

Méthodes d'évaluation empirique

Analyses statistiques

Aurélien Tabard

Méthodes d'évaluation

Matin

- Introduction ~ 45 min / 1h
 - Approches d'évaluation
 - Méthodes analytiques
 - Méthodes empiriques
- Concevoir une expérience 1h
 - Exemples
 - Les bases
 - La structure d'une expérience
 - Mener une expérience
 - Collecter les données
- Mise en pratique 1h

Before starting

Scaling

We now have access to large audiences :

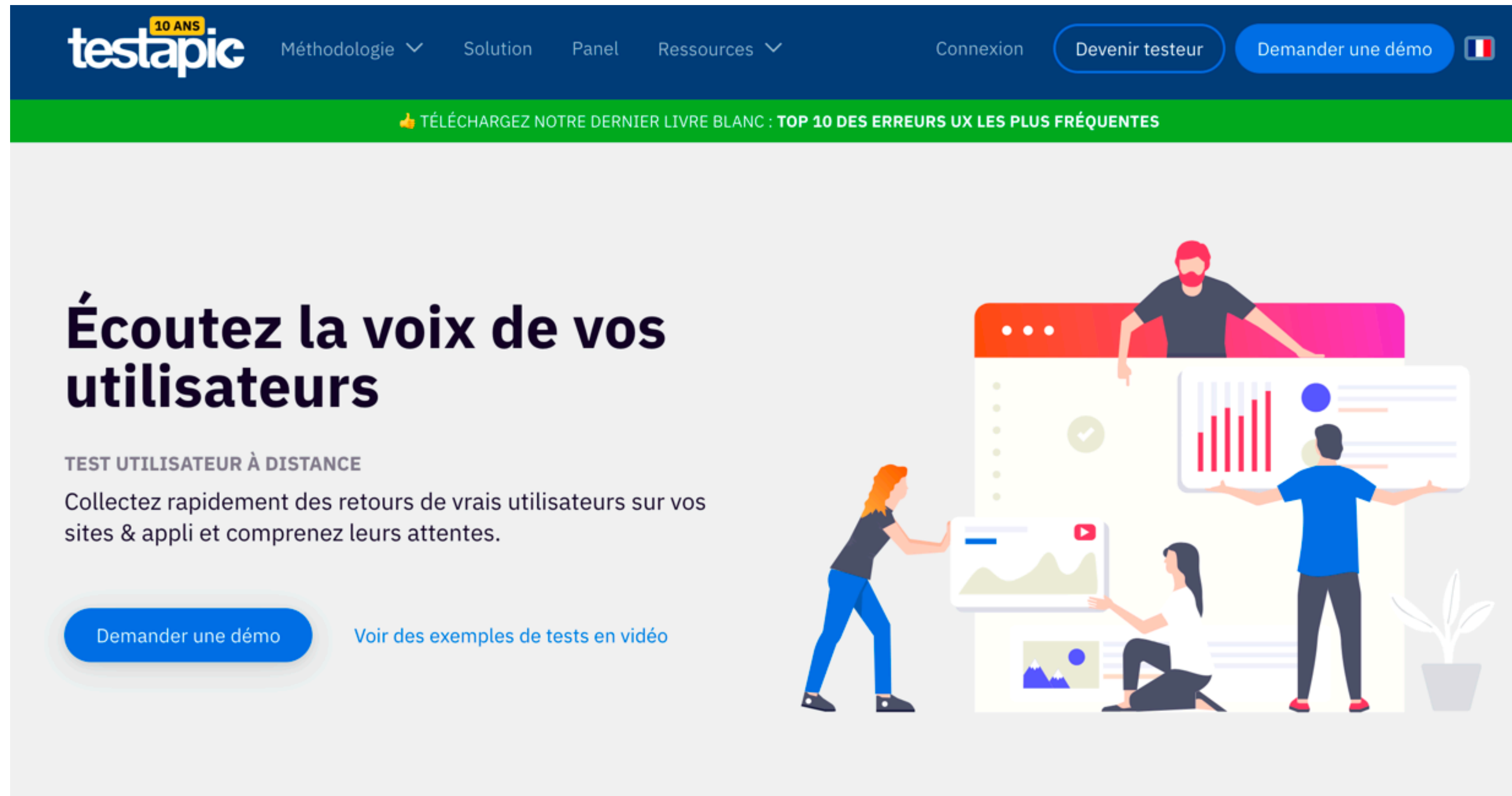
- On the Web
- On mobile platforms

With two interesting properties :

- Ease of distribution of updates
- Ease of logging


We can scale up studies, as the one discussed later (would deserve a lecture in itself)

Remote usability studies



The image shows the homepage of the Testapic website. At the top, there is a dark blue navigation bar with the Testapic logo (including a '10 ANS' badge), menu items for 'Méthodologie', 'Solution', 'Panel', and 'Ressources', and buttons for 'Connexion', 'Devenir testeur', and 'Demander une démo'. A green banner below the navigation bar promotes a white paper: 'TÉLÉCHARGEZ NOTRE DERNIER LIVRE BLANC : TOP 10 DES ERREURS UX LES PLUS FRÉQUENTES'. The main content area features the headline 'Écoutez la voix de vos utilisateurs' and the sub-headline 'TEST UTILISATEUR À DISTANCE'. Below this, a paragraph states: 'Collectez rapidement des retours de vrais utilisateurs sur vos sites & appli et comprenez leurs attentes.' There are two call-to-action buttons: 'Demander une démo' and 'Voir des exemples de tests en vidéo'. On the right side, there is an illustration of four people (three men and one woman) gathered around a large screen displaying various data visualizations like bar charts, line graphs, and a video player, representing a collaborative remote usability study session.

testapic 10 ANS

Méthodologie ▾ Solution Panel Ressources ▾ Connexion Devenir testeur Demander une démo 


👉 TÉLÉCHARGEZ NOTRE DERNIER LIVRE BLANC : TOP 10 DES ERREURS UX LES PLUS FRÉQUENTES

Écoutez la voix de vos utilisateurs

TEST UTILISATEUR À DISTANCE

Collectez rapidement des retours de vrais utilisateurs sur vos sites & appli et comprenez leurs attentes.

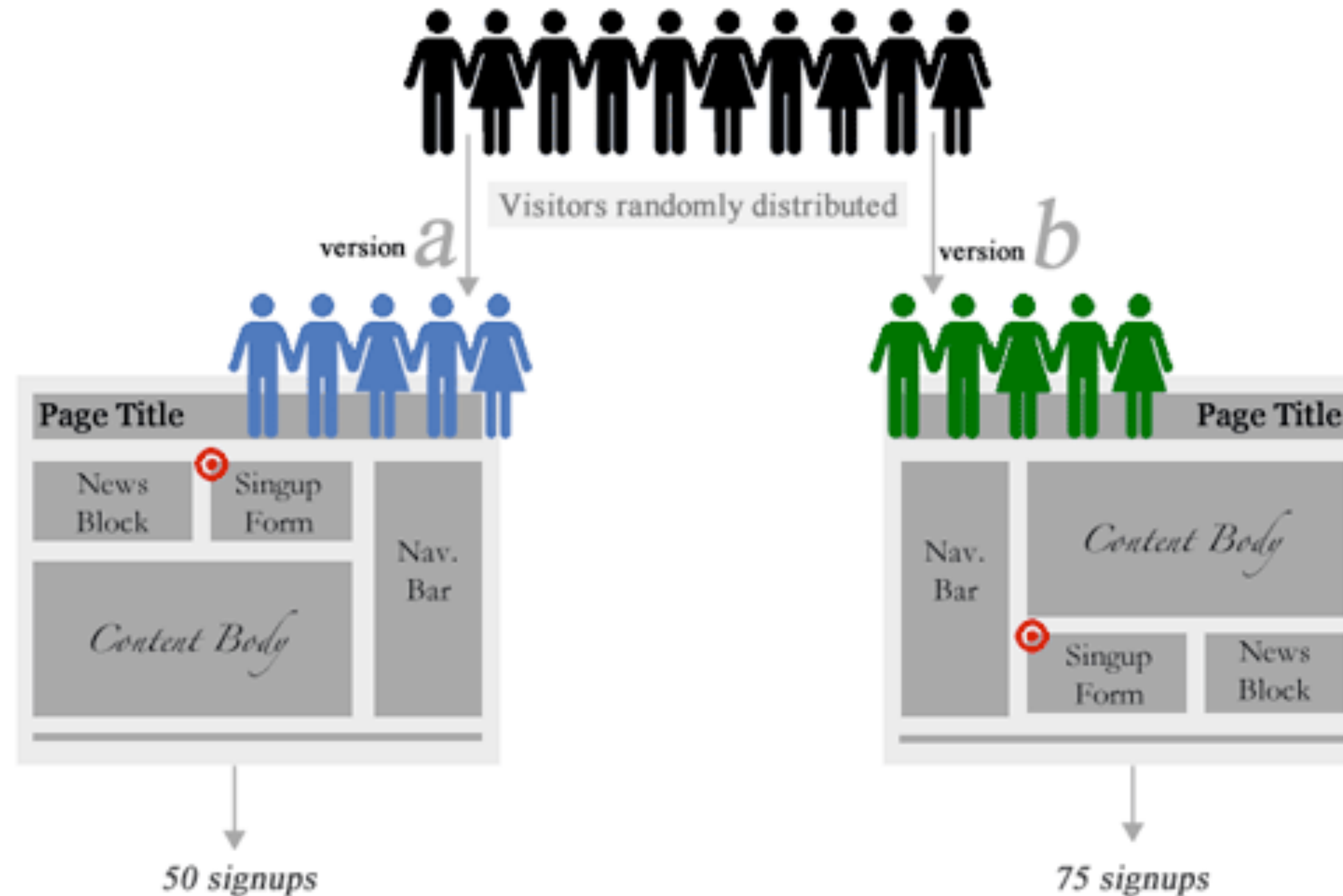
Demander une démo Voir des exemples de tests en vidéo



Controlled distribution of Beta versions

The screenshot shows the TestFlight website interface. At the top, the TestFlight logo is on the left, and navigation links for SDK, TestFlight Live, Support, Blog, About, Jobs, Log In, and Sign Up are on the right. Below the navigation is a banner with the text: "SDK + [icon] = TestFlight Live. Real-time dashboard for actions and revenue. Read more »". The main headline reads "The freedom to build better apps" with a "FREE" badge. Below this is the subtext: "A free testing service for mobile developers, managers and testers." The "How it works:" section features a flowchart with three steps: "Set up TestFlight" (represented by a control tower icon), "Distribute your beta" (represented by an airplane icon), and "Analyze usage" (represented by a box with "CHECKPOINTS", "CRASHES", and "FEEDBACK" labels). The final step is "Improve your app!" (represented by a drafting tool icon).

A/B testing



Version B is better than version A

A/B testing

e.g. [optimizely.com](https://www.optimizely.com)

Test alternative designs of webpages, or mobile screens. Helps identify :

- better form designs
- better conversion rates (e.g. for a newsletter)

Limits :

- You need significant traffic
- Does not replace user studies !
- Does not provide explanations / qualitative insights
- Arbitrary changes can be disturbing to users
- Complex when there is tailored and social content, e.g., Facebook
- Often used for incremental changes, complex for full redesigns

Statistical analysis

Afternoon

- Practice
- Checking your data
- Significance testing with t-tests
- Significance testing with Anova
- Measuring effect sizes
- Beyond significance testing

Statistical analysis

- **Practice**
- Checking your data
- Significance testing with t-tests
- Significance testing with Anova
- Measuring effect sizes
- Beyond significance testing

Practice

How could you test the effect of **two soporific drugs (independent variable)** on **amount of sleep (dependent variable)**



Practice

You want to test the effect of **two soporific drugs (independent variable)** on **amount of sleep(dependent variable)**.

Recruit 10 participants and make them sleep to get their basic (control) sleep time.

Then you give them drug 1 and note the difference of sleep time.

You do the same for drug 2.

sleep extra drug 1

1	0.7
2	-1.6
3	-0.2
4	-1.2
5	-0.1
6	3.4
7	3.7
8	0.8
9	0.0
10	2.0

sleep extra drug 2

1	1.9
2	0.8
3	1.1
4	0.1
5	-0.1
6	4.4
7	5.5
8	1.6
9	4.6
10	3.4

Practice

You want to test the effect of **two soporific drugs** (independent variable) on **amount of sleep** (dependent variable).

Two possibilities :

- Participants went through both conditions, i.e. **had both drugs**
= within subject experiment so the **data is paired**
- Participants were **split into 2 groups** (e.g. 10 new participants for drug 2)
= between subject experiment so the **data is unpaired**

Practice 2

Identify the best controller

You have created a new VR app, and have to decide which VR controller is better.

How do we proceed?



Statistical analysis

This week

- Practice
- **Checking your data**
- Significance testing with t-tests
- Significance testing with Anova
- Measuring effect sizes
- Beyond significance testing

Who is fastest?

It depends of:

- the median differences
- the data distribution (standard deviation)
- the sample size
- whether averages are significantly different

First : the exploratory part

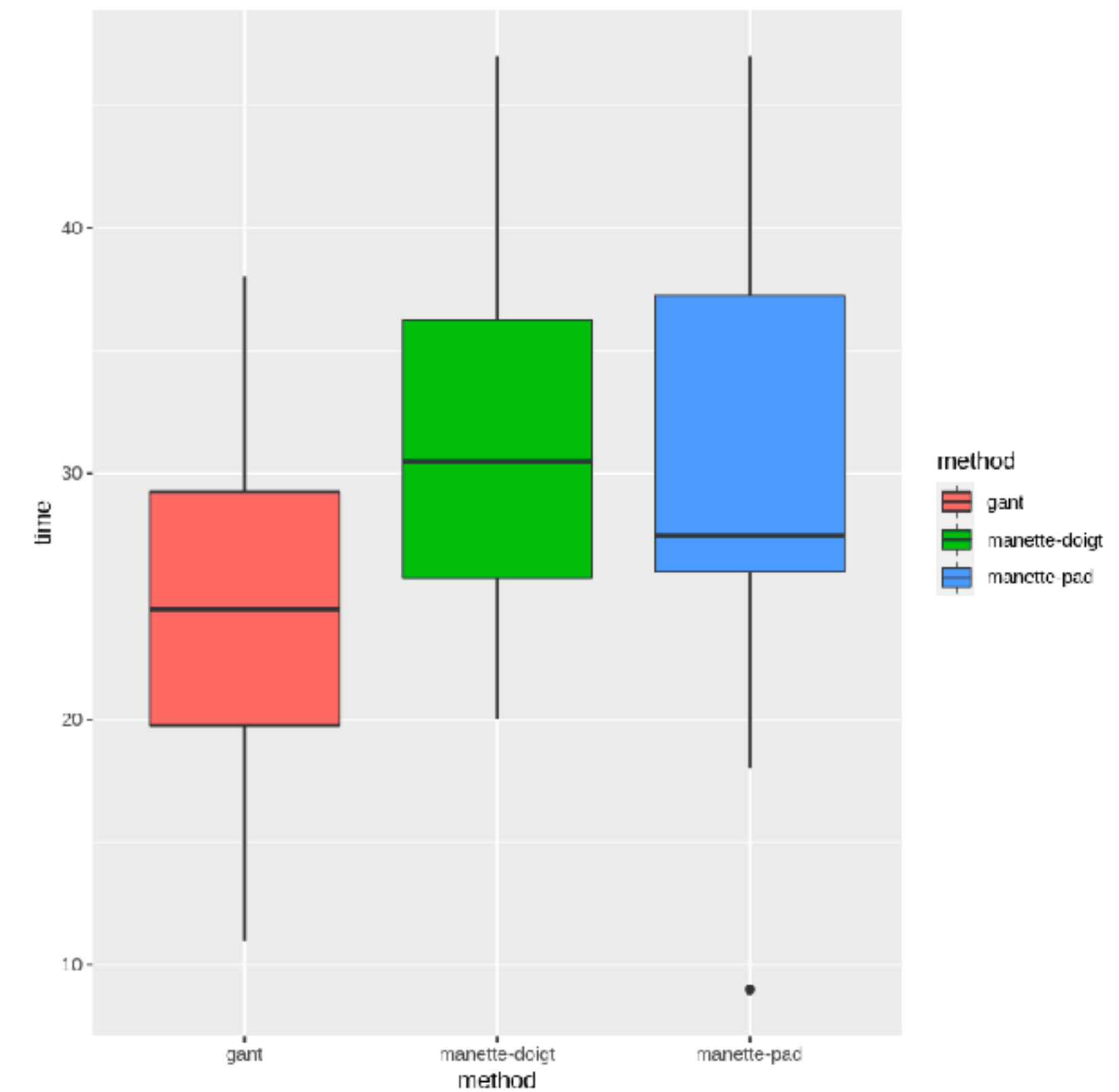
- look at the data with basic plots and statistics to get an idea.

