

# Experimental design and analysis

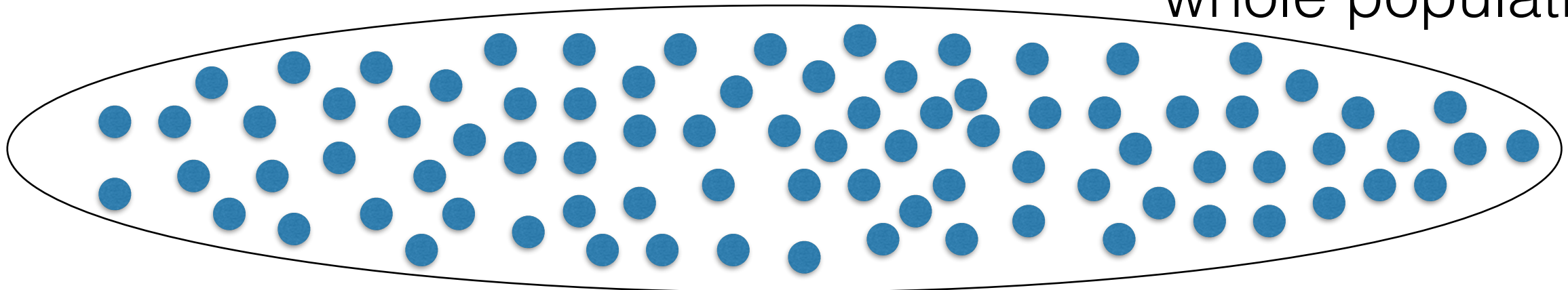
## Inferential statistics

<https://www.lri.fr/~appert/eval/>

Inferential statistics  
*use a sample to describe the  
whole population (with some  
uncertainty)*

# Experiment & inferential statistics

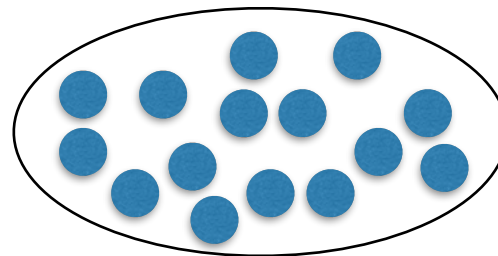
whole population



Experiment

data sample

(e.g., times for completing a pointing task)



inferential statistics use the sample to provide a description of the whole population that can be trusted only with a given probability

● one observation (e.g., time for completing a pointing task)

# Level of statistical significance

Inferential statistics uses the probability theory

A test based on inferential statistics outputs a  $p$  value. This  $p$  value provides an estimate of how often we would get the obtained statistic by chance, if in fact the null hypothesis were true.

How can we decide to reject a hypothesis? We decide on a level of significance, *i.e.*, the cutoff ( $\alpha$ ) before running the test. We reject the null hypothesis when  $p < \alpha$ .

Usually, the cutoff is set to 0.05 (as proposed by statistician Fisher)

# Null Hypothesis Significance Test

State the null hypothesis based on your research hypothesis  
( $H_0$ : the factor does not impact the measure value)

Decide the level of statistical significance  $\alpha$  (usually 0.05, i.e. 5%)

A Null Hypothesis Significance Test proceeds as follows:

- we know what the distribution of the value of a statistic (*e.g.*,  $t$ ,  $F$ , ...) (\*) is when the null hypothesis is true
- we compute this statistic for our sample of observations (*e.g.*,  $t$ ,  $F$ , ...) (\*)
- we use the known distribution to estimate the probability ( $p$  value) of observing this statistic if the null hypothesis were true for our sample
- If this probability is very low ( $p < \alpha$ ) , we reject the null hypothesis (based on the fact that there is very little chance to observe such a result if there was actually no difference).

*(\*) the specific statistic to consider depends on the type of your measure, the assumption about the distribution, and the type of design*

# Null Hypothesis Significance Test

State the null hypothesis based on your research hypothesis  
( $H_0$ : the factor does not impact the measure value)

Decide the level of statistical significance  $\alpha$  (usually 0.05, i.e. 5%)

A Null Hypothesis Significance Test proceeds as follows:

- we know what the distribution of the value of a statistic (*e.g.*,  $t$ ,  $F$ , ...) (\*) is when the null hypothesis is true
- we compute this statistic for our sample of observations (*e.g.*,  $t$ ,  $F$ , ...) (\*)
- we use the known distribution to estimate the probability ( $p$  value) of observing this statistic if the null hypothesis were true for our sample
- If this probability is very low ( $p < \alpha$ ) , we reject the null hypothesis (based on the fact that there is very little chance to observe such a result if there was actually no difference).

*(\*) the specific statistic to consider depends on the type of your measure, the assumption about the distribution, and the type of design*

# Null Hypothesis Significance Test

State the null hypothesis based on your research hypothesis  
( $H_0$ : the factor does not impact the measure value)

Decide the level of statistical significance  $\alpha$  (usually 0.05, i.e. 5%)

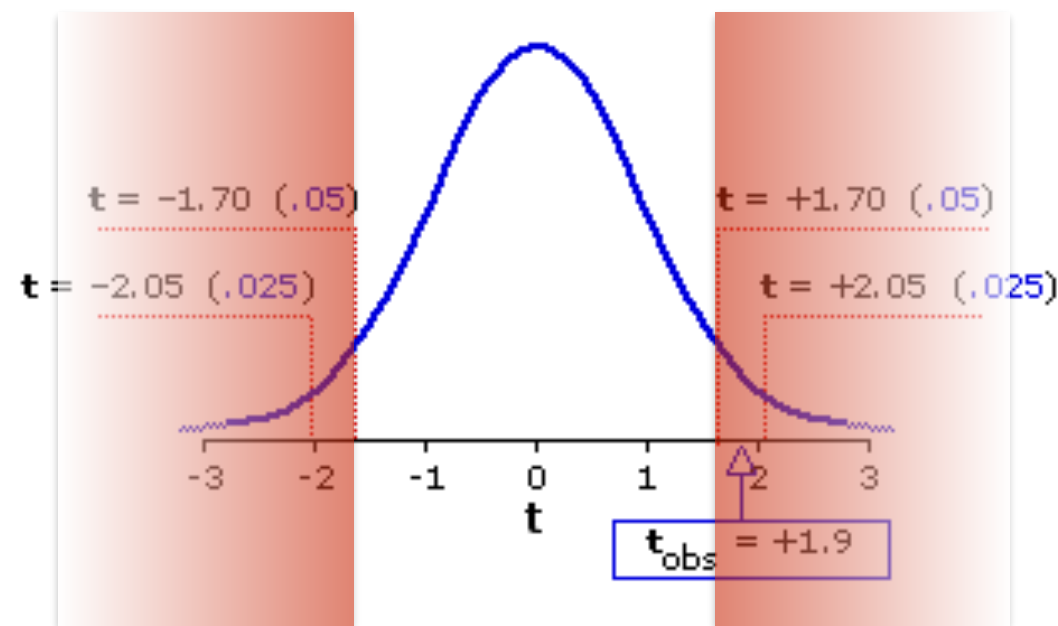
A Null Hypothesis Significance Test proceeds as follows:

- we know what the distribution of the value of a statistic (*e.g.*,  $r$ ,  $t$ ,  $F$ , ...) (\*) is  
when the null hypothesis is true

- we compute this statistic for our sample

- we use the known distribution to estimate the probability of observing this statistic if the null hypothesis were true  
the distribution of our statistic ( $t$  in that example)

- If this probability is very low ( $p < \alpha$ )  
that we know:  
the fact that there is very little chance of actually no difference).



(\*) the specific statistic to consider depends on the type of your measure, the assumption about the distribution, and the type of design

In red: values for our statistic that we are unlikely to observe if the null hypothesis is true (less than 5% of the cases if  $\alpha = 0.05$ ).

We decide to reject the null hypothesis in these cases (for example, if  $t = 1.9$ ).

