0.33]
[0.14, 0.08, 0.61, 0.53, 0.07, 0.97, 0.56, 0.41, 0.97, 0.32]
[0.10, 0.23, 0.95, 0.86, 0.28, 0.03, 0.88, 0.33, 0.47, 0.79]
[0.52, 0.37, 0.27, 0.08, 0.62, 0.39, 0.01, 0.80, 0.20, 0.46]
[0.81, 0.27, 0.83, 0.79, 0.29, 0.38, 0.39, 0.77, 0.22, 0.15]
[0.76, 0.34, 0.64, 0.58, 0.95, 0.75, 0.18, 0.14, 0.94, 0.26]
[0.81, 0.96, 0.87, 0.03, 0.57, 0.05, 0.44, 0.89, 0.71, 0.45]
[0.82, 0.94, 0.09, 0.67, 0.48, 0.64, 0.46, 0.73, 0.55, 0.24]
[0.20, 0.16, 0.39, 0.67, 0.82, 0.54, 0.35, 0.19, 0.50, 0.07]
[0.74, 0.26, 0.10, 0.23, 0.25, 0.36, 0.91, 0.86, 0.24, 0.02]
[0.09, 0.38, 0.46, 0.15, 0.10, 0.43, 0.81, 0.07, 0.60, 0.47]
[0.34, 0.20, 0.08, 0.36, 0.82, 0.86, 0.42, 0.57, 0.09, 0.08]
[0.74, 0.94, 0.90, 0.47, 0.22, 0.55, 0.70, 0.67, 0.56, 0.14]
[0.84, 0.79, 0.45, 0.32, 0.35, 0.59, 0.43, 0.62, 0.23, 0.58]
[0.93, 0.30, 0.05, 0.07, 0.96, 0.23, 0.57, 0.98, 0.72, 0.84]

- **We reject H_0 7 out of 20 times** (how many times would we expect with $\alpha = 0.05$?)!

[0.38, 0.28, 1.00, 0.54, 0.74, 0.34, 0.10, 0.39, 0.47, 0.78]
[0.58, 0.51, 0.94, 0.90, 0.71, 0.61, 0.48, 0.52, 0.93, 0.24]
[0.97, 0.99, 0.38, 0.59, 0.93, 0.08, 0.43, 0.04, 0.01, 0.72]
[0.50, 0.97, 0.45, 0.22, 0.50, 0.32, 0.07, 0.76, 0.94, 0.38]
[0.20, 0.60, 0.03, 0.86, 0.92, 0.99, 0.97, 0.66, 0.32, 0.01]
[0.89, 0.19, 0.34, 0.69, 0.73, 0.08, 0.12, 0.59, 0.43, 0.33]
[0.14, 0.08, 0.61, 0.53, 0.07, 0.97, 0.56, 0.41, 0.97, 0.32]
[0.10, 0.23, 0.95, 0.86, 0.28, 0.03, 0.88, 0.33, 0.47, 0.79]
[0.52, 0.37, 0.27, 0.08, 0.62, 0.39, 0.01, 0.80, 0.20, 0.46]
[0.81, 0.27, 0.83, 0.79, 0.29, 0.38, 0.39, 0.77, 0.22, 0.15]
[0.76, 0.34, 0.64, 0.58, 0.95, 0.75, 0.18, 0.14, 0.94, 0.26]
[0.81, 0.96, 0.87, 0.03, 0.57, 0.05, 0.44, 0.89, 0.71, 0.45]
[0.82, 0.94, 0.09, 0.67, 0.48, 0.64, 0.46, 0.73, 0.55, 0.24]
[0.20, 0.16, 0.39, 0.67, 0.82, 0.54, 0.35, 0.19, 0.50, 0.07]
[0.74, 0.26, 0.10, 0.23, 0.25, 0.36, 0.91, 0.86, 0.24, 0.02]
[0.09, 0.38, 0.46, 0.15, 0.10, 0.43, 0.81, 0.07, 0.60, 0.47]
[0.34, 0.20, 0.08, 0.36, 0.82, 0.86, 0.42, 0.57, 0.09, 0.08]
[0.74, 0.94, 0.90, 0.47, 0.22, 0.55, 0.70, 0.67, 0.56, 0.14]
[0.84, 0.79, 0.45, 0.32, 0.35, 0.59, 0.43, 0.62, 0.23, 0.58]
[0.93, 0.30, 0.05, 0.07, 0.96, 0.23, 0.57, 0.98, 0.72, 0.84]

- We reject H_0 7 out of 20 times (how many times would we expect with $\alpha = 0.05$?)!
- Compound hypotheses can increase the false positives rate.