Plotting your data

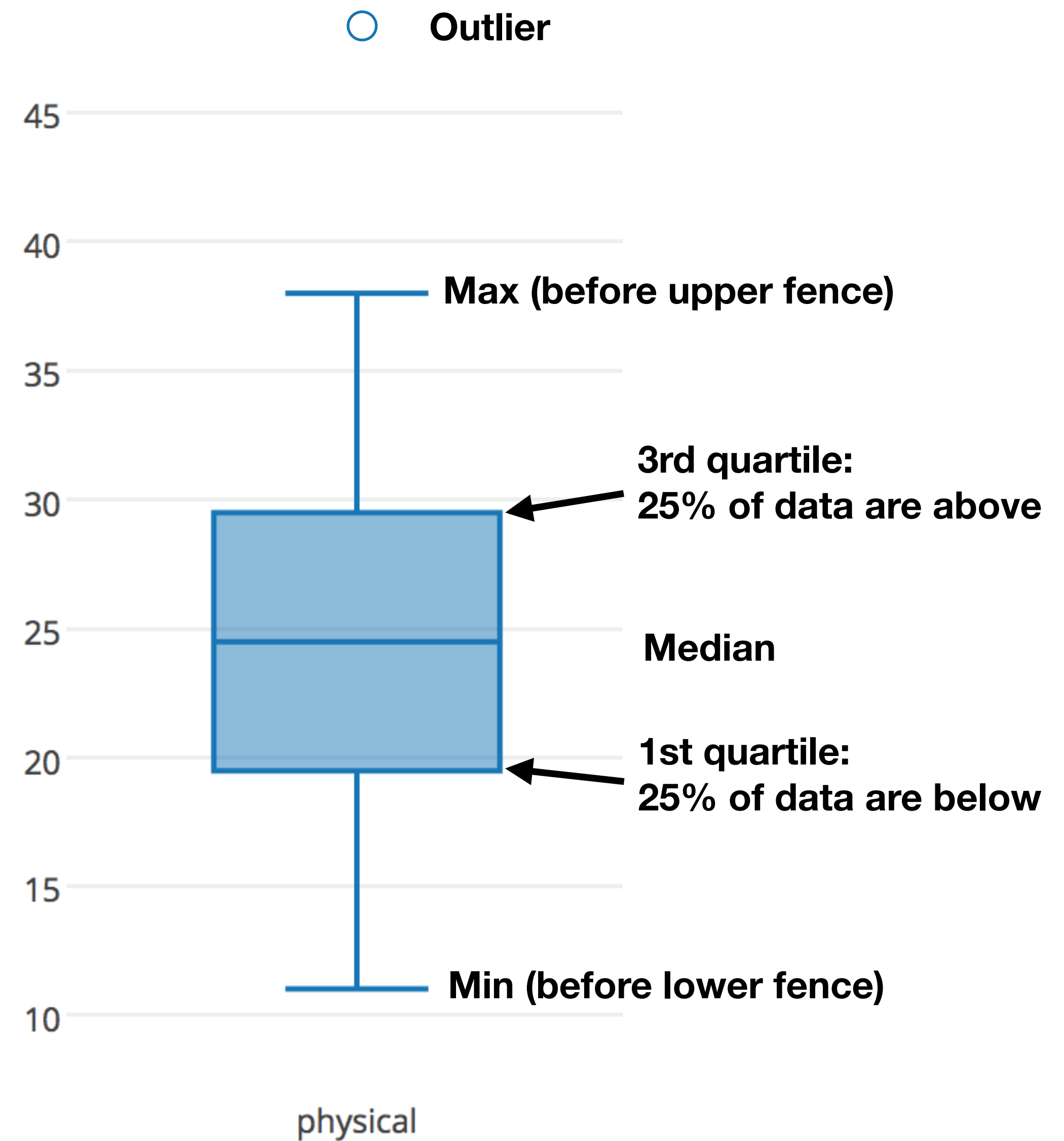
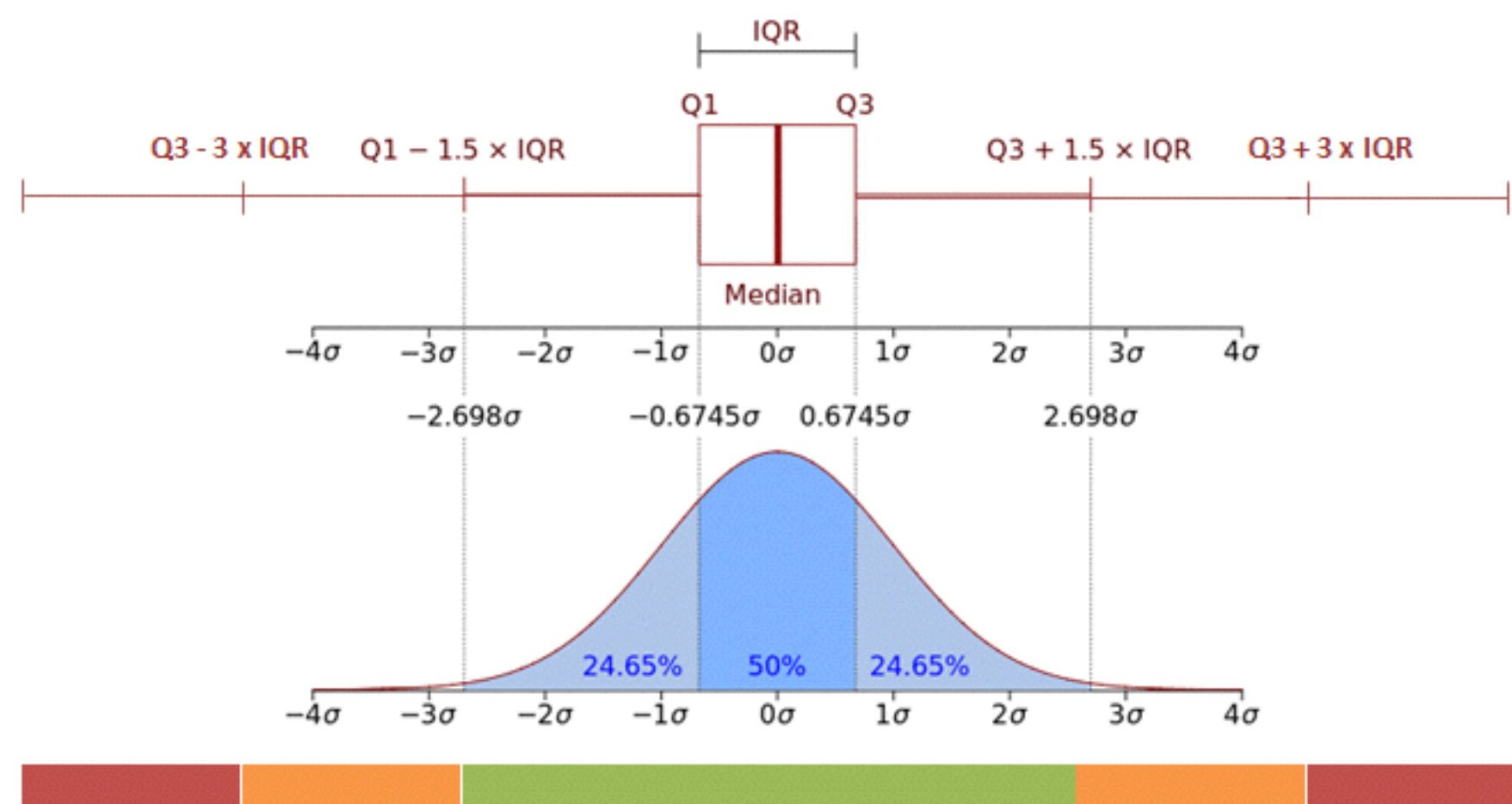
<https://colab.research.google.com/drive/1Is8hWFtlnLXOoHqpU7C5jPR2vHLO8y3D?usp=sharing>

```
head(data) #
```

```
ggplot(data, aes(x=method, y=time, fill=method)) +  
  geom_boxplot()
```



Reading a boxplot



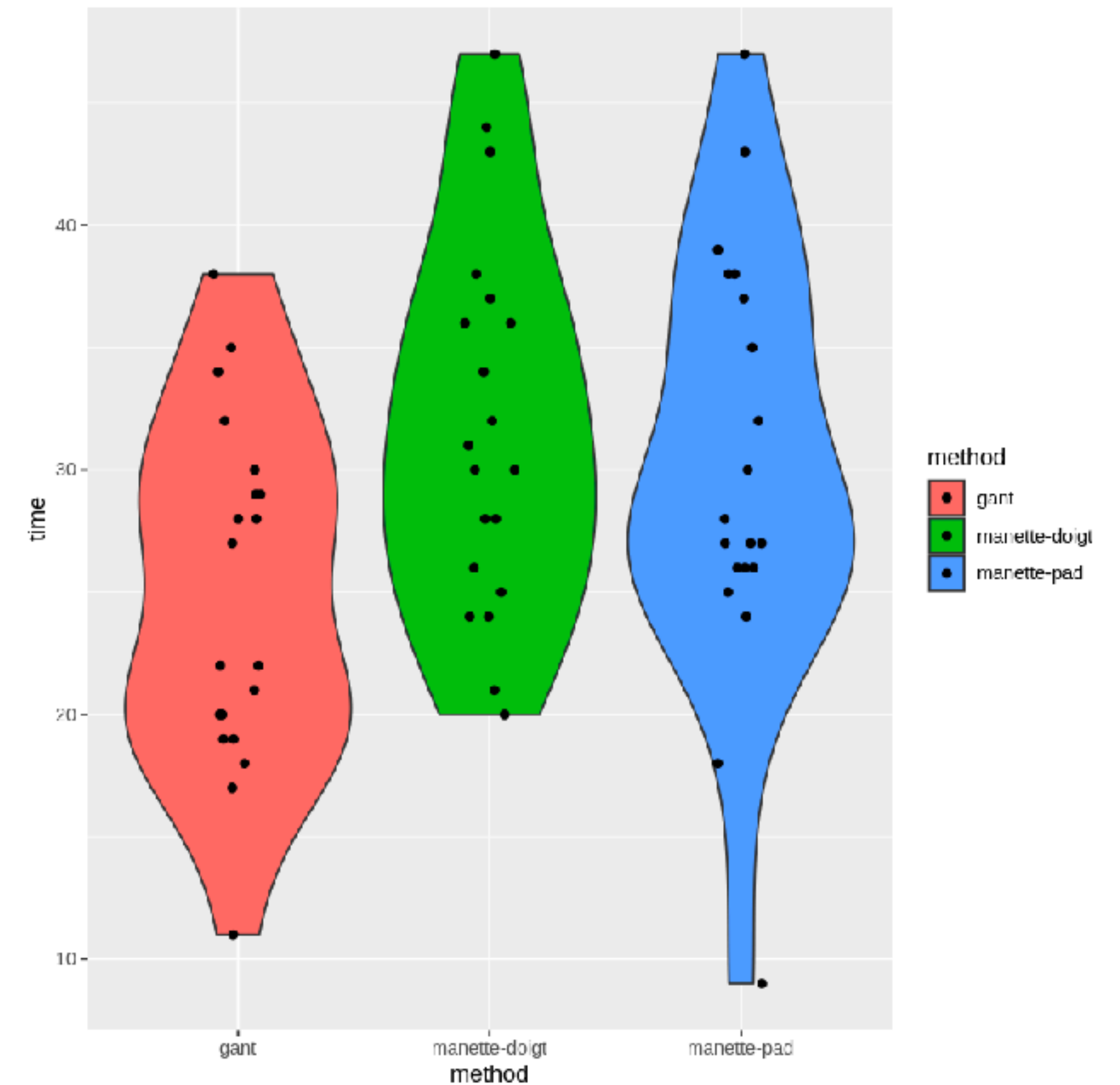
Prefer a violin plot

<https://colab.research.google.com/drive/1Is8hWFtlnLXOoHqpU7C5jPR2vHLO8y3D?usp=sharing>

```
head(data) #
```

```
ggplot(data, aes(x=method, y=time, fill=method)) +  
  geom_boxplot()
```