# Experiment that we use as an example

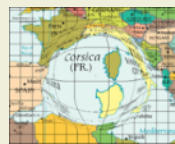
Pointing performance of different types of magnifying lenses

2 factors (5 x 5 design) - 10 participants

Lens type (5 levels):



**ML** (Manhattan Lens) ,



**FL** (Fisheye Lens) ,

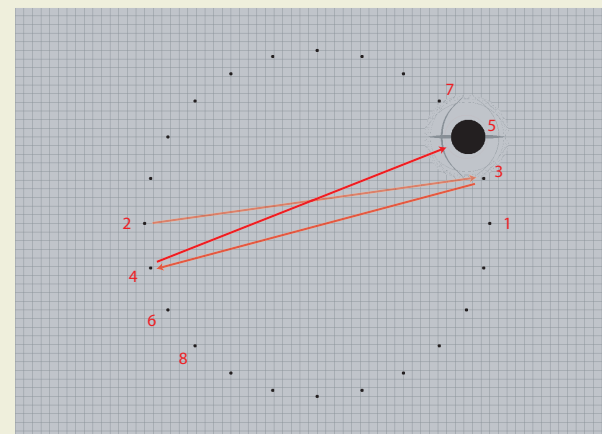
**SCF** , **SCB** ,



**BL** (Blending Lens)

Lens' magnification (5 levels): **2, 4, 6, 10, 14**

Task



target acquisition

Measure

Pointing time (in ms)



# Experiment that we use as an example

Pointing performance of different types of magnifying lenses

2 factors (5 x 5 design) - 10 participants

2 Factors

1 Measure

## Collected data

(log file `lens_experiment.csv`)

Participant	Block	Trial	Lens	Magnification	ID	PointingTime
1	4	0	FL	6	6.0035549	2297
1	4	0	FL	6	6.0035549	1485
1	4	0	FL	6	6.0035549	2000
1	4	0	FL	6	6.0035549	1843
1	4	0	FL	6	6.0035549	1813

...

10	2	9	SCF	6	6.0035549	2375
10	2	9	SCF	6	6.0035549	2359
10	2	9	SCF	6	6.0035549	2313
10	2	9	SCF	6	6.0035549	2453
10	2	9	SCF	6	6.0035549	2187
10	2	9	SCF	6	6.0035549	2875
10	2	9	SCF	6	6.0035549	2688

**Note:** When we analyze collected results, all logs are in a single file ( $\neq$  one file per participant)

# Inferential statistics

Testing the effect of nominal factor(s) on a continuous measure

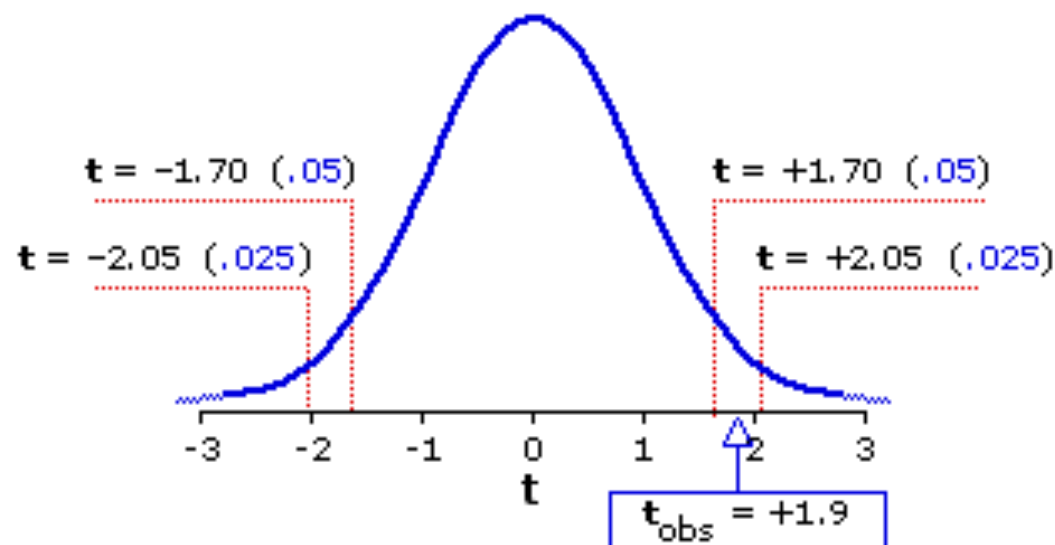
t-test  
anova

# t-test (Student test)

## When should we use a t-test?

When comparing two groups (i.e., when testing the effect of a nominal factor that has two levels)

A t-test consists of computing the  $t$  statistic (a function of the difference between the two means), and watching where our computed  $t$  ( $t_{obs}$  in figure below) lies in the theoretical  $t$ -distribution when the null hypothesis is true

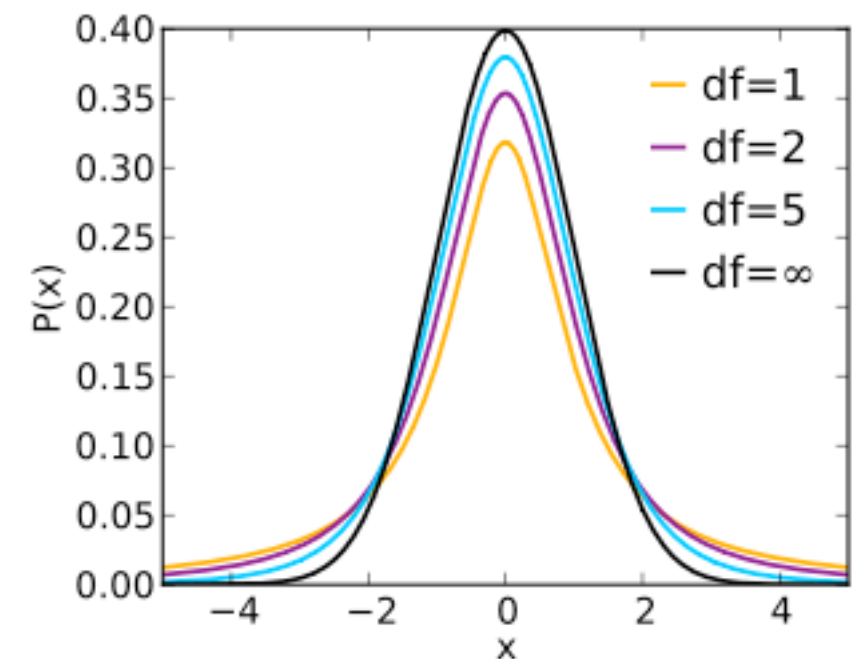


If our computed  $t$  is unlikely to happen ( $p < 0.05$ ) if there was no difference between the two groups ( $t$  lies in the tails of the theoretical distribution), we reject the null hypothesis.

# Theoretical t-distribution

A t-distribution shows the probability of observing a given t-value when the null hypothesis is true

There are several t distributions. The particular form of a t distribution is determined by its degrees of freedom ( $df$ )



$x$  is the value of t-ratio

# Paired vs. unpaired t-test

## paired t-test (or "repeated measures" t-test)

if the two groups are correlated. For example, participants have been measured under the two technique conditions (within-subject design)

## unpaired t-test

if the two groups are independent. For example, two groups of participants have been measured under the two different technique conditions (between-subject design)