[0.38, 0.28, 1.00, 0.54, 0.74, 0.34, 0.10, 0.39, 0.47, 0.78]
[0.58, 0.51, 0.94, 0.90, 0.71, 0.61, 0.48, 0.52, 0.93, 0.24]
[0.97, 0.99, 0.38, 0.59, 0.93, 0.08, 0.43, 0.04, 0.01, 0.72]
[0.50, 0.97, 0.45, 0.22, 0.50, 0.32, 0.07, 0.76, 0.94, 0.38]
[0.20, 0.60, 0.03, 0.86, 0.92, 0.99, 0.97, 0.66, 0.32, 0.01]
[0.89, 0.19, 0.34, 0.69, 0.73, 0.08, 0.12, 0.59, 0.43, 0.33]
[0.14, 0.08, 0.61, 0.53, 0.07, 0.97, 0.56, 0.41, 0.97, 0.32]
[0.10, 0.23, 0.95, 0.86, 0.28, 0.03, 0.88, 0.33, 0.47, 0.79]
[0.52, 0.37, 0.27, 0.08, 0.62, 0.39, 0.01, 0.80, 0.20, 0.46]
[0.81, 0.27, 0.83, 0.79, 0.29, 0.38, 0.39, 0.77, 0.22, 0.15]
[0.76, 0.34, 0.64, 0.58, 0.95, 0.75, 0.18, 0.14, 0.94, 0.26]
[0.81, 0.96, 0.87, 0.03, 0.57, 0.05, 0.44, 0.89, 0.71, 0.45]
[0.82, 0.94, 0.09, 0.67, 0.48, 0.64, 0.46, 0.73, 0.55, 0.24]
[0.20, 0.16, 0.39, 0.67, 0.82, 0.54, 0.35, 0.19, 0.50, 0.07]
[0.74, 0.26, 0.10, 0.23, 0.25, 0.36, 0.91, 0.86, 0.24, 0.02]
[0.09, 0.38, 0.46, 0.15, 0.10, 0.43, 0.81, 0.07, 0.60, 0.47]
[0.34, 0.20, 0.08, 0.36, 0.82, 0.86, 0.42, 0.57, 0.09, 0.08]
[0.74, 0.94, 0.90, 0.47, 0.22, 0.55, 0.70, 0.67, 0.56, 0.14]
[0.84, 0.79, 0.45, 0.32, 0.35, 0.59, 0.43, 0.62, 0.23, 0.58]
[0.93, 0.30, 0.05, 0.07, 0.96, 0.23, 0.57, 0.98, 0.72, 0.84]

- **We reject H_0 7 out of 20 times** (how many times would we expect with $\alpha = 0.05$?)!
- Compound hypotheses can increase the false positives rate.
- One way to account for multiple testing is to **divide α by the number of tests** or **multiply each p-value by the number of tests**.
- This is called **Bonferroni correction**.

[0.38, 0.28, 1.00, 0.54, 0.74, 0.34, 0.10, 0.39, 0.47, 0.78]
[0.58, 0.51, 0.94, 0.90, 0.71, 0.61, 0.48, 0.52, 0.93, 0.24]
[0.97, 0.99, 0.38, 0.59, 0.93, 0.08, 0.43, 0.04, 0.01, 0.72]
[0.50, 0.97, 0.45, 0.22, 0.50, 0.32, 0.07, 0.76, 0.94, 0.38]
[0.20, 0.60, 0.03, 0.86, 0.92, 0.99, 0.97, 0.66, 0.32, 0.01]
[0.89, 0.19, 0.34, 0.69, 0.73, 0.08, 0.12, 0.59, 0.43, 0.33]
[0.14, 0.08, 0.61, 0.53, 0.07, 0.97, 0.56, 0.41, 0.97, 0.32]
[0.10, 0.23, 0.95, 0.86, 0.28, 0.03, 0.88, 0.33, 0.47, 0.79]
[0.52, 0.37, 0.27, 0.08, 0.62, 0.39, 0.01, 0.80, 0.20, 0.46]
[0.81, 0.27, 0.83, 0.79, 0.29, 0.38, 0.39, 0.77, 0.22, 0.15]
[0.76, 0.34, 0.64, 0.58, 0.95, 0.75, 0.18, 0.14, 0.94, 0.26]
[0.81, 0.96, 0.87, 0.03, 0.57, 0.05, 0.44, 0.89, 0.71, 0.45]
[0.82, 0.94, 0.09, 0.67, 0.48, 0.64, 0.46, 0.73, 0.55, 0.24]
[0.20, 0.16, 0.39, 0.67, 0.82, 0.54, 0.35, 0.19, 0.50, 0.07]
[0.74, 0.26, 0.10, 0.23, 0.25, 0.36, 0.91, 0.86, 0.24, 0.02]
[0.09, 0.38, 0.46, 0.15, 0.10, 0.43, 0.81, 0.07, 0.60, 0.47]
[0.34, 0.20, 0.08, 0.36, 0.82, 0.86, 0.42, 0.57, 0.09, 0.08]
[0.74, 0.94, 0.90, 0.47, 0.22, 0.55, 0.70, 0.67, 0.56, 0.14]
[0.84, 0.79, 0.45, 0.32, 0.35, 0.59, 0.43, 0.62, 0.23, 0.58]
[0.93, 0.30, 0.05, 0.07, 0.96, 0.23, 0.57, 0.98, 0.72, 0.84]

P-Hacking

- Reporting only significant results or searching for significant results and reporting only these is called **p-hacking**.
- It can drastically increase the false positives rate.

P-Hacking

- Reporting only significant results or searching for significant results and reporting only these is called **p-hacking**.
- It can drastically increase the false positives rate.

Multiple testing

- Testing compound hypotheses that are made up of many single tests is called “multiple testing”
- Rejecting H_0 as soon as one test fails can dramatically increase the false positives rate.
- There are methods to correct for that. One (very conservative one) is Bonferroni correction.

Thanks for listening! Questions?