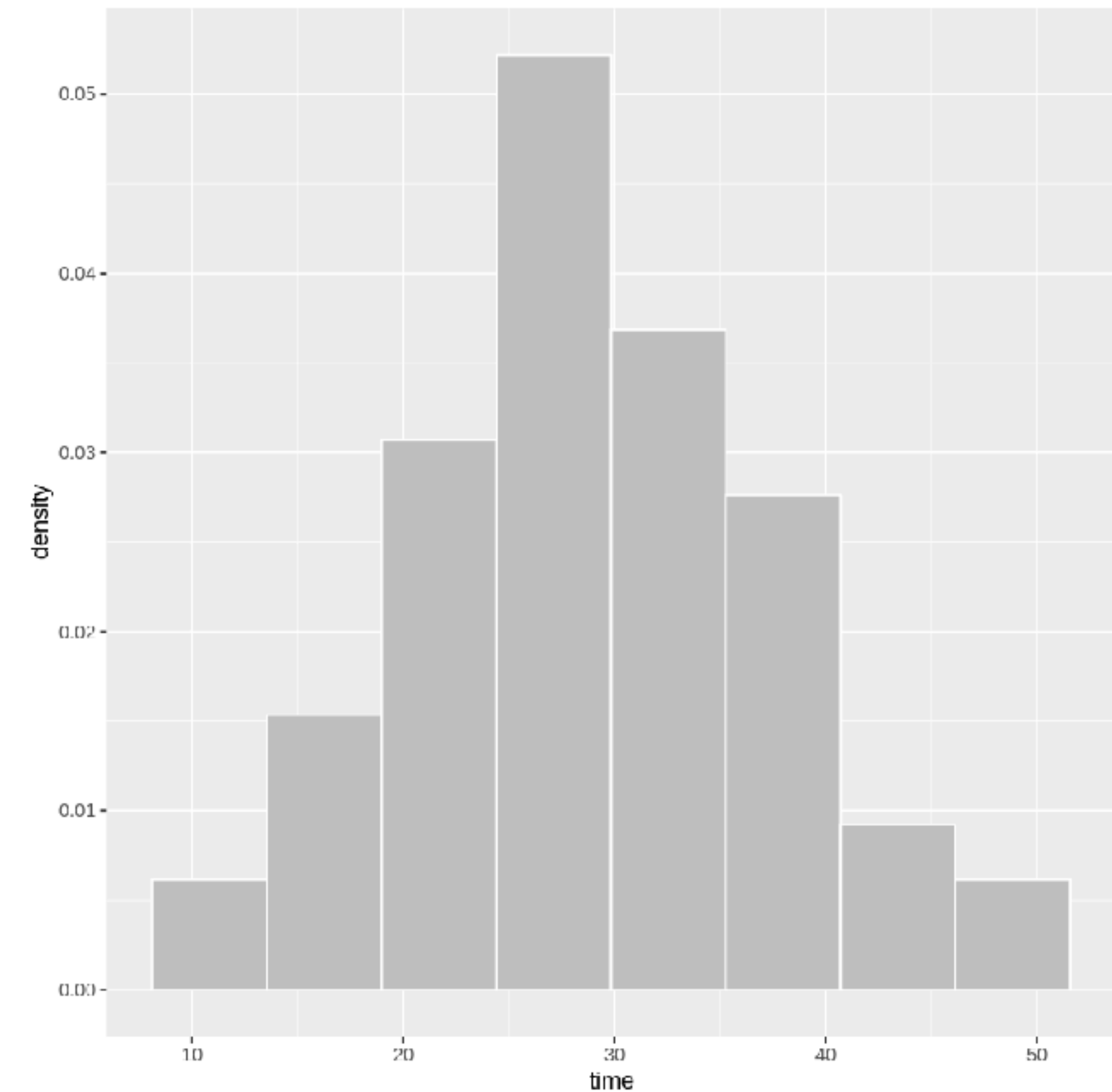
```
ggplot(data, aes(x=method, y=time, fill=method)) +  
  geom_violin() +  
  geom_jitter(height = 0, width = 0.1)
```



Looking at your data distribution

<https://colab.research.google.com/drive/1Is8hWFtlnLXOoHqpU7C5jPR2vHLO8y3D?usp=sharing>

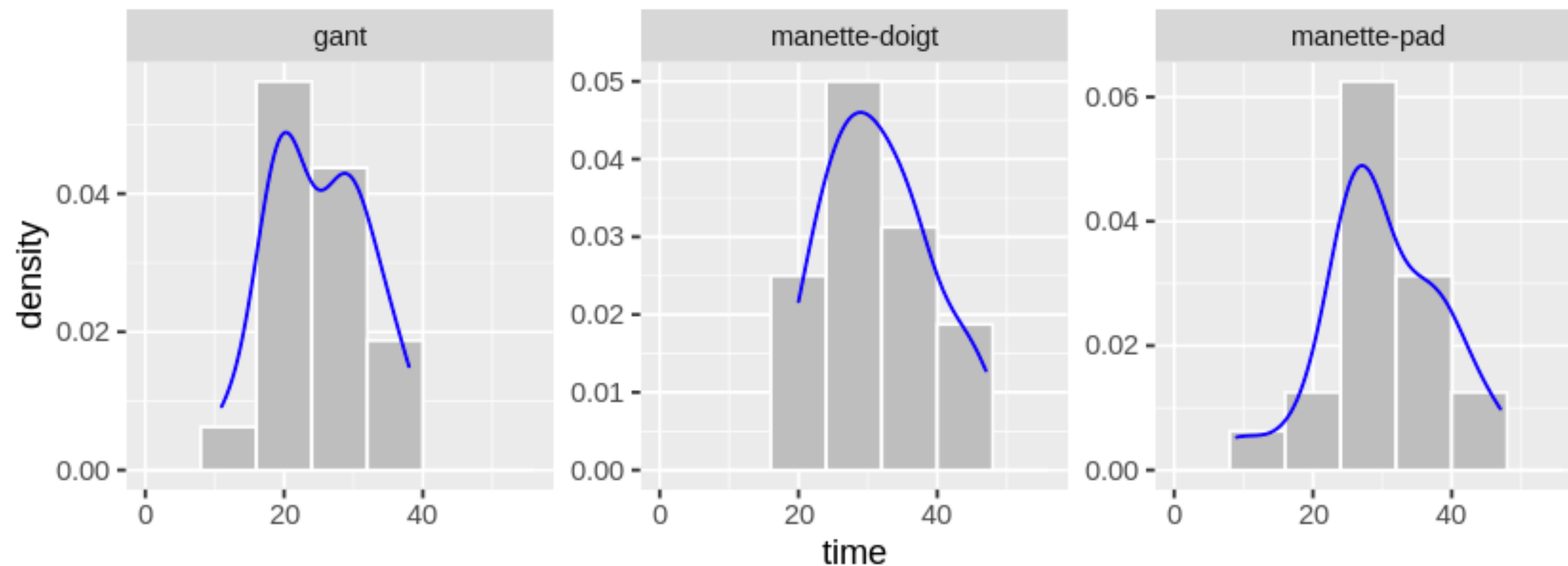
```
ggplot(data, aes(x = time)) +  
  geom_histogram(aes(y = ..density..),  
                bins=8, # or specify manually :  
                # breaks = seq(0, 60, by = 10),  
                colour = "white", fill="grey75")
```



Looking at your data distribution

<https://colab.research.google.com/drive/1Is8hWFtlnLXOoHqpU7C5jPR2vHLO8y3D?usp=sharing>

```
ggplot(data, aes(x=time)) +  
  geom_histogram( aes(y=..density..),  
                 breaks = seq(0, 60, by = 8),  
                 colour = "white", fill="grey75") +  
  facet_wrap(~method, scales = "free") +  
  geom_density(aes(y=..density..), colour="blue")
```



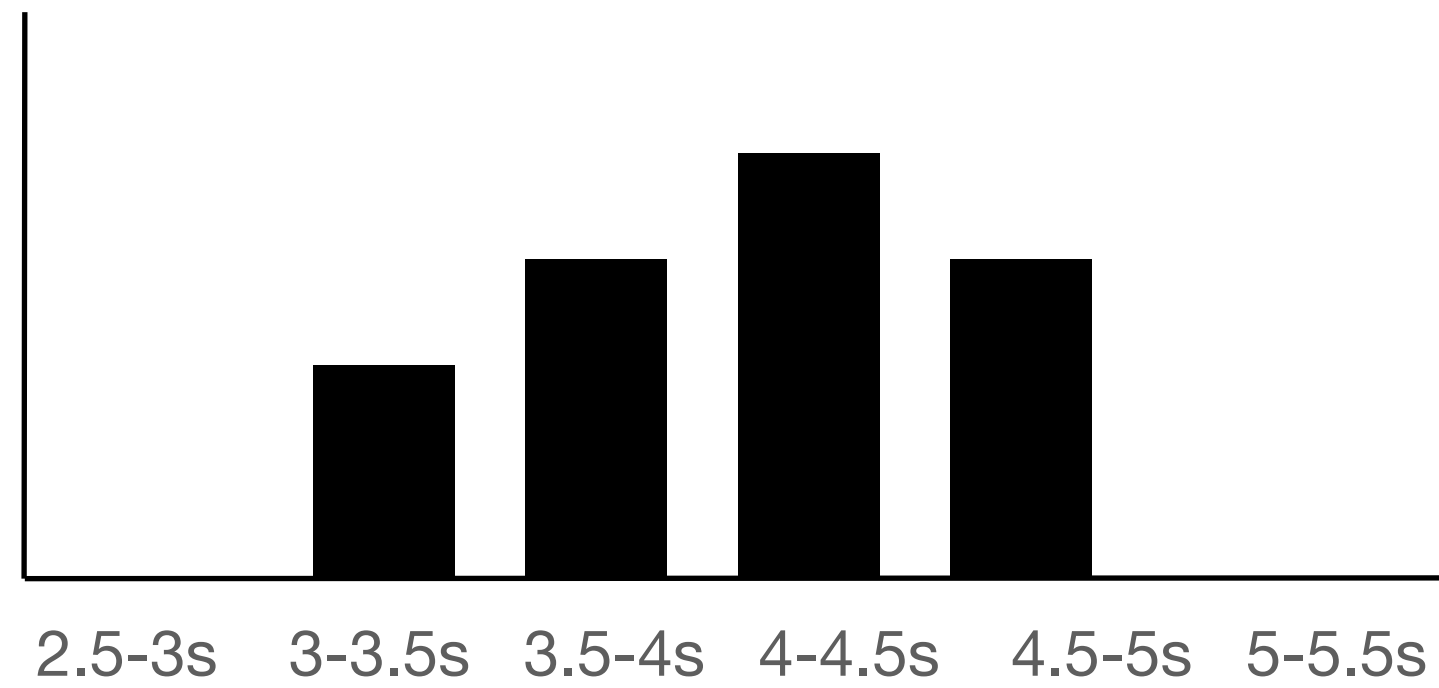
Statistical analysis

- Practice
- Checking your data
- **Significance testing**
 - with t-tests
 - with Anova
- Measuring effect sizes
- Beyond significance testing

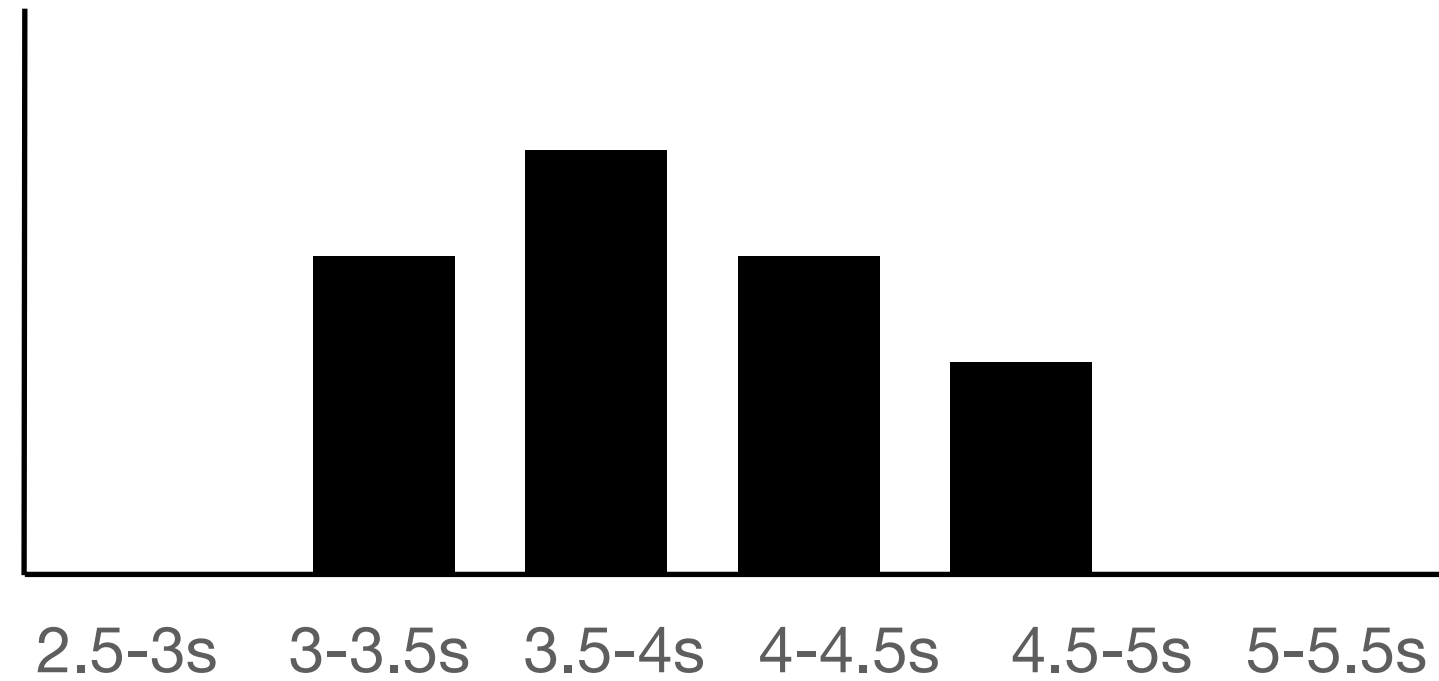
Statistical significance

**A result is called statistically significant
if it is unlikely to have occurred by chance**

Une différence significative ?