**additional assumption for unpaired t-test:**

the variances of the population of the two groups are equal

# Degrees of freedom

## Paired t-test

size group 1 = size group 2 = size group

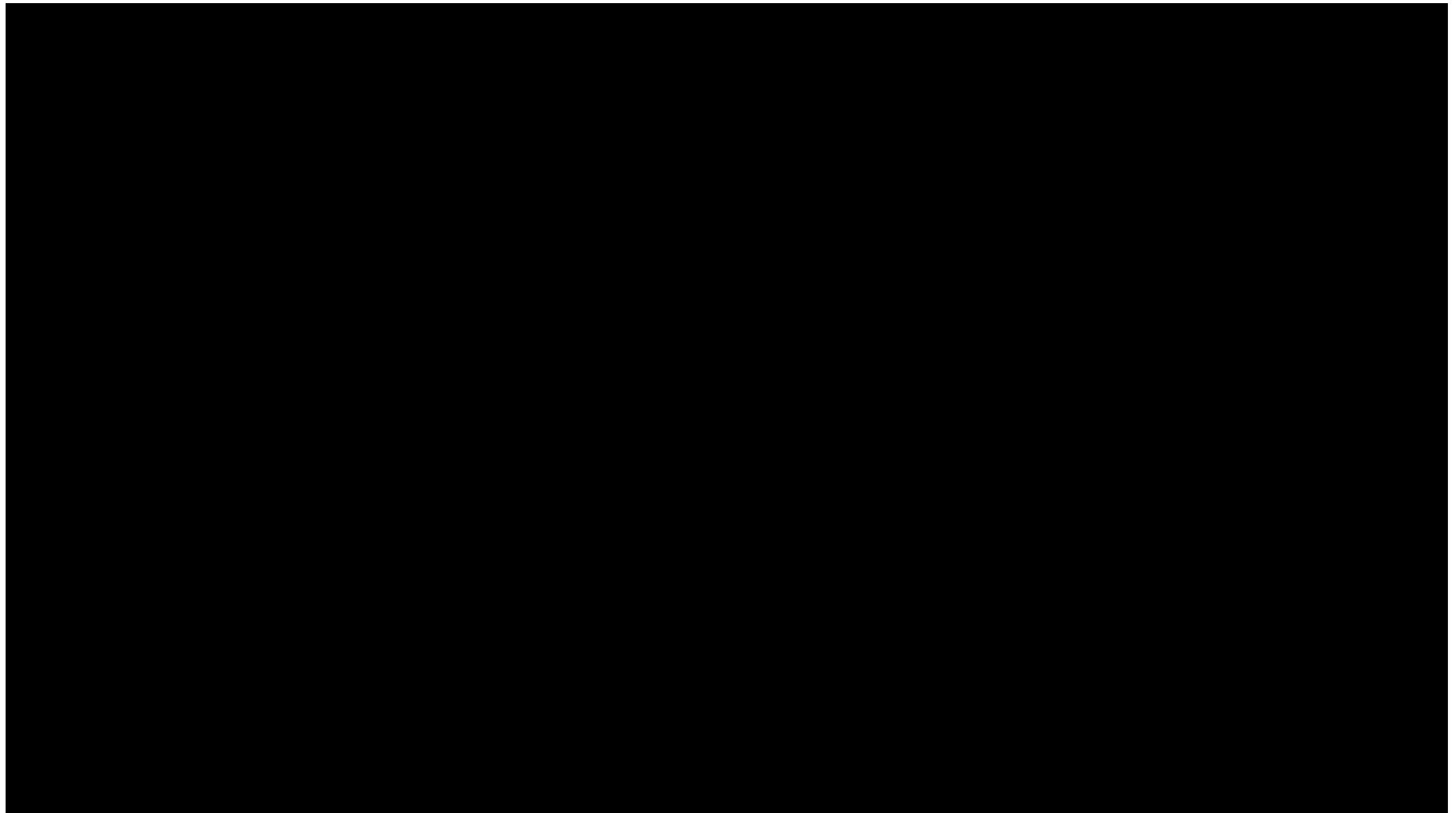
$$df = \text{size group} - 1$$

$$= \text{number of participants} - 1$$

## Unpaired t-test

$$df = \text{size group 1} + \text{size group 2} - 2$$

# t-test explained again



<https://www.youtube.com/watch?v=5Dnw46eC-0o>

# t-test and effect size

The  $t$  statistic allows you to tell if the difference in means is significant or not but does not give the size of the difference

Two kinds of effect size metrics for a t-test

Cohen's  $d$  (paired and unpaired)

Pearson's  $r$  (unpaired only)

	small size	medium size	large size
Cohen's $d$	0.2	0.5	0.8
Pearson's $r$	0.1	0.3	0.5



# Effect size for a paired t-test

Cohen's  $d$  for a paired t test

$$d = \frac{|M|}{SD}$$

where  $M$  is the mean of differences, and  $SD$  is the standard deviation of differences

It represents the difference in terms of standard deviations (normalized)

# Effect size for an unpaired t-test

Pearson's  $r$  for an unpaired t test

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

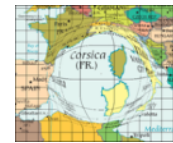
where  $t$  is the value of the test (t-ratio), and  $df$  is the number of degrees of freedom.

# t-test with `pingouin`



**ML** (Manhattan Lens)

VS



**FL** (Fisheye Lens)

We believe that the more occlusion lenses cause, the higher the pointing time. In particular, we make the following research hypothesis:

H: Pointing Time is greater with ML lenses than with FL lenses.

Null hypothesis: Pointing Time is the same with ML and FL.

We are in the context of a within-subject design (paired groups), we run a paired t-test using the `pingouin` library.

# t-test with pingouin

A paired t-test takes as input two vector of data points. There is one vector per condition (in our case: *FL* and *ML*), and each vector contains the mean measure value for each participant (in our case, 10 mean pointing times as we have 10 participants):

```
data_fl = data.query('Lens == "FL"')
data_fl = data_fl.groupby('Participant', as_index=False)['PointingTime'].mean()
display(data_fl)
data_ml = data.query('Lens == "ML"')
data_ml = data_ml.groupby('Participant', as_index=False)['PointingTime'].mean()
display(data_ml)
```

data\_fl

	Participant	PointingTime
0	1	2234.900000
1	10	2349.350000
2	2	2401.108333
3	3	2721.108333
4	4	2412.704167
5	5	2171.679167
6	6	2359.762500
7	7	2586.983333
8	8	2482.025000
9	9	2278.516667

data\_ml

	Participant	PointingTime
0	1	3091.016667
1	10	3899.479167
2	2	3429.166667
3	3	4932.154167
4	4	3533.025000
5	5	3261.845833
6	6	3284.375000
7	7	3277.541667
8	8	3205.920833
9	9	3071.362500

# t-test with pingouin

We use function `ttest` from library `pingouin`

```
ttest = pg.ttest(data_fl['PointingTime'], data_ml['PointingTime'], paired=True)
ttest
```

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	-7.490894	9	two-sided	0.000037	[-1430.59, -766.96]	2.669193	664.76	1.0

The  $t$  value for our observations is  $-7.49$ . This value is very unlikely to happen if the null hypothesis were true ( $p = 0.000037$ ). We thus reject the null hypothesis : the difference between the two groups is significant. Moreover, this difference is large ( $d=2.6$ )

```
mean_fl = data_fl['PointingTime'].mean()
mean_ml = data_ml['PointingTime'].mean()
print('Mean pointing time for Lens=FL: ', mean_fl)
print('Mean pointing time for Lens=ML: ', mean_ml)
```

```
Mean pointing time for Lens=FL: 2399.8137500000003
```

```
Mean pointing time for Lens=ML: 3498.5887500000003
```

## Final report:

We found a significant effect of factor Lens on PointingTime ( $t(9) = -7.5$ ,  $p < 0.001$ , Cohen's  $d=2.6$ ), with lens FL (2399ms on average) outperforming lens ML (3498ms on average).

# t-test with pingouin

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	-7.490894	9	two-sided	0.000037	[-1430.59, -766.96]	2.669193	664.76	1.0

*t*

*df*

*Cohen's d*

*p*

Mean pointing time for Lens=FL: 2399.8137500000003  
Mean pointing time for Lens=ML: 3498.5887500000003

## Final report:

We found a significant effect of factor Lens on PointingTime ( $t(9) = -7.5$ ,  $p < 0.001$ , *Cohen's d*=2.6), with lens FL (2399ms on average) outperforming lens ML (3498ms on average).

# Analysis of Variance (anova)

## When should we use an anova?

When comparing more than two groups (i.e., when testing the effect of a nominal factor that has three levels or more, or when testing the effect of more than one factor)

An anova test consists of computing the  $F$  statistic (a function of the difference between the means), and watching where our computed  $F$  lies in the theoretical F-distribution when the null hypothesis is true (i.e., no difference between any pair of groups)

# ANOVA test

ANOVA generalizes the t-test to an arbitrary number of groups

Making Student tests between pairs of groups does not give correct results as it increases the chance of making a Type I error

Group 1  $\neq$  Group 2 (95% confidence)

Group 1  $\neq$  Group 3 (95% confidence)

=> Group 1  $\neq$  Groups 2 and 3 (90% confidence)



# ANOVA and *df* (*degrees of freedom*)

The  $F$  statistic is a function of how much of the total observed variance is due to the variance between groups (*i.e.*, ANOVA separates the internal variability in each group and the variability between groups)

$$F = \frac{MS_{bg}}{MS_{wg}} = \frac{\text{a measure of the aggregate differences among the means of the } \mathbf{k} \text{ groups}}{\text{a measure of the amount of random variability that exists inside the } \mathbf{k} \text{ groups}}$$

$df_{bg} = k - 1$

$df_{wg} = \text{size group} - k$

$\vdots$

group means are set

also called  $df_{error}$  or  $df_{residuals}$

# ANOVA and effect size

The  $F$  statistic allows you to tell if there is at least one pair of groups that significantly differs but it says how large the difference is

Effect size for an anova is computed with  $\eta^2$  (eta squared)

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}$$

where  $SS_{effect}$  is the sum of square for the factor and  $SS_{total}$  is the total sum of square

	small size	medium size	large size
eta squared	0.01	0.06	0.14

# ANOVA test

The computation of  $F$  depends on your design

If the groups are paired (e.g., a within-subject design),  
use a repeated measures ANOVA test

One-way ANOVA

analyze one factor

n-way ANOVA

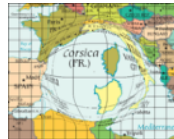
analyze two or more factors

# One-way anova-test with `pingouin`

Five different lenses:



**ML** (Manhattan Lens) ,



**FL** (Fisheye Lens) ,

**SCF** , **SCB** ,



**BL** (Blending Lens)

We believe that, because of their different design properties, their pointing time differs.

H: Pointing Time differs depending on the type of Lens used.

Null hypothesis: Pointing Time is the same regardless of the Lens used.

We test one factor (Lens) in the context of a within-subject design (paired groups), we run a repeated measures anova test using the `pingouin` library.

```
import pingouin as pg
```

# One-way anova-test with pingouin

We use function `rm_anova` from library `pingouin`

```
aovrm1way = pg.rm_anova(data=data, dv='PointingTime', within='Lens', subject='Participant')  
aovrm1way
```

	Source	ddof1	ddof2	F	p-unc	p-GG-corr	ng2	eps	sphericity	W-spher	p-spher
0	Lens	4	36	62.408495	1.077454e-15	0.000001	0.794287	0.337259	False	0.004401	0.000011

The  $F$  value for our observations is 62. This value is very unlikely to happen if the null hypothesis were true ( $p = 1.077454e-15$ ). We thus reject the null hypothesis : the difference at least two groups is significant. Moreover, this difference is large ( $\eta^2=0.79$ )

## Final report:

We found a significant effect of factor Lens on PointingTime ( $F(4,36) = 62, p < 0.001, \eta^2=0.79$ ).

+ chart (cf. next slides)

# One-way anova-test with `pingouin`

	Source	ddof1	ddof2	F	p-unc	p-GG-corr	ng2	eps	sphericity	W-spher	p-spher
0	Lens	4	36	62.408495	1.077454e-15	0.000001	0.794287	0.337259	False	0.004401	0.000011

## Final report:

We found a significant effect of factor Lens on PointingTime ( $F(4,36) = 62, p < 0.001, \eta^2=0.79$ ).

+ pairwise comparisons and charts for identifying and visualizing where these differences are (cf. next slides)

# Post-hoc tests

ANOVA says that there are significant effects

BUT does not say which group significantly differs from which other group

Post-hoc tests are used to find where the differences between groups are

The most common post hoc tests used in HCl are the Tukey's test and the multiple pairwise t-test (with potentially some correction method like Holm or Bonferroni)

# Post-hoc tests with pingouin

We use function `pairwise_tests` from library `pingouin`

```
posthoc = pg.pairwise_tests(data=data, dv='PointingTime', within=['Lens'], subject='Participant',
                             parametric=True, padjust='fdr_bh', effsize='hedges')
posthoc
```

	Contrast	A	B	Paired	Parametric	T	dof	alternative	p-unc	p-corr	p-adjust	BF10	hedges
0	Lens	BL	FL	True	True	4.345090	9.0	two-sided	1.863562e-03	2.070625e-03	fdr_bh	24.371	1.552562
1	Lens	BL	ML	True	True	-5.614069	9.0	two-sided	3.283107e-04	4.103884e-04	fdr_bh	104.046	-1.936085
2	Lens	BL	SCB	True	True	31.577207	9.0	two-sided	1.572325e-10	1.572325e-09	fdr_bh	3.503e+07	5.194532
3	Lens	BL	SCF	True	True	6.595215	9.0	two-sided	9.982369e-05	1.426053e-04	fdr_bh	285.907	1.777954
4	Lens	FL	ML	True	True	-7.490894	9.0	two-sided	3.728208e-05	7.456416e-05	fdr_bh	664.76	-2.556411
5	Lens	FL	SCB	True	True	9.341874	9.0	two-sided	6.287323e-06	1.571831e-05	fdr_bh	3096.594	3.432268
6	Lens	FL	SCF	True	True	0.645461	9.0	two-sided	5.347356e-01	5.347356e-01	fdr_bh	0.368	0.231850
7	Lens	ML	SCB	True	True	9.851450	9.0	two-sided	4.052381e-06	1.350794e-05	fdr_bh	4536.747	3.832725
8	Lens	ML	SCF	True	True	6.786651	9.0	two-sided	8.025315e-05	1.337553e-04	fdr_bh	344.493	2.647644
9	Lens	SCB	SCF	True	True	-11.895149	9.0	two-sided	8.297272e-07	4.148636e-06	fdr_bh	1.812e+04	-3.151003

$p=0.53 (> 0.05)$   
 $\Rightarrow$  the difference  
 between FL and SCF is  
 not significant

Each line presents the comparison between a pair of groups. A low *p-value* (p-corr) indicates that groups in the pair significantly differ. We can see here that all *p-values* are low ( $<0.05$ ) except for the pair (FL, SCF). So all pairs significantly differ, except (FL, SCF).

## Final report:

We found a significant effect of factor Lens on PointingTime ( $F(4,36) = 62, p < 0.001, \eta^2=0.79$ ). All pairs of lenses significantly differ (all *p*'s  $< 0.05$ ), except FL and SCF ( $p = 0.5$ ).

+ chart for identifying and visualizing where these differences are (cf. next slide)



# Visualizing descriptive stats for the groups that we compare

```
import plotly.express as px
```

We compute some descriptive stats for visualizing our dataset (function `summarizeDF` is provided in the notebook example on the website)

```
import math

def summarizeDF(df, factors, measure):
    [...]
    return summary
```

```
stats = summarizeDF(data, ['Lens'], 'PointingTime')
stats
```

	Lens	Mean	Count	Std	ci95_hi	ci95_lo	errorry_hi	errorry_lo
0	BL	2666.405417	2400	1400.267213	2722.426743	2610.384090	56.021327	56.021327
1	FL	2399.813750	2400	1266.792679	2450.495081	2349.132419	50.681331	50.681331
2	ML	3498.588750	2400	3156.100429	3624.856745	3372.320755	126.267995	126.267995
3	SCB	1881.055417	2400	658.170719	1907.387246	1854.723588	26.331829	26.331829
4	SCF	2359.847083	2400	1162.838646	2406.369464	2313.324703	46.522380	46.522380

# Visualizing descriptive stats for the groups that we compare

```
import plotly.express as px
```

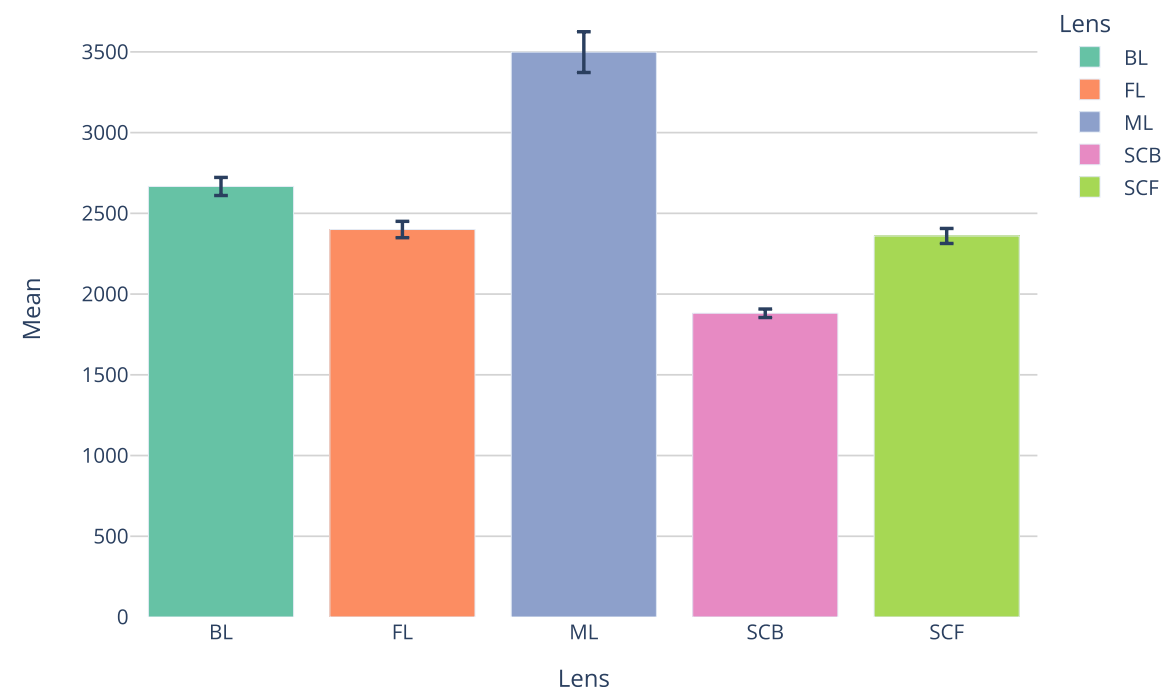
We use function `bar` from library `plotly.express` to produce a bar chart depicting the different groups

```
nice_color_palette = ['#66c2a5', '#fc8d62', '#8da0cb', '#e78ac3', '#a6d854']

fig = px.bar(stats, x='Lens', y='Mean', error_y="error_y_hi", error_y_minus="error_y_lo", color='Lens',
color_discrete_sequence=nice_color_palette)

fig.update_layout({
    'plot_bgcolor' : 'rgba(0,0,0,0)'
})
fig.update_yaxes(showgrid=True, gridwidth=1, gridcolor='LightGray')

fig.show()
```