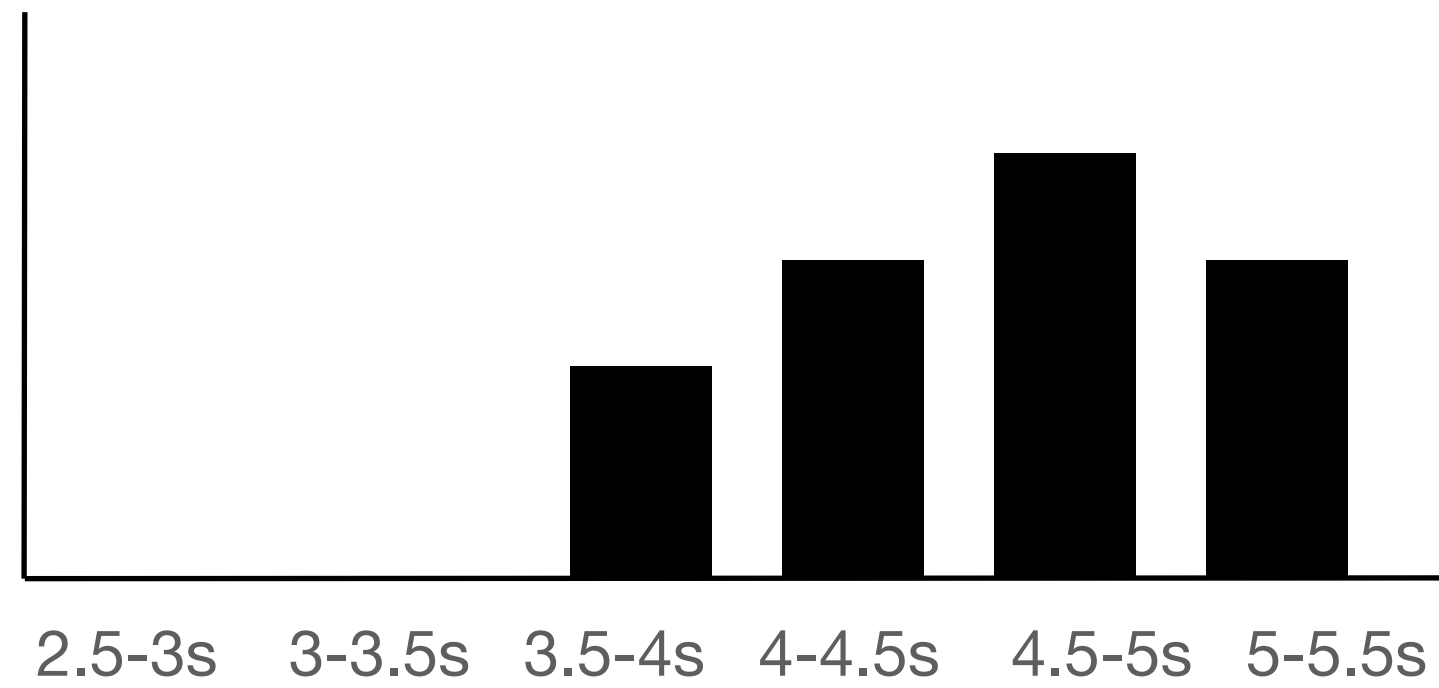


45% chance that the mean of a sample from each group is similar

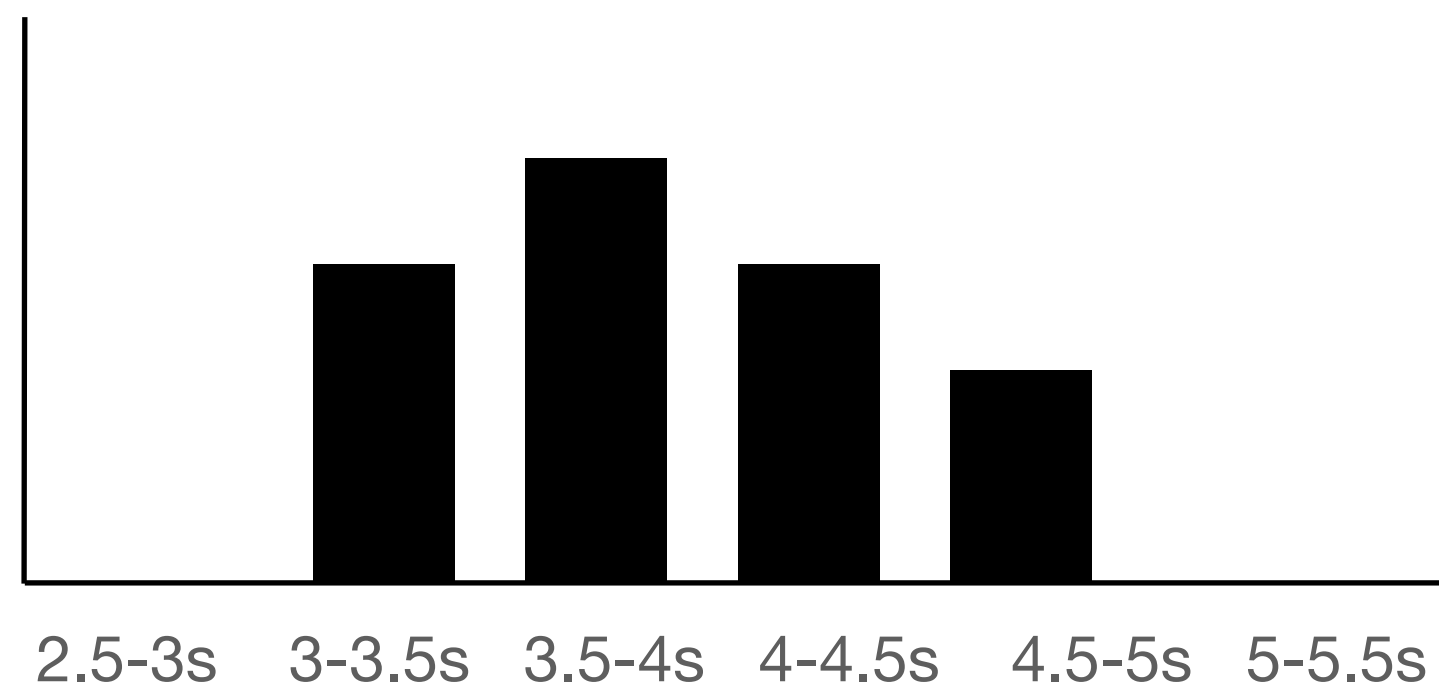


TTEST pvalue = 0.4548

Une différence significative ?

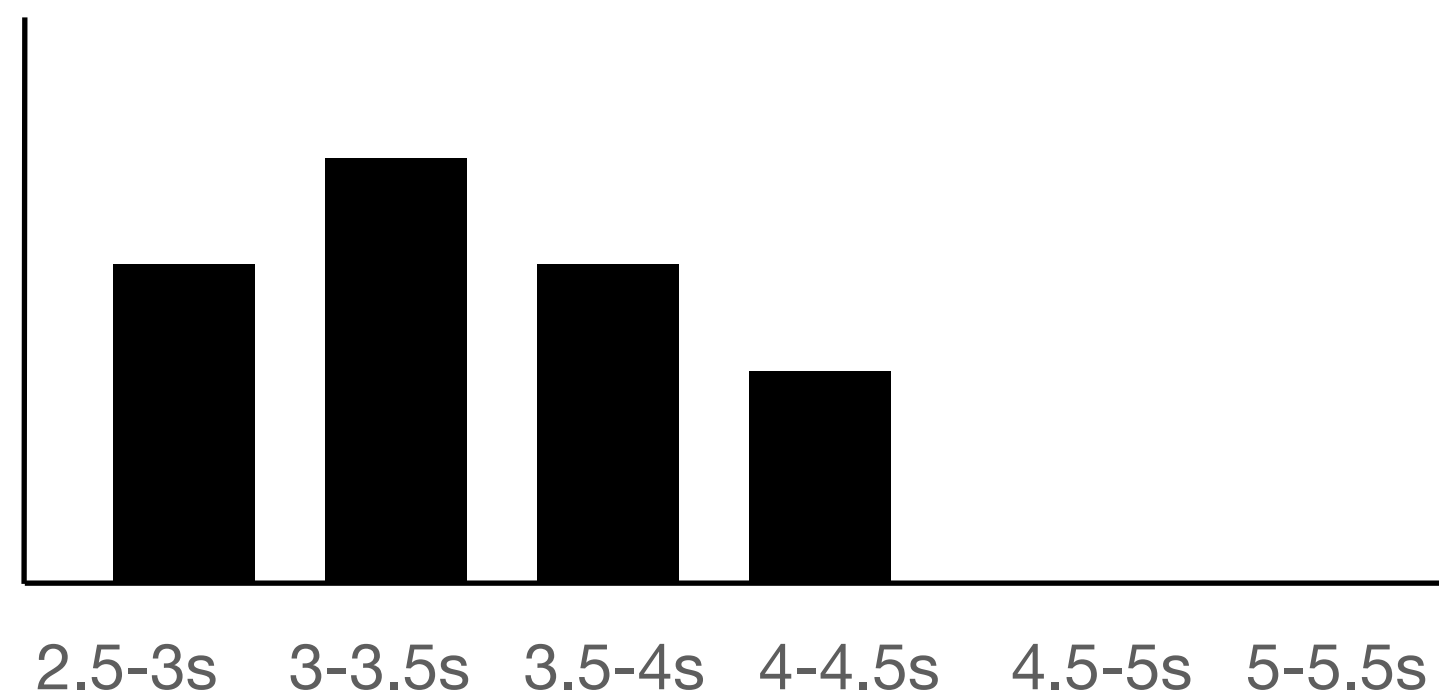
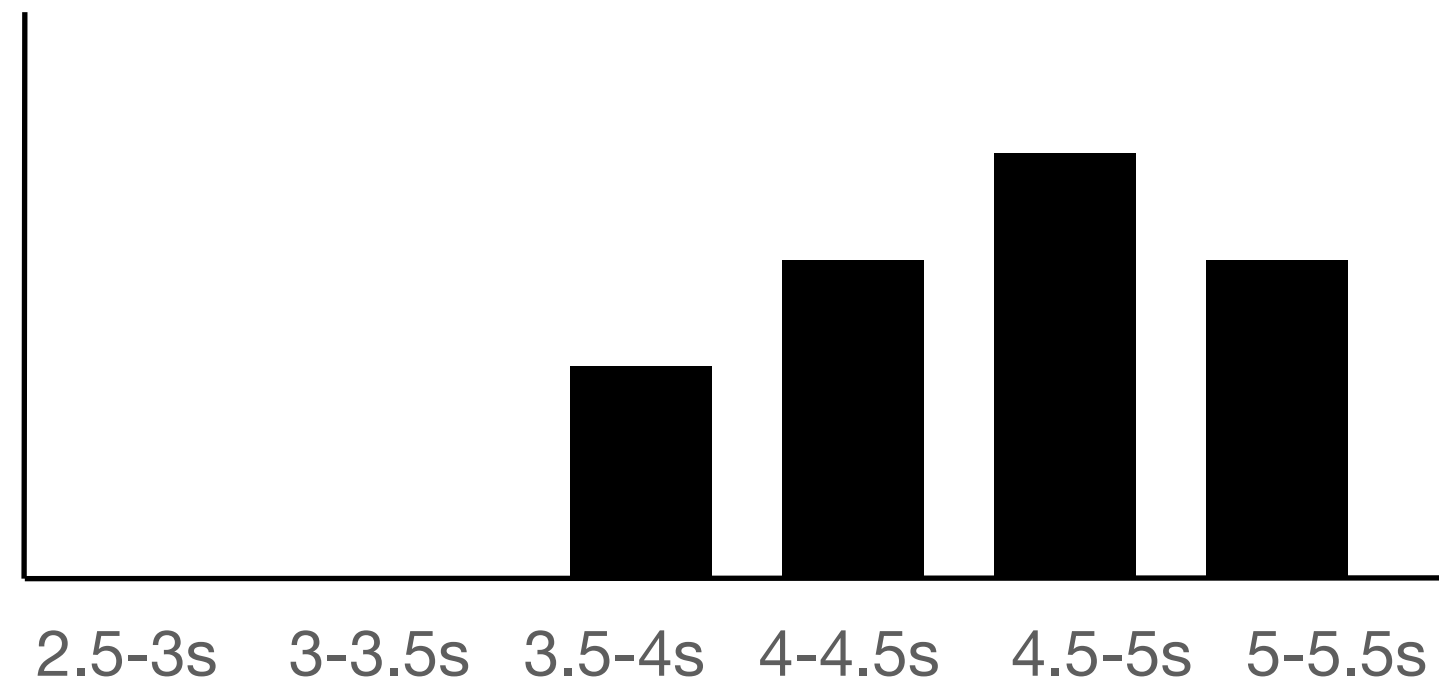


14% chance that the mean of a sample from each group is similar



TTEST pvalue = 0.14432

Une différence significative ?



unlikely two samples
will have the same mean

TTEST pvalue = 0.00097

Significance level

If a test of significance gives a **p-value lower** than the significance level, such results are informally referred to as 'statistically significant'.

Popular levels of significance are :

- 10% (0.1),
- **5% (0.05),**
- 1% (0.01),
- 0.5% (0.005), and
- 0.1% (0.001).

Significance testing

Gives p:

- The probability that two population have the same mean
- Not probability the result is due to chance...

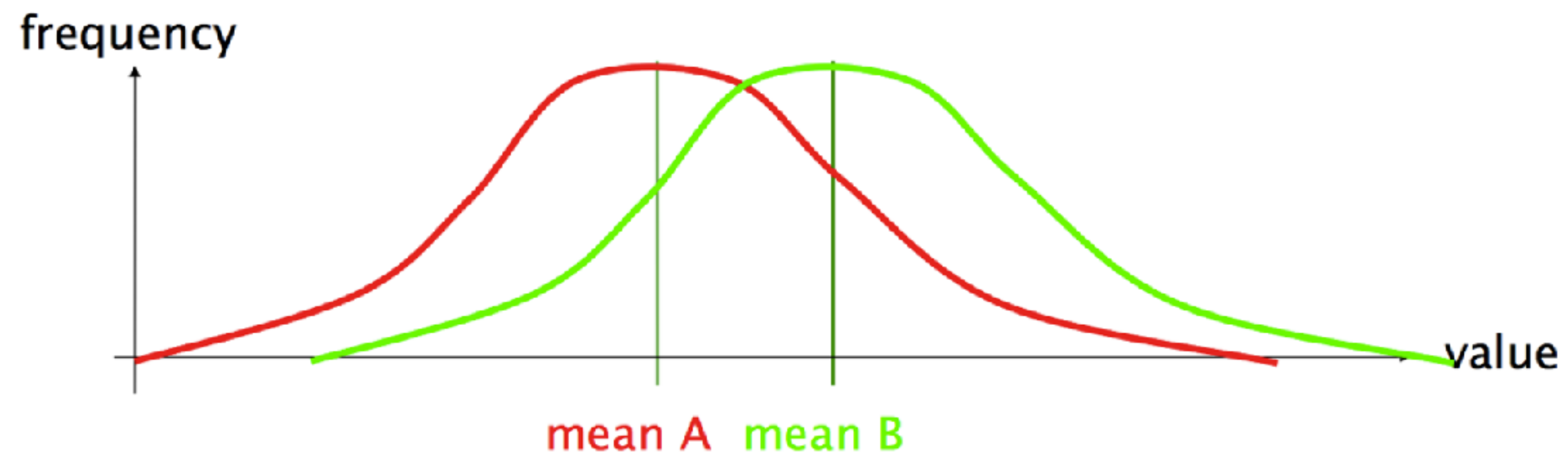
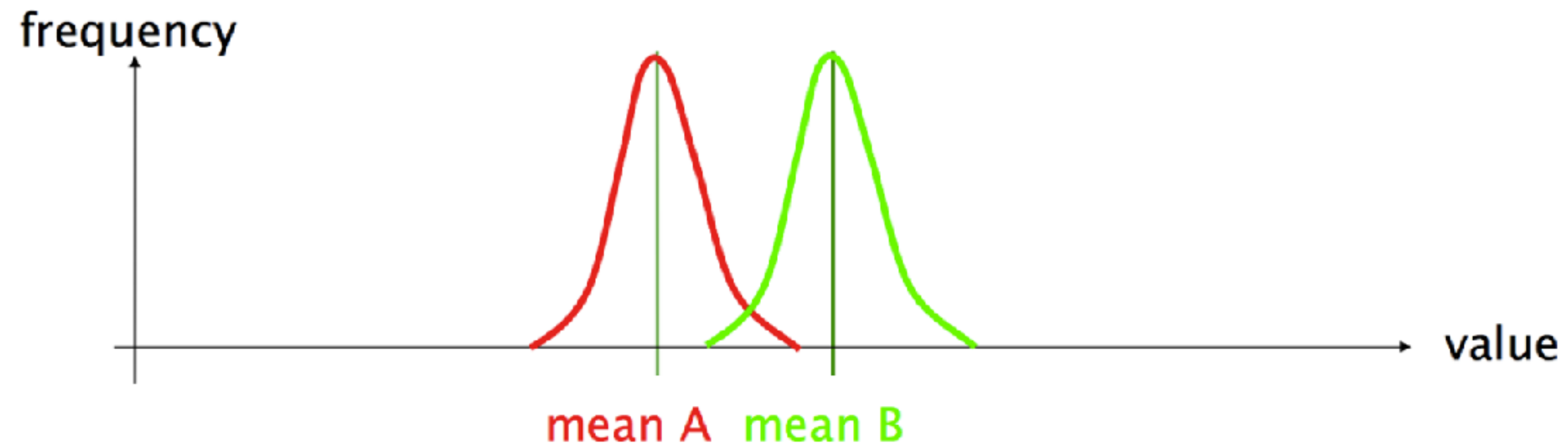
In HCl:

- $p < 0.05$ (= 5% probability) is a convention (or 0.01)
- a smaller p (e.g. 0.00001) doesn't make the result more significant.
- a significant result is different from an important result

Comparing values

via <http://www.medien.ifi.lmu.de/lehre/ws1213/mmi2/uebung/slides10.pdf>

Is there a significant difference between two measures?



DO NOT

If $p > 0.05$ say:

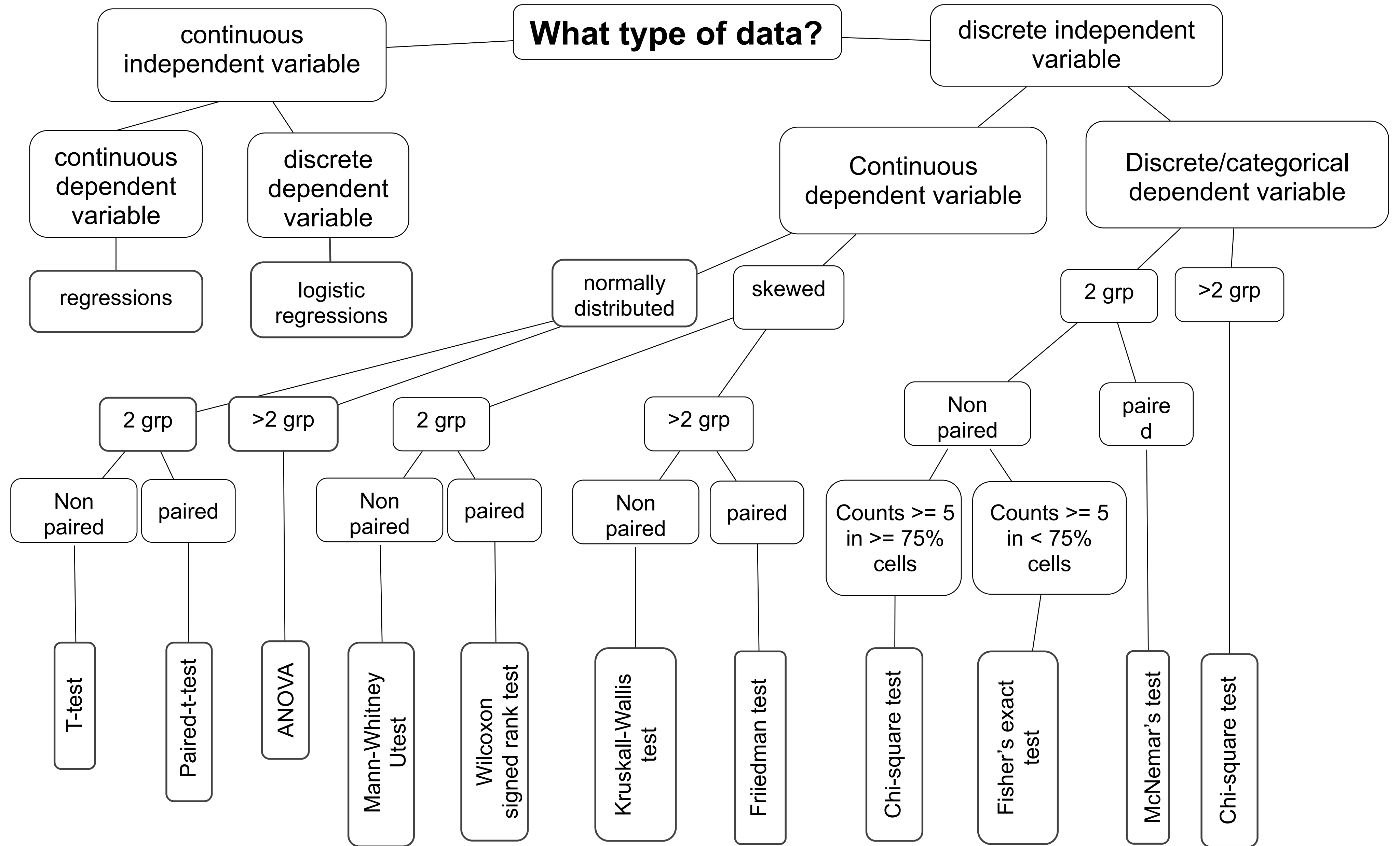
- “our tests showed that there was no difference”
- significant difference -> impact
- no significant difference -> nothing

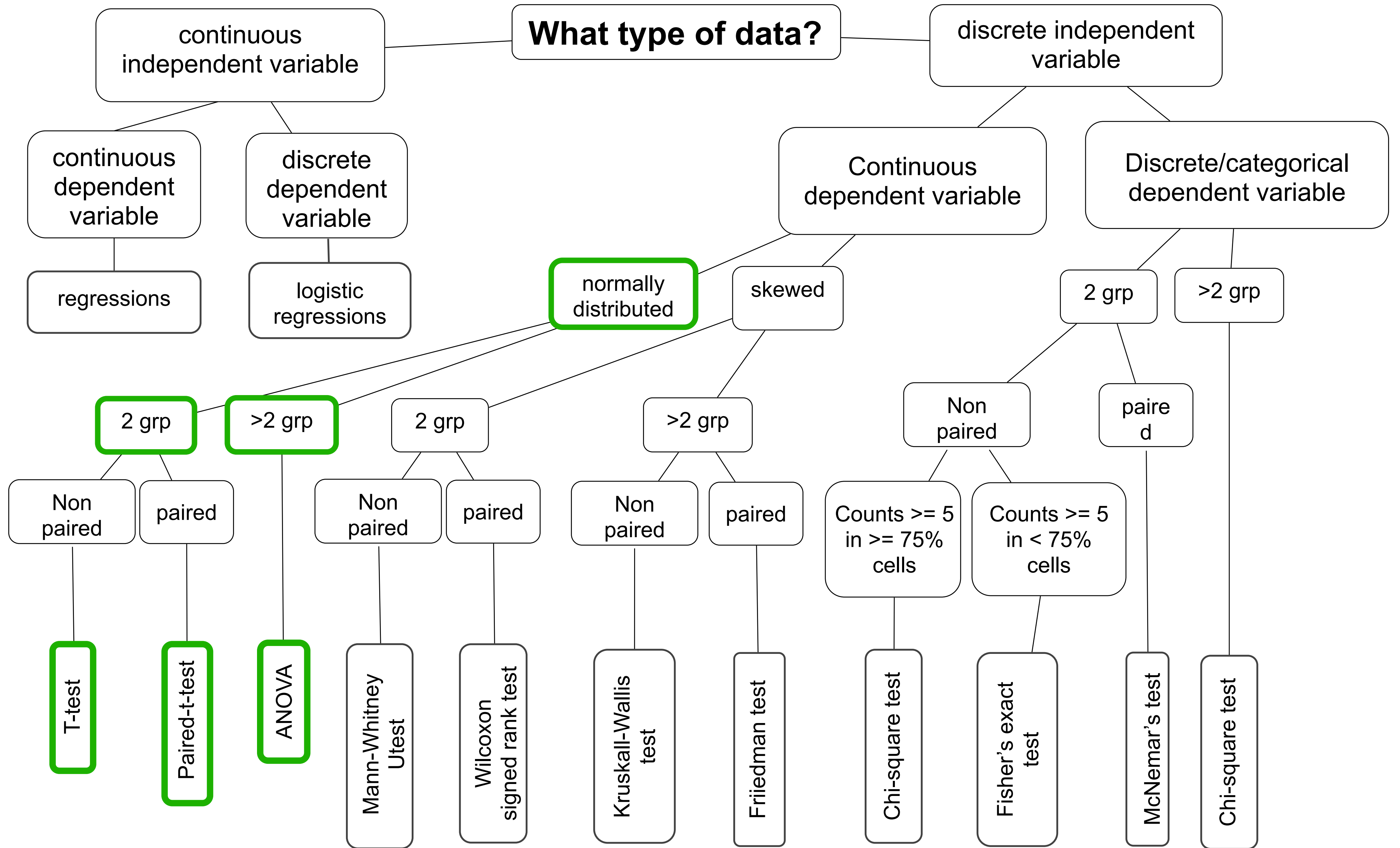
It only means that there is **not enough evidence to reject the null hypothesis** (it fails to reject the null hypothesis).

With significance testing, you cannot show that there is no difference!

Statistical analysis

- Practice
- Checking your data
- **Significance testing with t-tests**
- Significance testing with Anova
- Measuring effect sizes
- Beyond significance testing



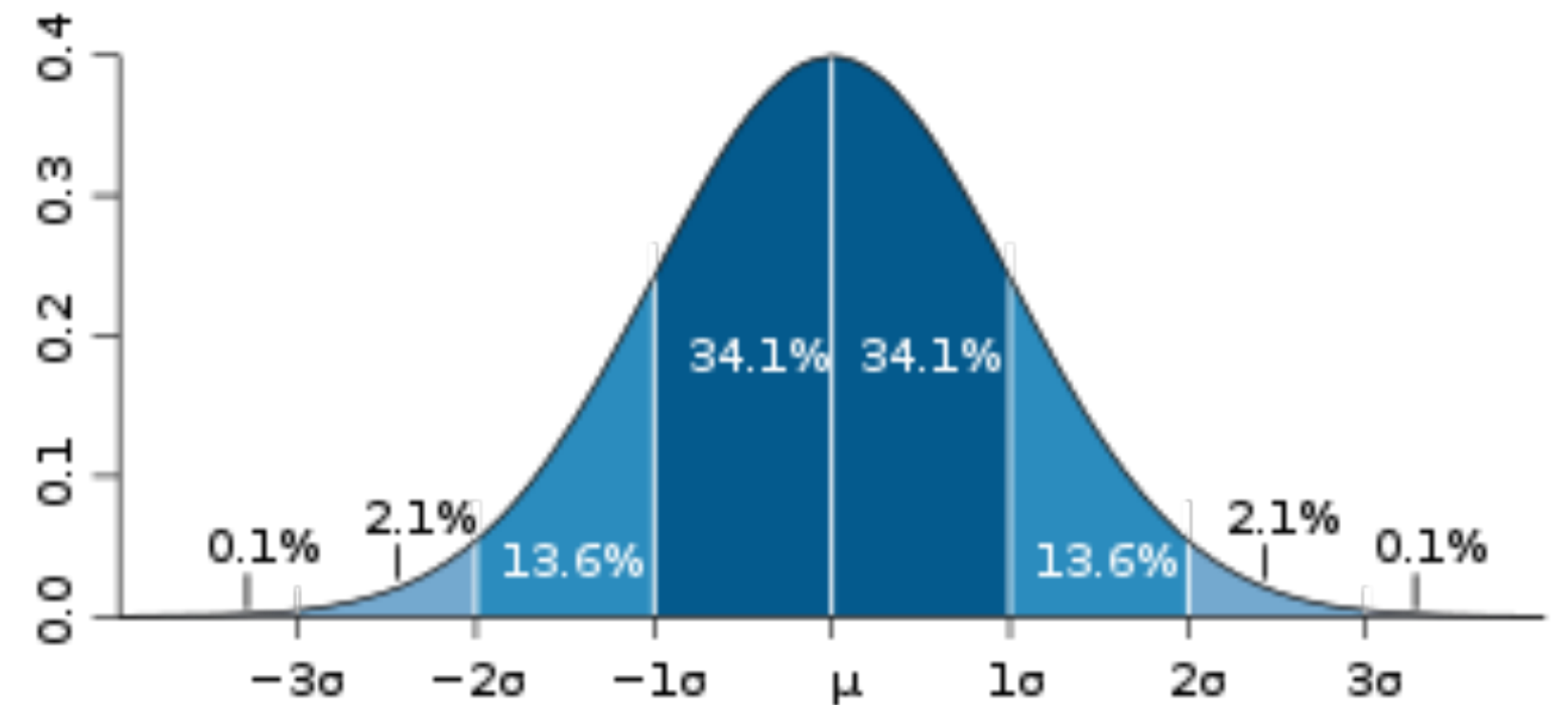


(Student's) t-test

Looks at the relationship between two data sets

Designed for

- small sample (= few measurements)
- unknown (mean and) standard deviation
- but has to be normally distributed



Parametric tests

Les tests paramétriques font généralement les hypothèses suivantes :

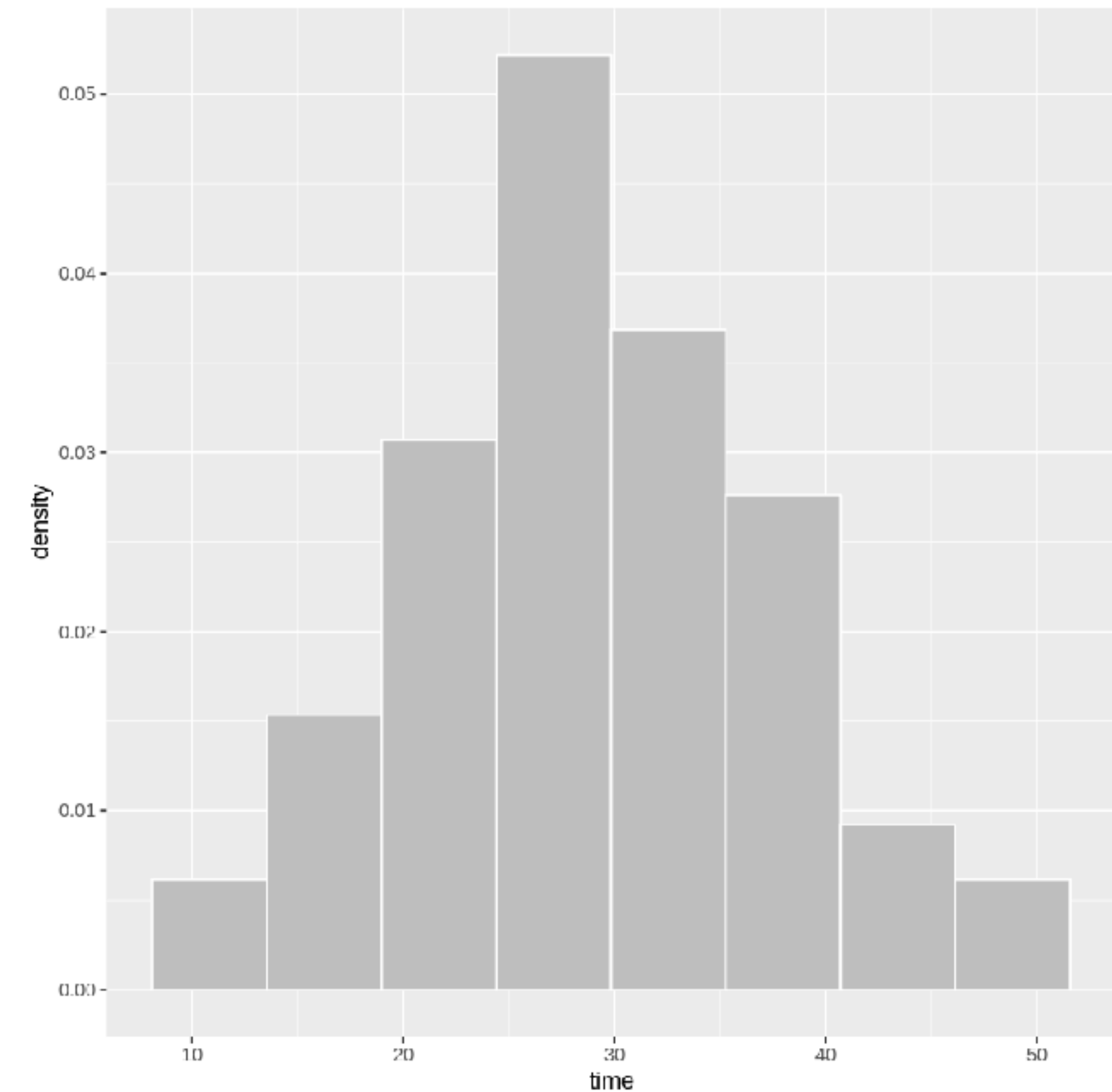
1. les points de données doivent être indépendants les uns des autres
2. supposer que les données sont normalement distribuées
3. homogénéité de la variance