See documentation for customizing your chart  
<https://plotly.com/python/bar-charts/>

# Visualizing descriptive stats for the groups that we compare

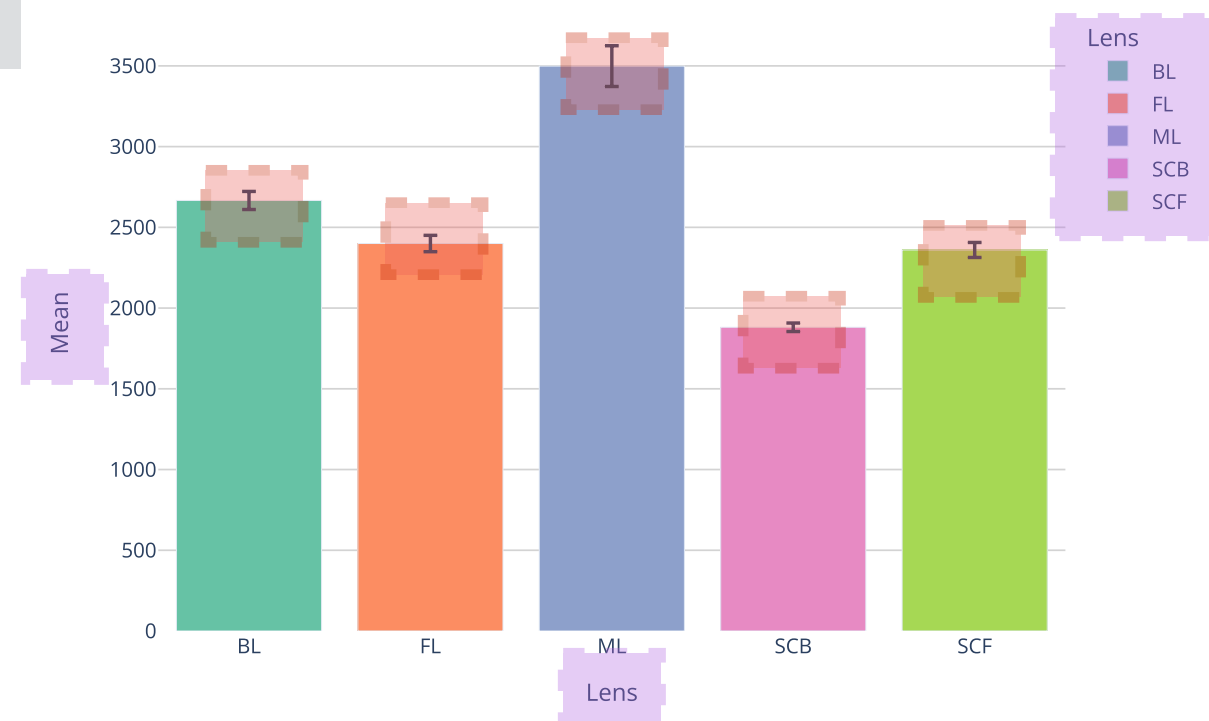
```
import plotly.express as px
```

```
nice_color_palette = ['#66c2a5', '#fc8d62', '#8da0cb', '#e78ac3', '#a6d854']

fig = px.bar(stats, x='Lens', y='Mean', error_y="error_hi", error_y_minus="error_lo", color='Lens',
color_discrete_sequence=nice_color_palette)

fig.update_layout({
    'plot_bgcolor' : 'rgba(0,0,0,0)'
})
fig.update_yaxes(showgrid=True, gridwidth=1, gridcolor='LightGray')

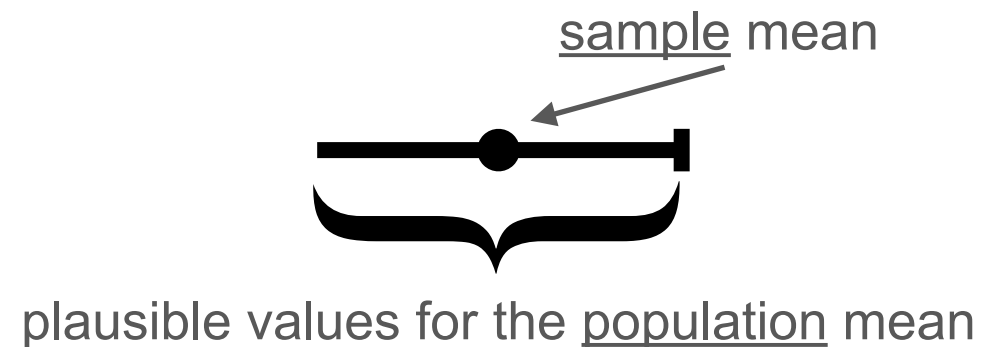
fig.show()
```



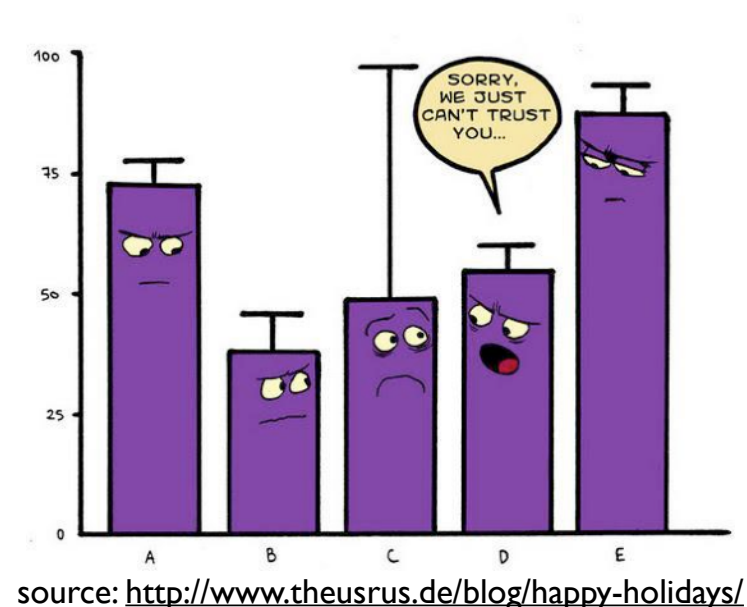
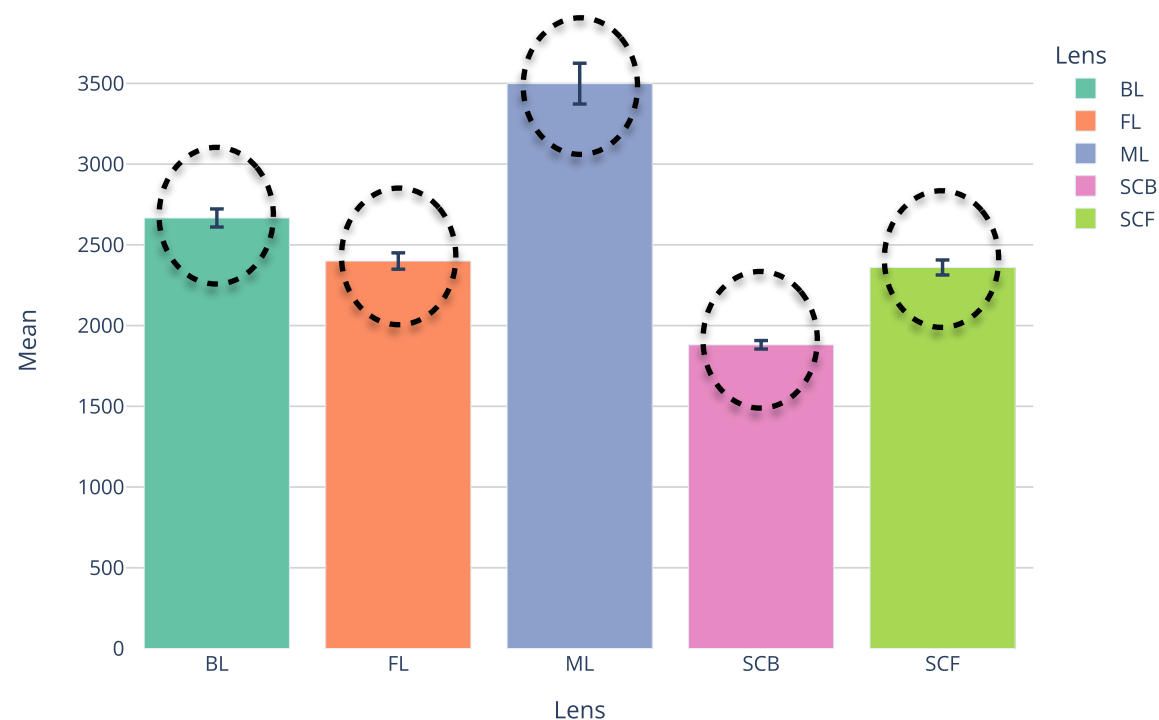
# Confidence Intervals

Error bars on bar charts usually represent *Confidence Intervals (CI)*. A CI is a range of plausible values for the population mean, calculated from a sample

e.g., a CI with a 95% confidence level has a 95% chance of capturing the population mean. (This also means that our confidence interval *might not* include the population mean...)