Les tests non paramétriques ne font aucune hypothèse sur la distribution normale

Looking at your data distribution

<https://colab.research.google.com/drive/1Is8hWFtlnLXOoHqpU7C5jPR2vHLO8y3D?usp=sharing>

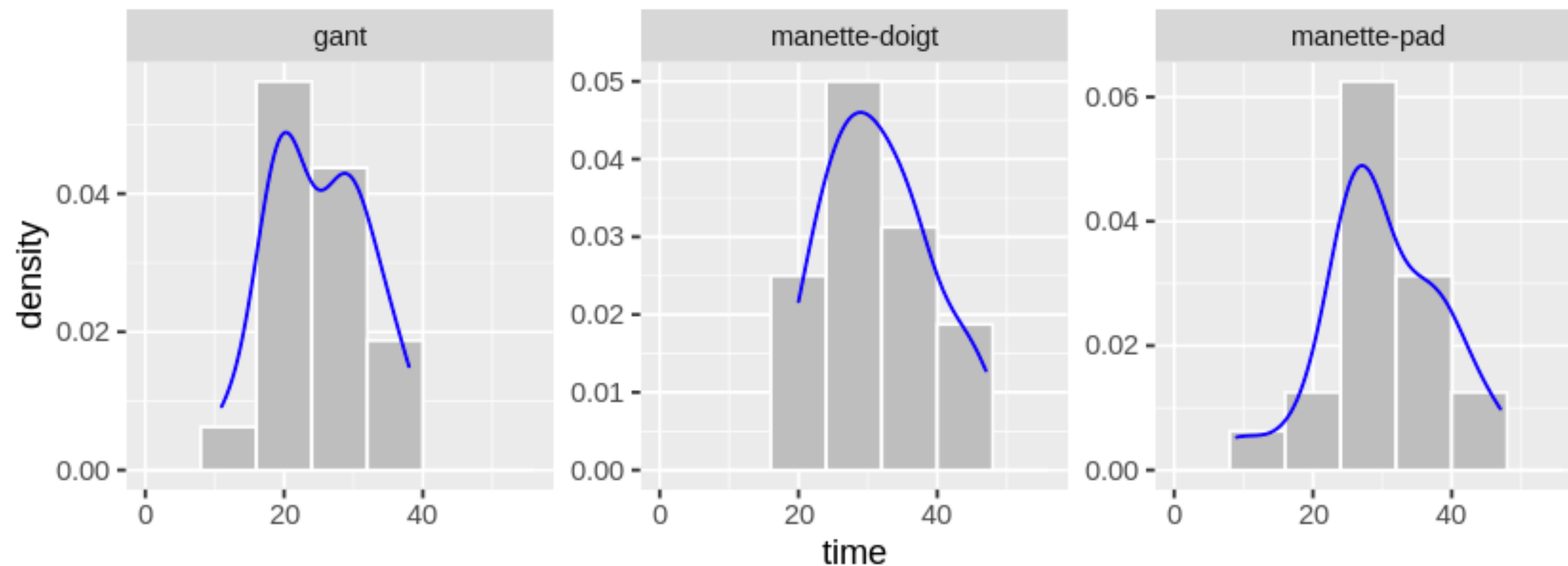
```
ggplot(data, aes(x = time)) +  
  geom_histogram(aes(y = ..density..),  
                 bins=8, # or specify manually :  
                 # breaks = seq(0, 60, by = 10),  
                 colour = "white", fill="grey75")
```



Looking at your data distribution

<https://colab.research.google.com/drive/1Is8hWFtlnLXOoHqpU7C5jPR2vHLO8y3D?usp=sharing>

```
ggplot(data, aes(x=time)) +  
  geom_histogram( aes(y=..density..),  
                 breaks = seq(0, 60, by = 8),  
                 colour = "white", fill="grey75") +  
  facet_wrap(~method, scales = "free") +  
  geom_density(aes(y=..density..), colour="blue")
```



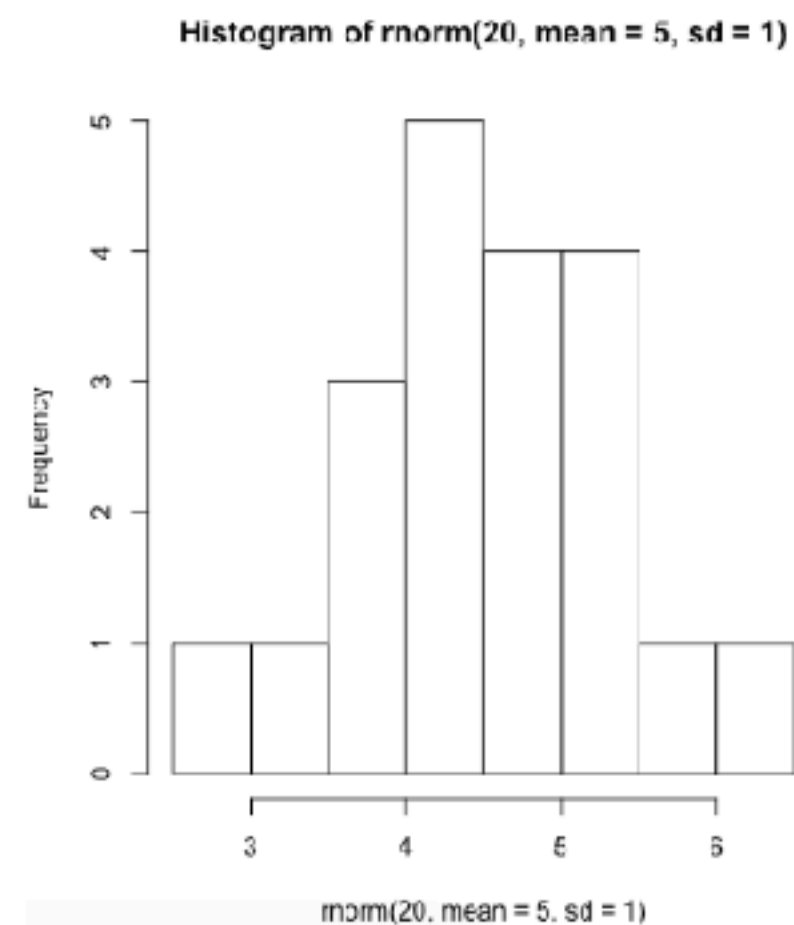
Shapiro-Wilk

pour tester la normalité

```
> shapiro.test(rnorm(20, mean=5, sd=1))
```

Shapiro-Wilk normality test

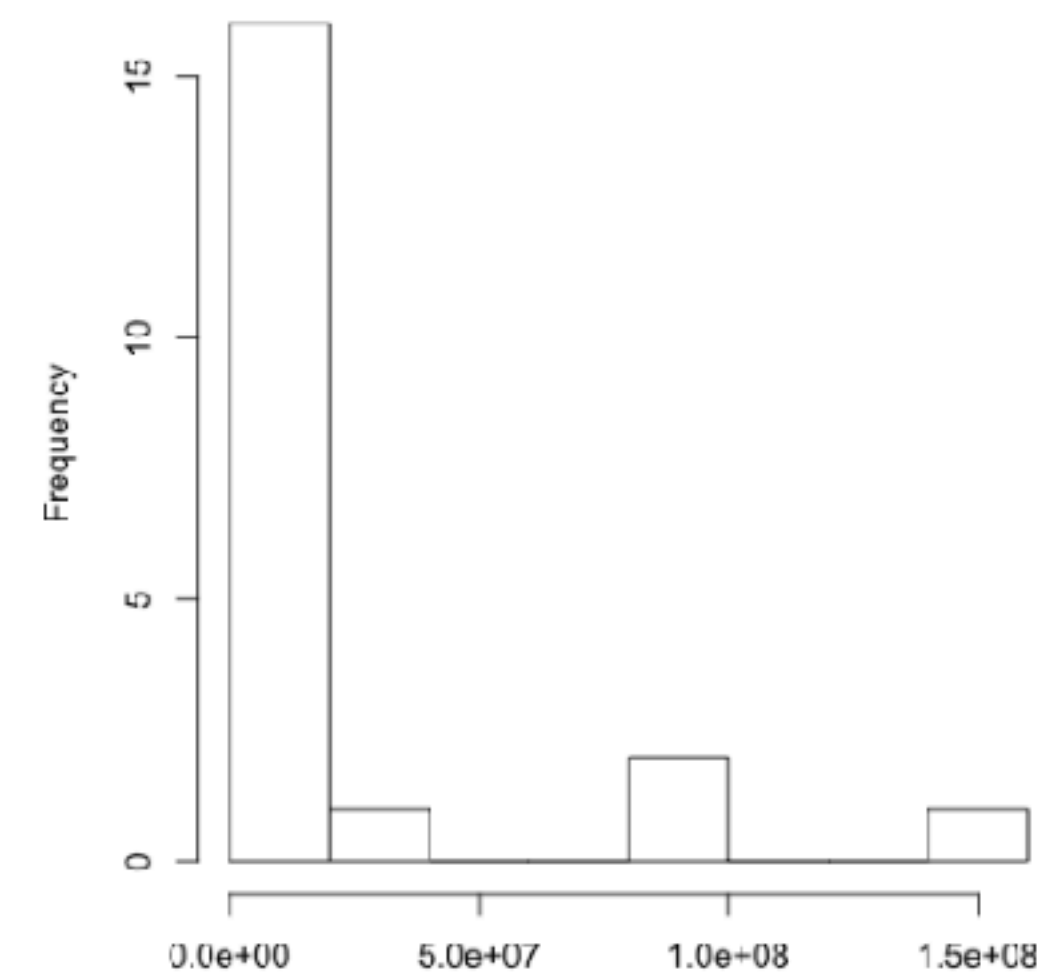
```
data:  rnorm(20, mean = 5, sd = 1)  
W = 0.96325, p-value = 0.6106
```



```
> shapiro.test(rnorm(20, mean=5, sd=1)^10)
```

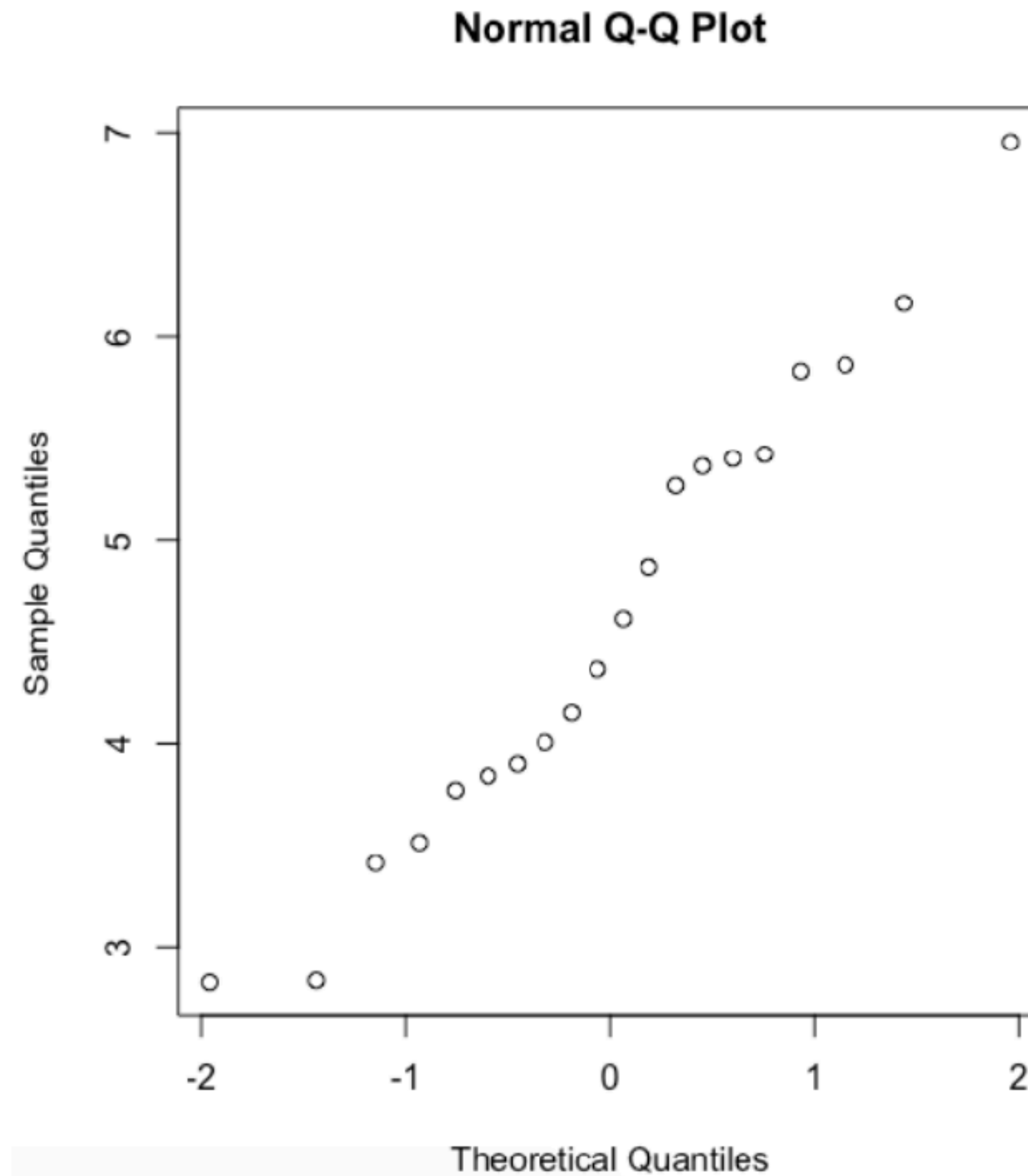
Shapiro-Wilk normality test

```
data:  rnorm(20, mean = 5, sd = 1)^10  
W = 0.81394, p-value = 0.001405
```



QQPlot

tracé quantile-quantile - contrôle visuel (subjectif)



t-test in R

paired or unpaired

between subject experiment

```
t.test( data[data["method"]=="gant",3],  
        data[data["method"]=="manette-pad",3])  
# unpaired
```

within subject experiment

```
t.test( data[data["method"]=="gant",3],  
        data[data["method"]=="manette-pad",3],  
        paired = TRUE))  
# paired
```


Un-paired t-test

Between subject

```
t.test( data[data["method"]=="gant",3], data[data["method"]=="manette-pad",3] )
```

```
data: data[data["method"] == "gant", 3] and data[data["method"] == "manette-  
pad", 3]
```

```
t = -2.0438, df = 36.271, p-value = 0.04828
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
-10.25904782 -0.04095218
```

```
sample estimates:
```

```
mean of x mean of y
```

```
24.95      30.10
```

“An unpaired student t-test showed no significant difference between the two devices.”