A confidence interval is a function of the *Standard Error of the mean (SE)*, i.e., the estimate of the standard deviation that would be obtained from the means of a large number of samples drawn from that population



# Two-way anova-test

A two-way anova test analyzes the effect of two factors at the same time.

Why analyzing the effect of more than one factor at the same time?

Because of possible interaction effects

# Interaction effect: a simple example

measure

factor 1

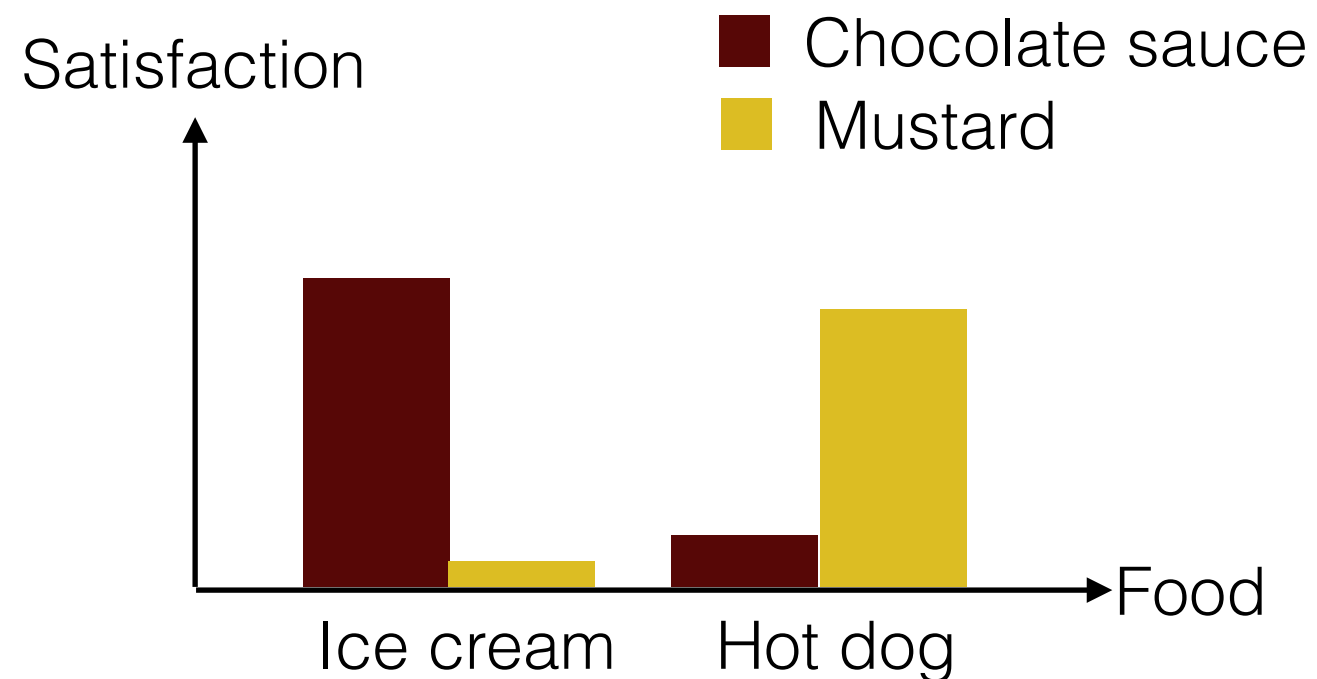
factor 2

Satisfaction = Food x Condiment

Experiment: ask participants “How much do you appreciate condiment X on food Y?”

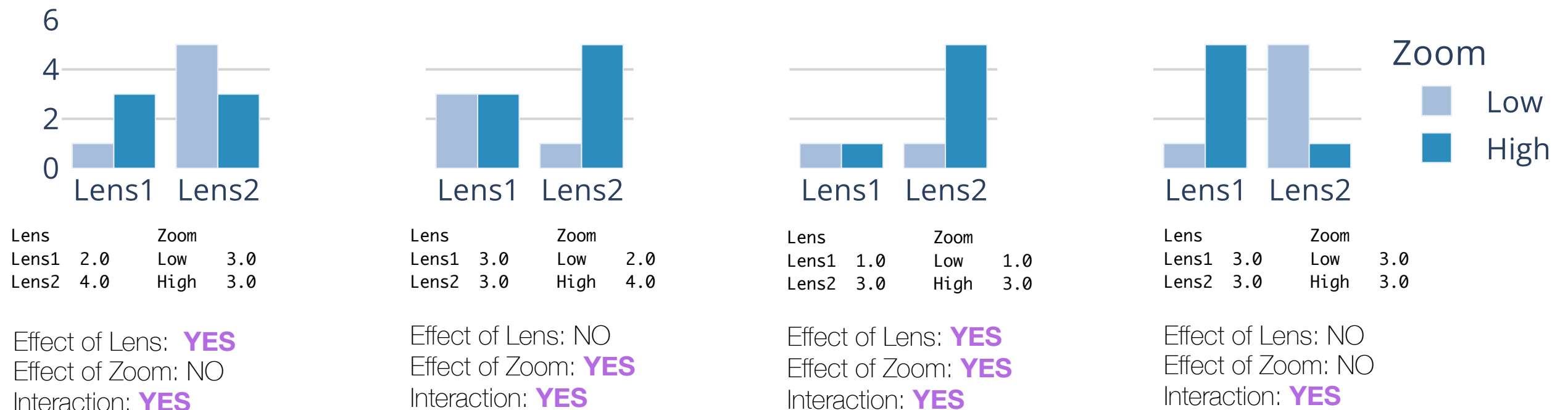
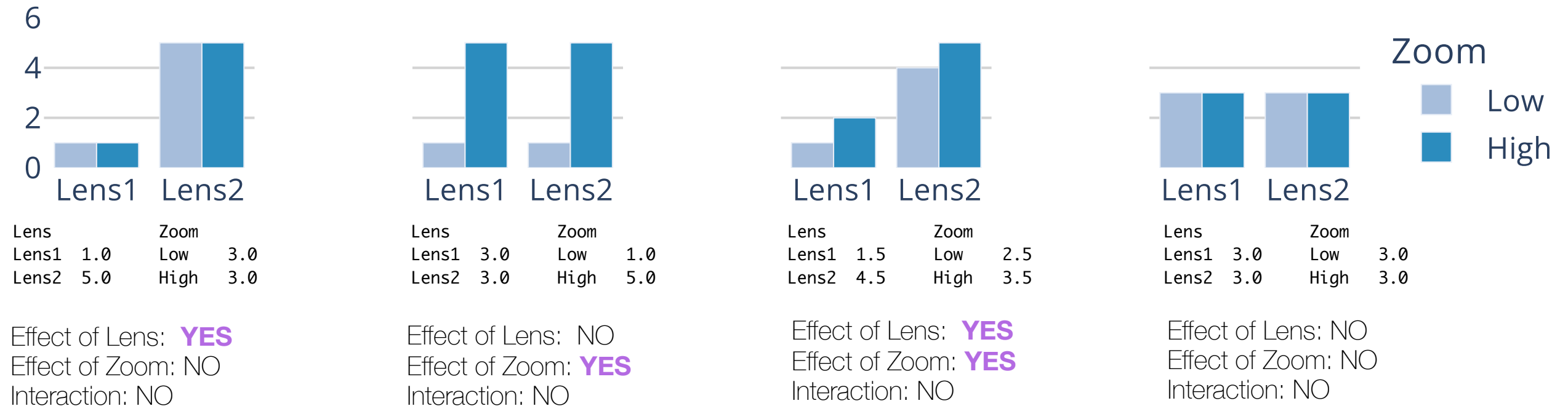
“It depends on the type of food!”

*interaction effect between food and condiment*



from <http://statisticsbyjim.com/regression/interaction-effects/>

# Simple and Interaction effects: many possible situations



# Two-way anova-test with pingouin

We use function `rm_anova` from library `pingouin`

```
aovrm2way = pg.rm_anova(data=data, dv='PointingTime', within=['Lens', 'Magnification'], subject='Participant')
aovrm2way
```

	Source	SS	ddof1	ddof2	MS	F	p-unc	p-GG-corr	ng2	eps
0	Lens	7.094755e+07	4	36	1.773689e+07	62.408495	1.077454e-15	1.464553e-06	0.685343	0.337259
1	Magnification	3.302858e+08	4	36	8.257145e+07	666.205094	3.309458e-33	6.874744e-14	0.910230	0.381525
2	Lens * Magnification	9.525647e+07	16	144	5.953529e+06	88.046197	6.510501e-66	6.160824e-09	0.745180	0.108297

We get a table with both simple and interaction effects. One line per effect where we can read all relevant information (degrees of freedom, p-value,  $F$ , effect size  $\eta^2$ )

## Final report:

An ANOVA test revealed a significant effect of Lens on PointingTime ( $F(4,36) = 62, p < 0.001, \eta^2=0.68$ ), a significant effect of Magnification on PointingTime ( $F(4,36) = 666, p < 0.001, \eta^2=0.91$ ), as well as a significant Lens x Magnification interaction effect ( $F(16,144) = 88, p < 0.001, \eta^2=0.74$ ).

+ post-hoc tests (see notebook) and charts (cf. next slides)



# Visualizing descriptive stats for the groups that we compare

```
import plotly.express as px
```

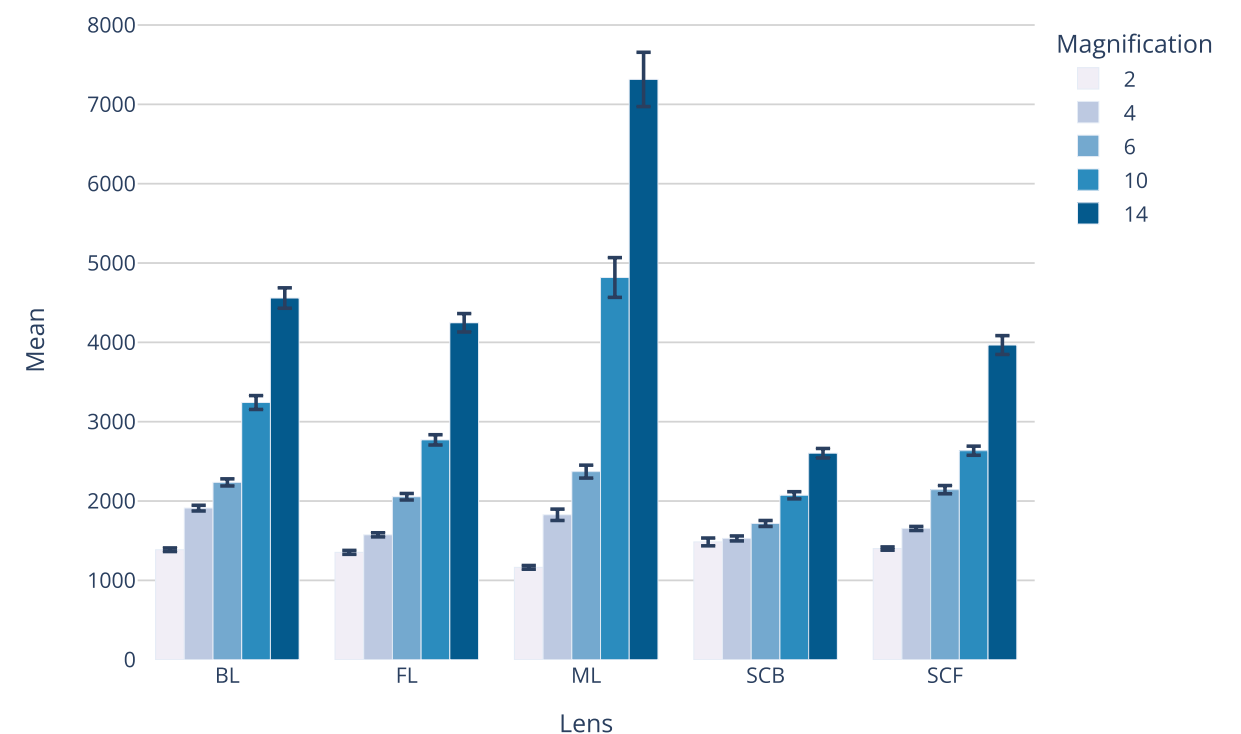
We use function `bar` from library `plotly.express` to produce a bar chart depicting the different groups

```
nice_color_palette = ['#f1eef6', '#bdc9e1', '#74a9cf', '#2b8cbe', '#045a8d']
stats['Magnification'] = stats['Magnification'].astype('str')

fig = px.bar(stats, x='Lens', y='Mean', error_y="error_y_hi", error_y_minus="error_y_lo", color='Magnification', barmode='group',
color_discrete_sequence=nice_color_palette)

fig.update_layout({
    'plot_bgcolor' : 'rgba(0,0,0,0)'
})
fig.update_yaxes(showgrid=True, gridwidth=1, gridcolor='LightGray')

fig.show()
```



# Inferential statistics

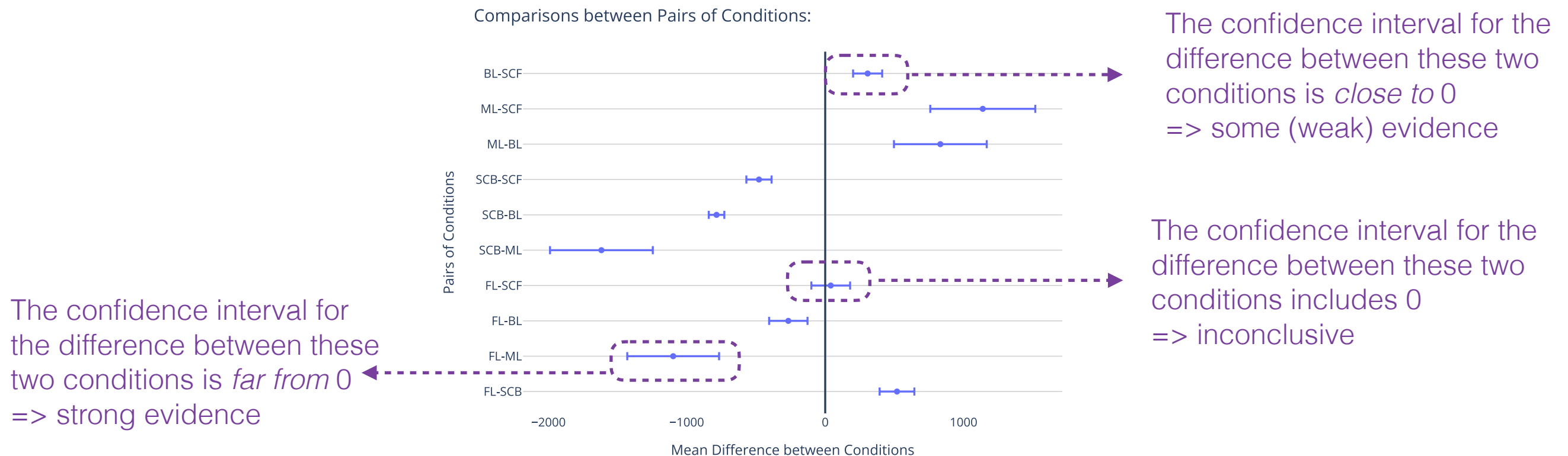
An alternative to null-hypothesis reasoning: confidence intervals

# The null hypothesis reasoning: critics

Dichotomy regarding how to accept or reject the null hypothesis on an arbitrary cut-off can be criticized (and actually is criticized)