Paired t-test

Within subject

```
t.test( data[data["method"]=="gant",3], data[data["method"]=="manette-pad",3],  
        paired = TRUE))
```

```
data: data[data["method"] == "gant", 3] and data[data["method"] == "manette-  
pad", 3]
```

```
t = -5.6248 df = 19, p-value = 2.008e-05
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
-7.066334 -3.233666
```

```
sample estimates:
```

```
mean of the differences
```

```
-5.15
```

“A paired student t-test showed significant difference between the two devices (two-tailed $t(19)=-5.6248$, $p < 0.05$)”

Tails in t-tests

= effect direction

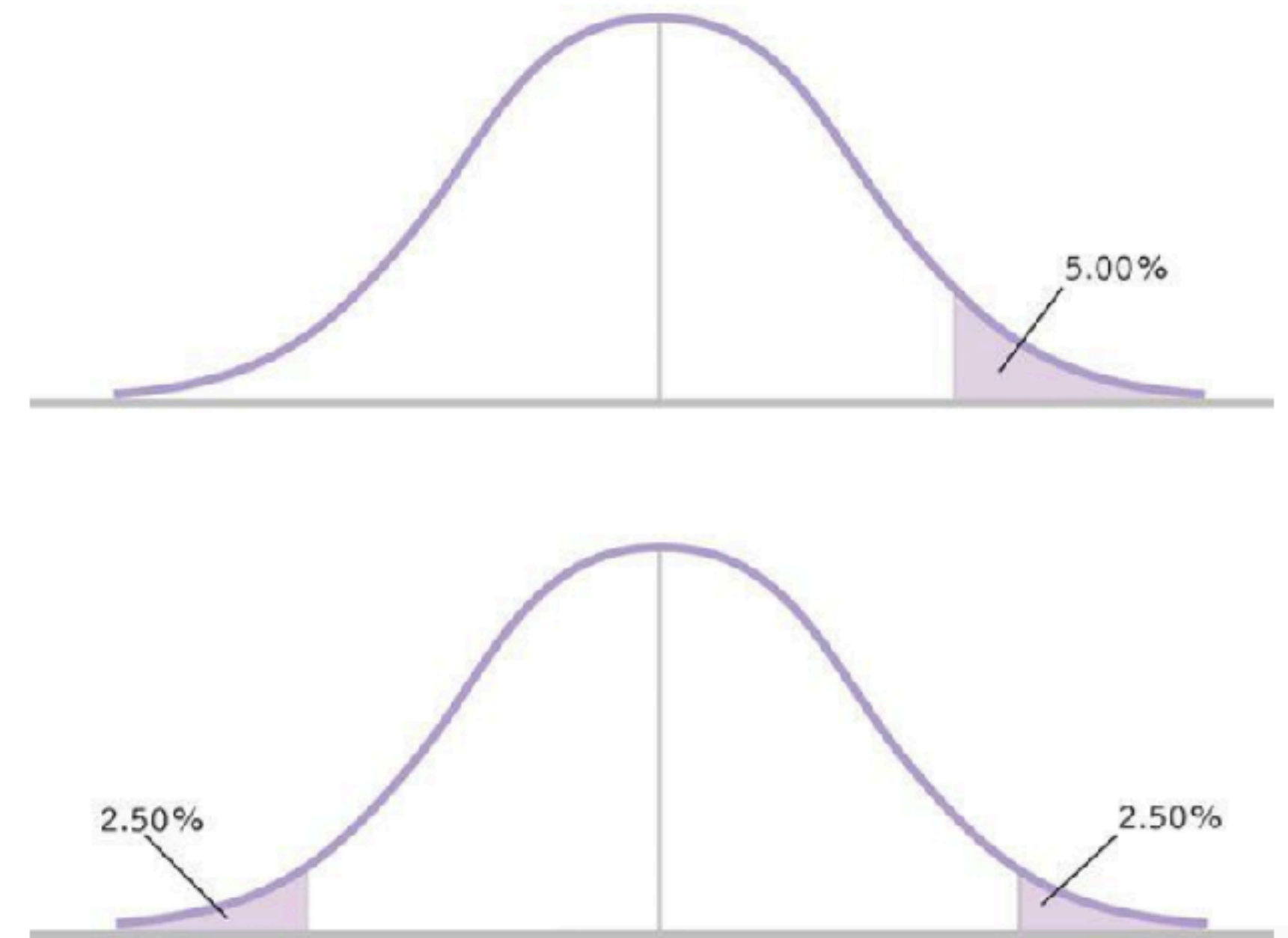
one-tail: only one side of the effect, i.e.

effect of shampoo 1 $>$ shampoo 2 (less)

or

effect of shampoo 1 $<$ shampoo 2 (greater)

two-tails: effect of drug 1 is $>$ and/or $<$ Drug 2



Summary so far

- Explain what is hypothesis testing
- Identify the limit of hypothesis testing (we cannot prove that things are similar)
- Explain what is a p value and a significance value
- Explain what is a t-test and when to use it
- Explain the difference between within and between subject studies
- Explain what is a Bonferroni correction and find the new significance level given an experimental design

Practice 2

Identify the best controller

What if we have more than two variables?



Practice 2

Identify the best controller



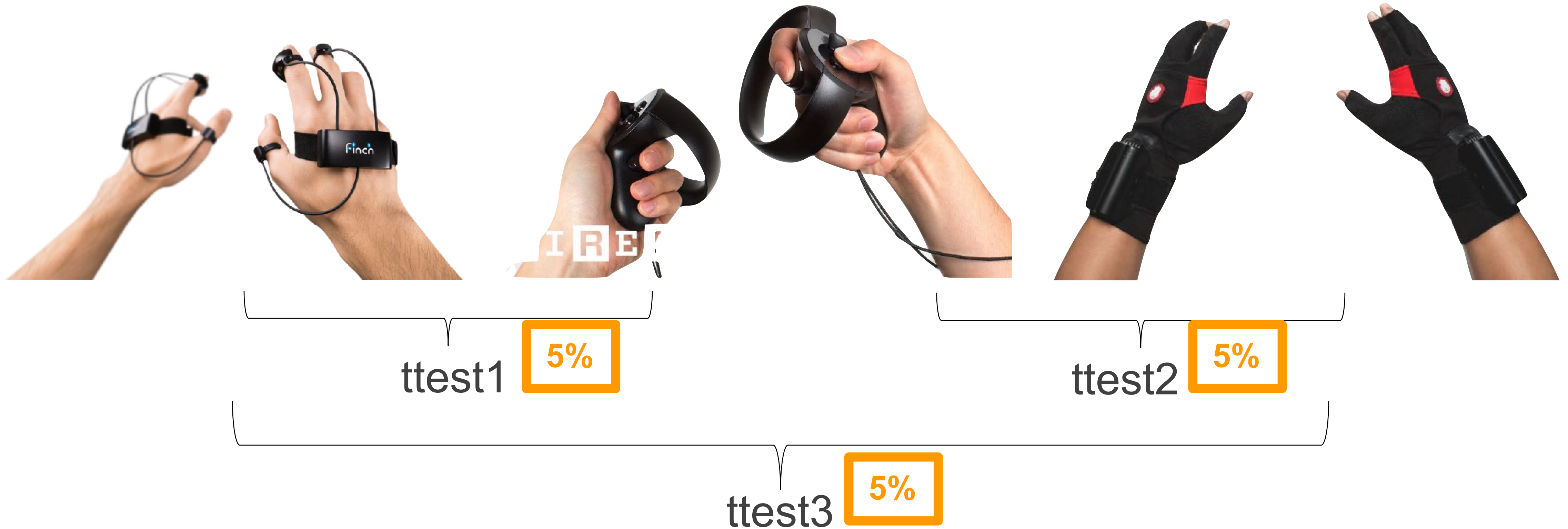
ttest1

ttest2

ttest3

Practice 2

Identify the best controller



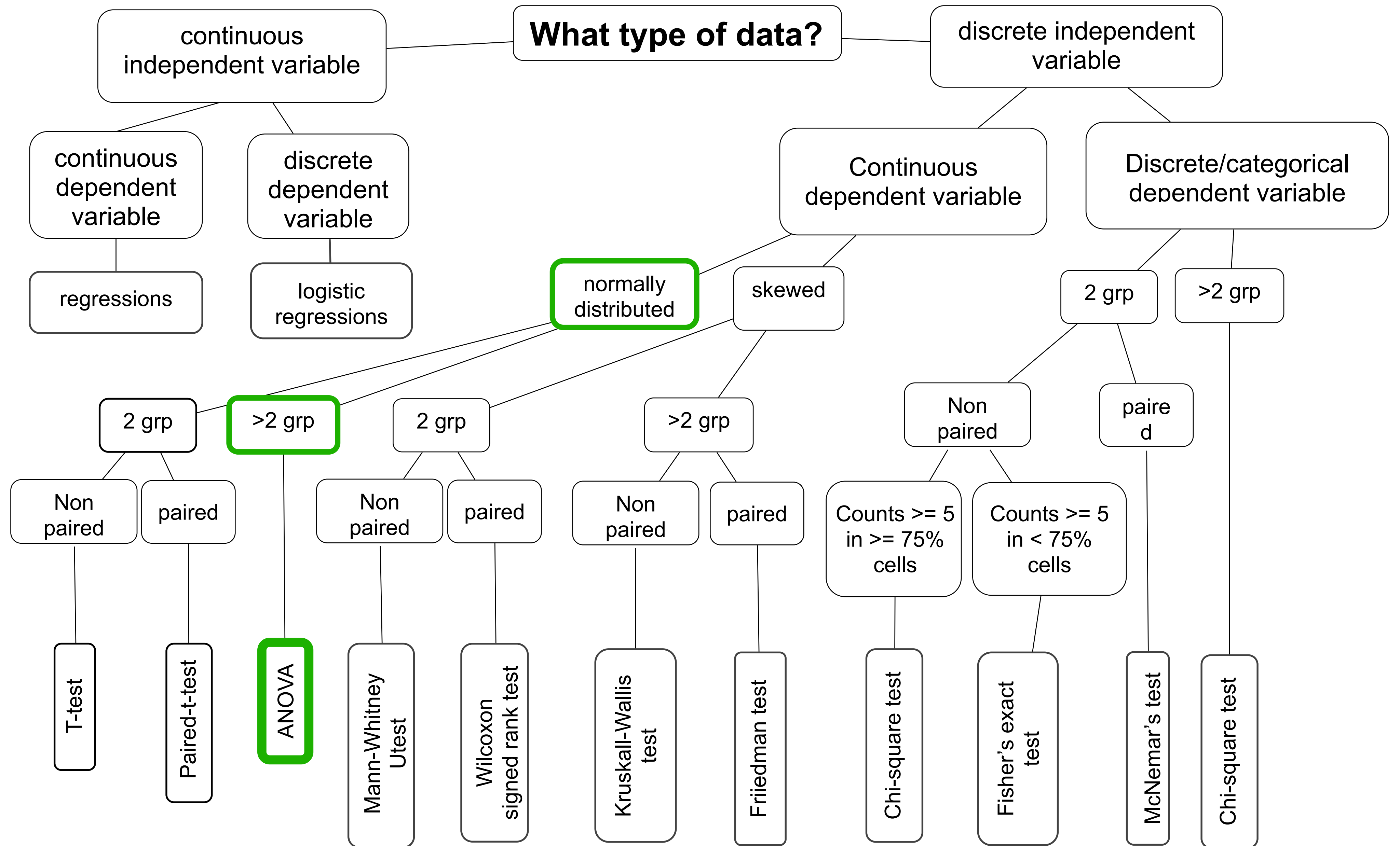
problem: any given test has a 5% chance of lying to you so when you use them multiple time you increase your risk of having errors (statisticians call this a “type I error”)

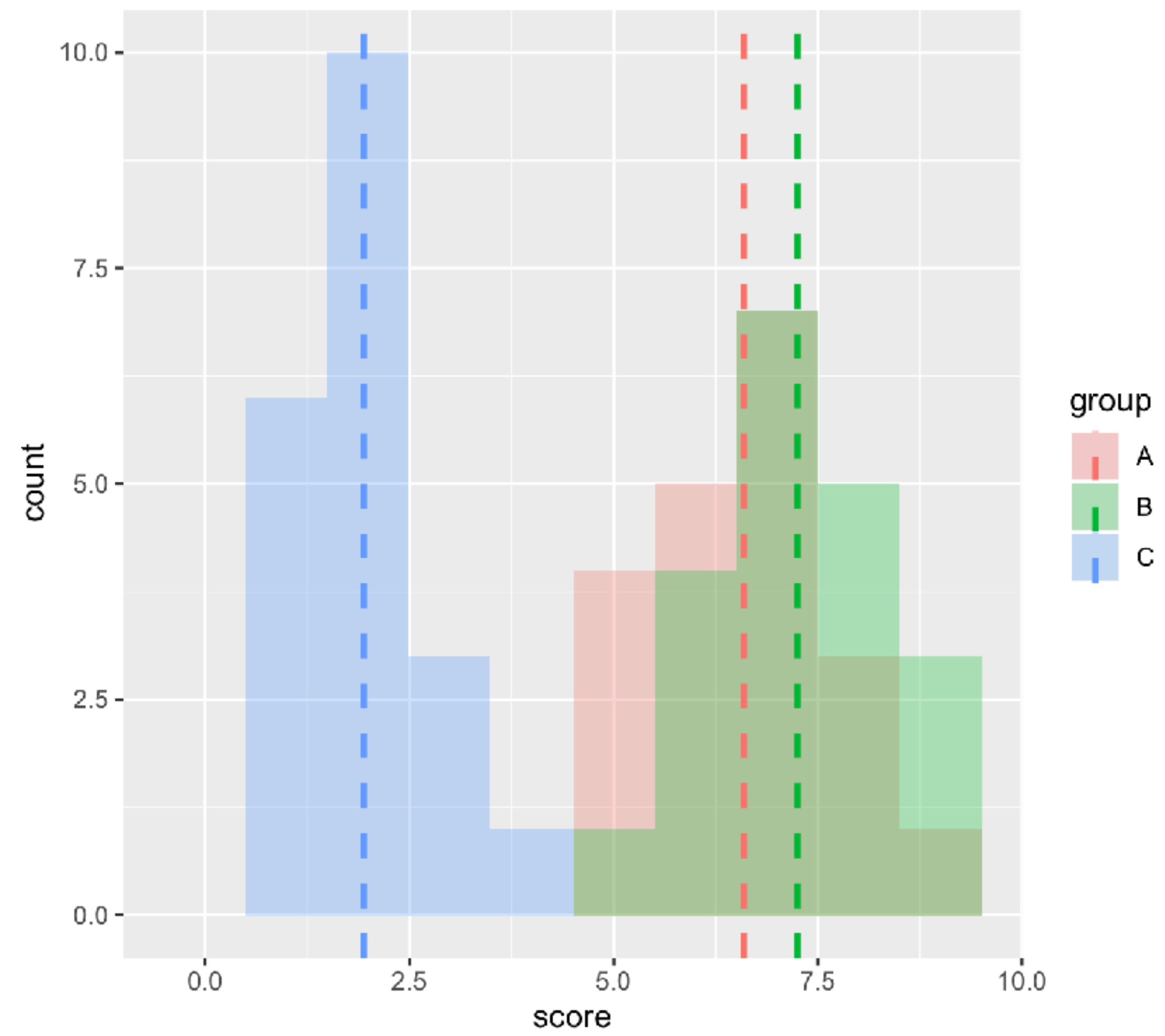
bonferroni correction

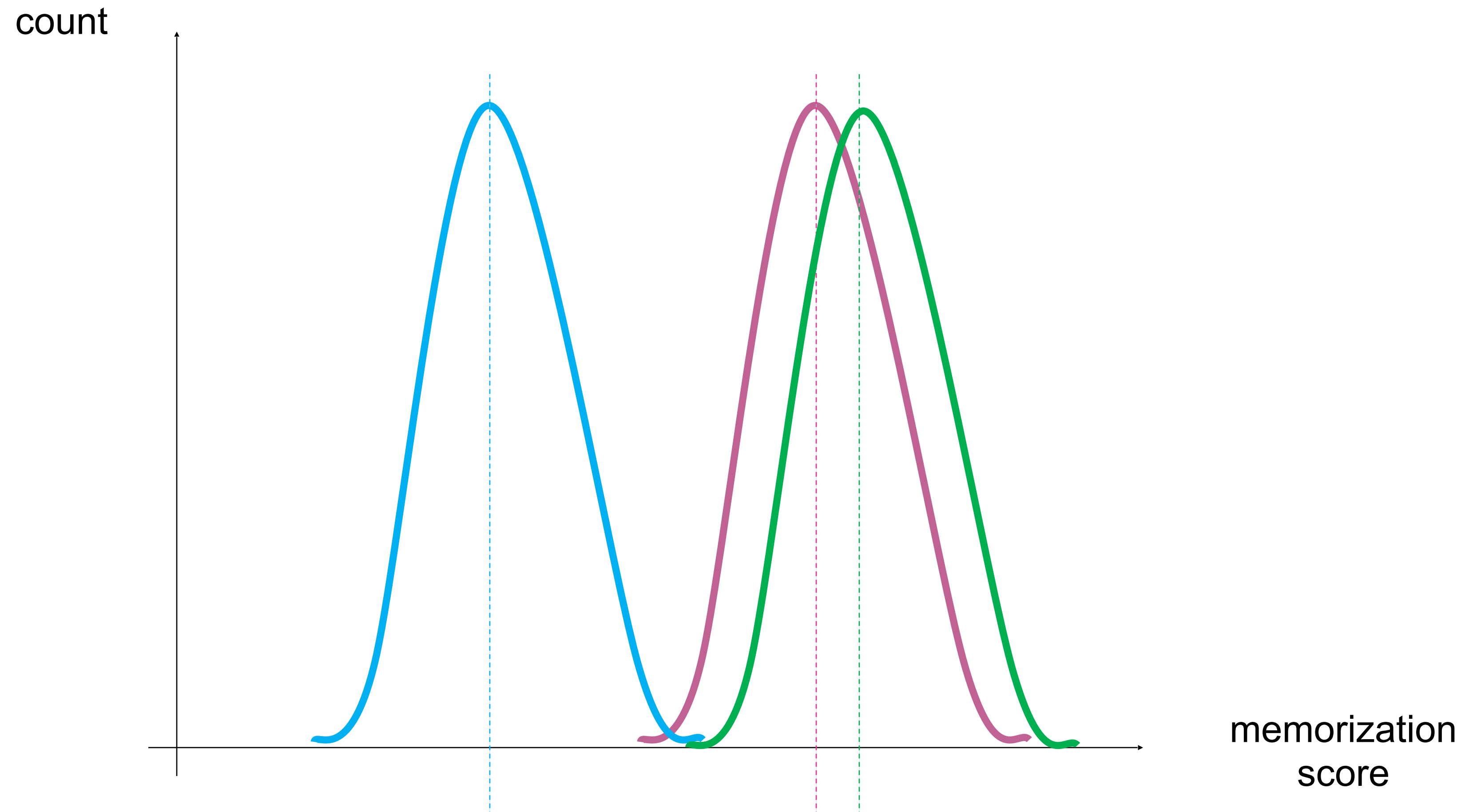
- when testing n hypotheses, test each one against $0.05/n$
- in our example we would need to use $0.05/3$ as a significant threshold instead of 0.05

Statistical analysis

- Practice
- Checking your data
- Significance testing with t-tests
- **Significance testing with Anova**
- Measuring effect sizes
- Beyond significance testing







(let's assume again these are normally distributed)

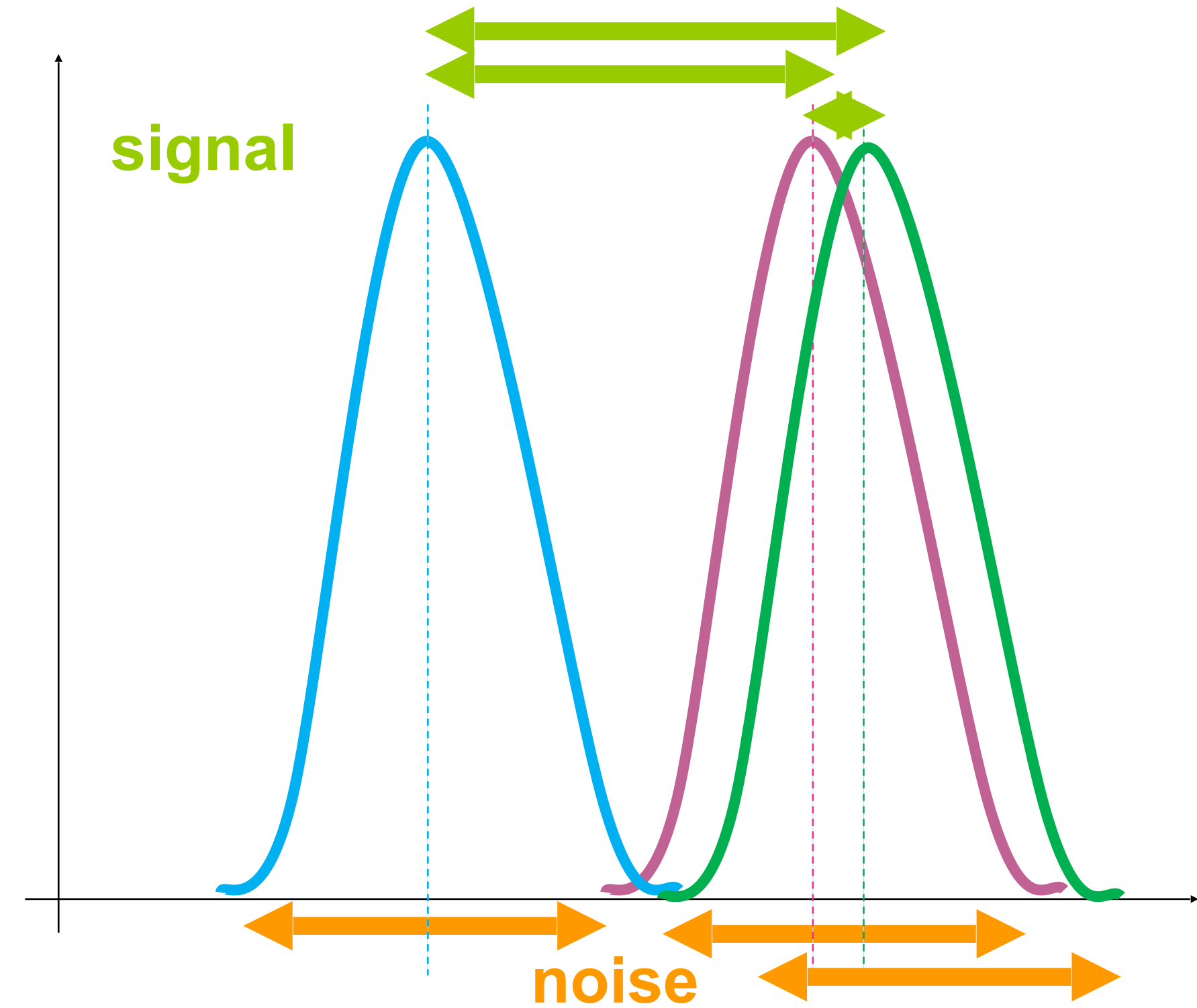
Any statistical tests

signal

noise

Anova

difference between group means
variability of groups



Anova types

Plan d'expérience	Variables indépendantes (IV)	Nombre de niveaux pour chaque IV	Types de test
Between-group	1	2	Independent-samples <i>t</i> test
	1	3 ou plus	One-way ANOVA
	2 ou plus	2 ou plus	Factorial ANOVA
Within-group	1	2	Paired-samples <i>t</i> test
	1	3 ou plus	Repeated measures ANOVA
	2 ou plus	2 ou plus	Repeated measures ANOVA
Mixed-group	2 ou plus	2 ou plus	Mixed-design ANOVA

Anova

```
results = afex::aov_ez(  
  data = data,  
  id = 'subject', # subject id column  
  dv = 'time', # dependent variable  
  within = c('method'), # within-subject independent variables  
  between = NULL, # between-subject independent variables  
  fun_aggregate = mean, # average multiple repetitions together for each  
  subject*condition  
  anova_table = list(es = 'ges') # effect size = generalized eta squared  
)
```

Anova Table (Type 3 tests)

Response: time

Effect	df	MSE	F	ges	p.value
1 method	1.54, 29.31	14.82	21.77 ***	.124	<.001

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1

Sphericity correction method: GG

Statistical analysis

- Practice
- Checking your data
- Significance testing with t-tests
- Significance testing with Anova
- **Measuring effect sizes**
- Beyond significance testing

Anova

```
results = afex::aov_ez(  
  data = data,  
  id = 'subject', # subject id column  
  dv = 'time', # dependent variable  
  within = c('method'), # within-subject independent variables  
  between = NULL, # between-subject independent variables  
  fun_aggregate = mean, # average multiple repetitions together for each  
  subject*condition  
  anova_table = list(es = 'ges') # effect size = generalized eta squared  
)
```

Anova Table (Type 3 tests)

Response: time

Effect	df	MSE	F	ges	p.value
1 method	1.54, 29.31	14.82	21.77 ***	.124	<.001

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1

Sphericity correction method: GG

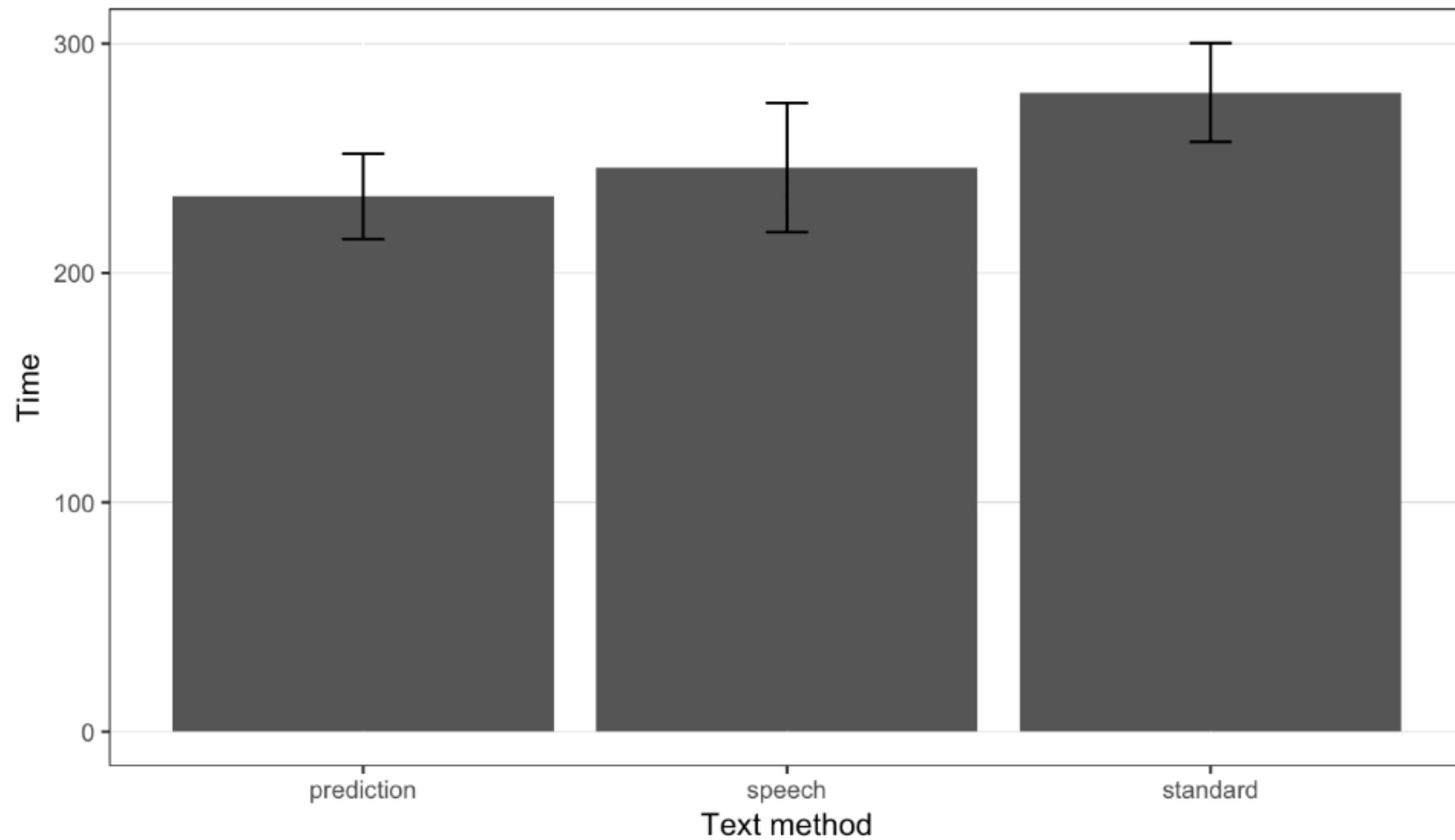
Effect sizes

Table 5.1 Effect sizes commonly used for null hypothesis significance testing and their values considered small, medium, and large effect sizes. This table was created based on existing literature (Cohen 1998; Field 2009; Mizumoto and Takeuchi 2008)

Statistical methods	Effect size	Small	Medium	Large
t-test	Cohen's d	0.2	0.5	0.8
ANOVA	η^2 and η_p^2	0.01	0.06	0.14
Non-parametric tests	R	0.1	0.3	0.5
Correlation	R	0.1	0.3	0.5

[Modern statistical methods for HCI, p.92]

95% confidence intervals (CI)



Post-hoc tests

When more than 2 levels, identify where the significant effect comes from.

Statistical analysis

- Practice
- Checking your data
- Significance testing with t-tests
- Significance testing with Anova
- Measuring effect sizes
- Beyond significance testing

Going further

<https://www.coursera.org/learn/designexperiments>

The screenshot shows the Coursera interface for the course 'Designing, Running, and Analyzing Experiments' by the University of California, San Diego. The page features a dark sidebar with navigation options: Home, Course Content, Assignments, Discussions, Classmates, and Course Info. The main content area includes a search bar, a course title, and a description. A sidebar on the right provides session information, including the current session (February 22 - May 1) and the upcoming session (March 21 - May 30), with a 'Switch sessions' button. A 'Help Center' button is located at the bottom right.

coursera Catalog Search catalog Institutions AT

UC San Diego

Designing, Running, and Analyzing Experiments

University of California, San Diego

Part of a 8-course series, the [Interaction Design Specialization](#)

About this Course

You will never know whether you have an effective user experience until you have tested it with users. In this course, you'll learn how to design experiments, how to run experiments, and how to analyze data from these experiments in order to evaluate and validate user experiences. You will work through real-world examples of experiments from the fields of IxD and HCI, understanding issues in experiment design and analysis. You will analyze multiple data sets using recipes given to you in the R statistical programming language -- no prior programming experience is assumed or required. By the end of the course, you will be able to knowledgeably design, run, and analyze your own experiments for putting empirical and statistical weight behind your designs.

- Subtitles available in English
- 9 weeks, 2-3 hours/week

You're currently enrolled in this session:
February 22 - May 1

Upcoming session:
March 21 - May 30

[Switch sessions](#)

Following session begins April 4

Financial Aid is available for learners who cannot afford the fee. [Learn more and apply.](#)

Course Ratings [Help Center](#)