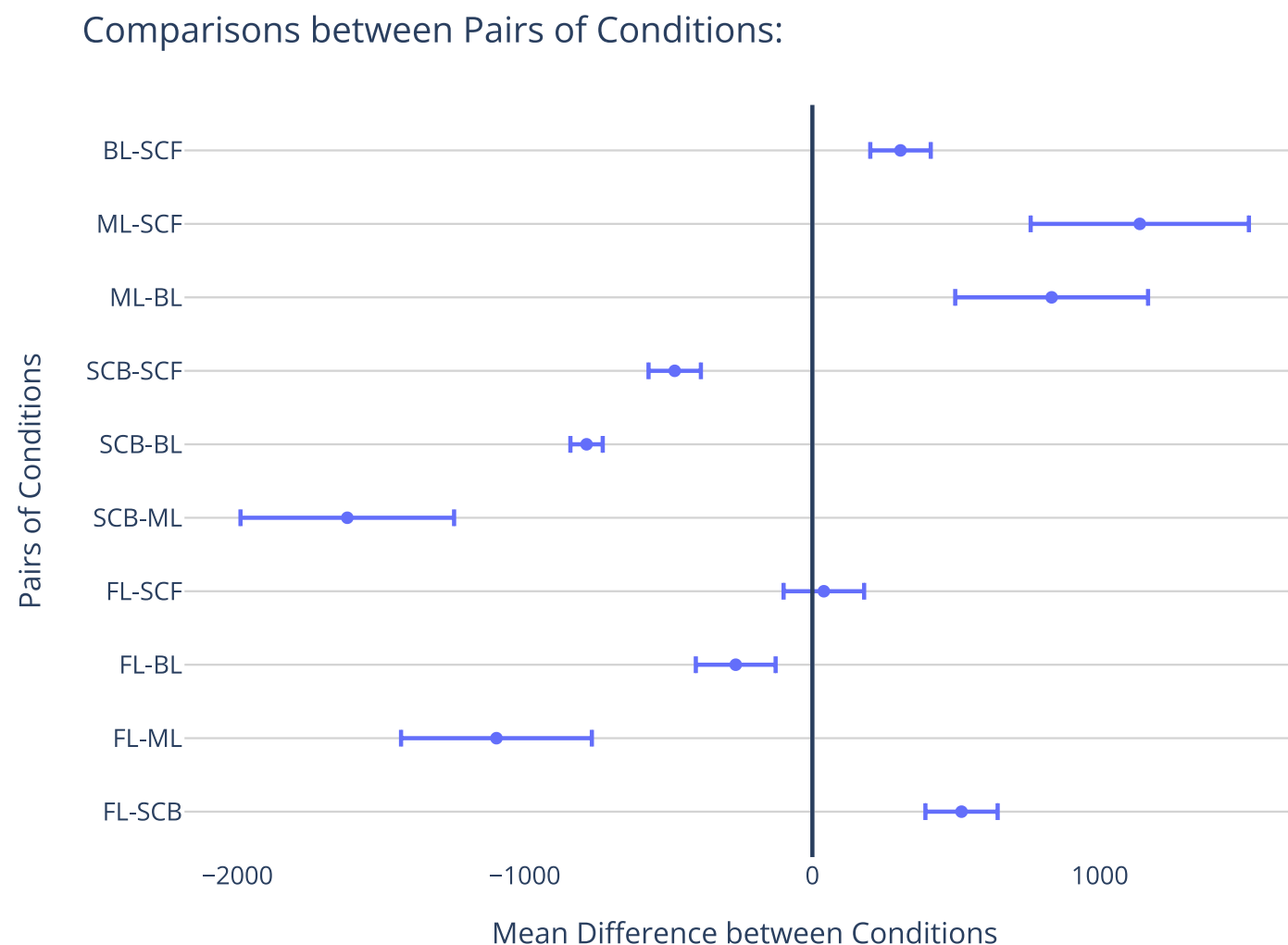
$p = 0.051 \Rightarrow$  reject,  $p = 0.049 \Rightarrow$  accept, mmm...

Some analysts prefer to focus on confidence intervals regarding the difference in means to look at their data and make more nuanced conclusions.



# Confidence Intervals of the Mean Difference between two conditions

See section **Confidence Intervals of the Mean Difference between two conditions** in notebook to get details about how to produce such a chart



# Inferential statistics

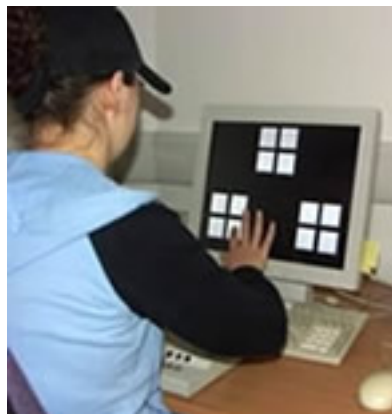
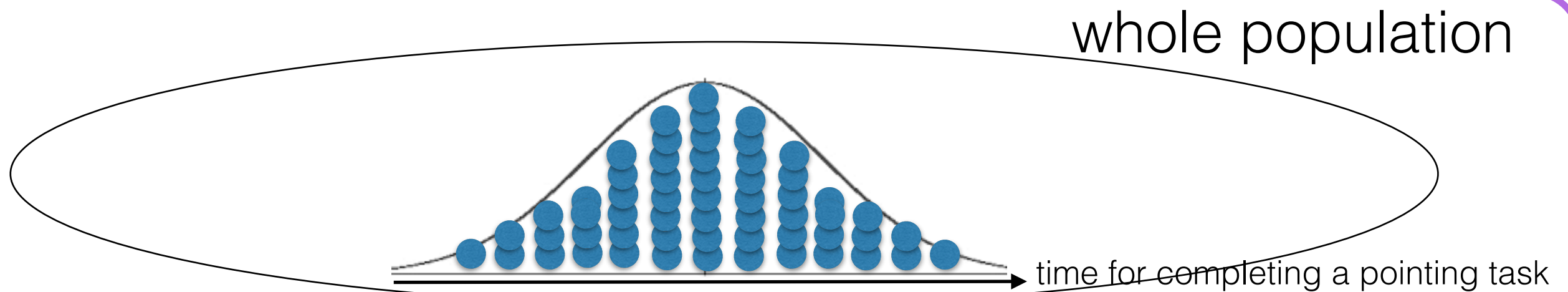
Non-parametric tests

## Parametric and non-parametric statistics

	Descriptive describe, show or summarize a <u>data sample</u>	Inferential make generalization about the <u>whole population</u> based on one data sample
Parametric make <u>assumptions</u> about the <u>data</u> <u>distribution</u> for the whole population	Descriptive parametric statistics	Inferential parametric statistics
Non parametric <u>no assumption</u> about the <u>data</u> <u>distribution</u>	Descriptive non-parametric statistics	Inferential non-parametric statistics

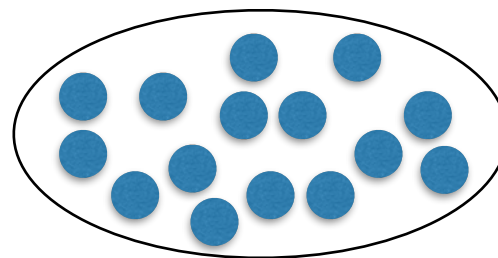
**assumption about the data  
distribution for the population**

# Experiment & parametric statistics



Experiment

data sample



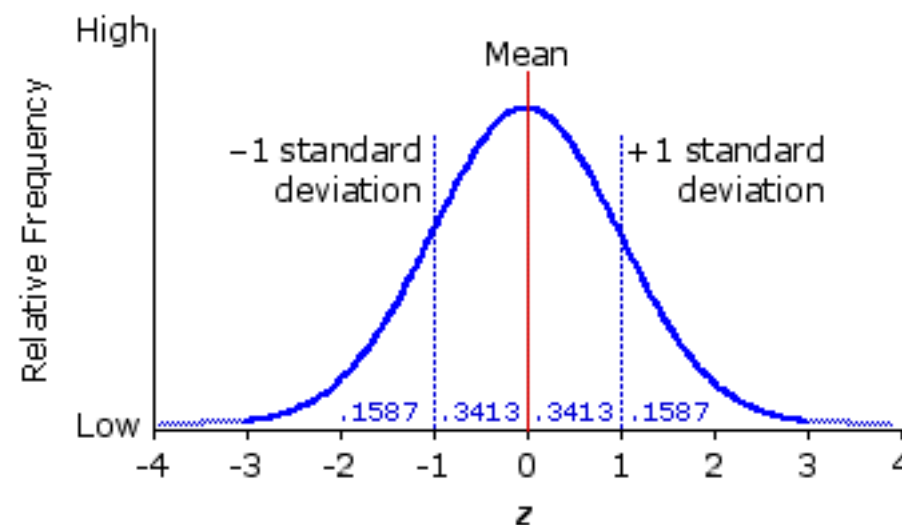
parametric statistics  
assumes that the data  
sample comes from **a  
population that follows a  
probability distribution**  
based on a fixed set of  
parameters  
(for example, a normal  
distribution)

# Parametric tests and the normal distribution

t-test, ANOVA test, Post-hoc tests

They all test the effect of nominal factors on a continuous measure. The assumption is that the continuous measure follows a normal distribution

Normal distribution: The "bell curve" (Mean = Median = Mode)



This graphics uses the standard deviation to scale the distribution: one unit on the x-axis is equal to one standard deviation.

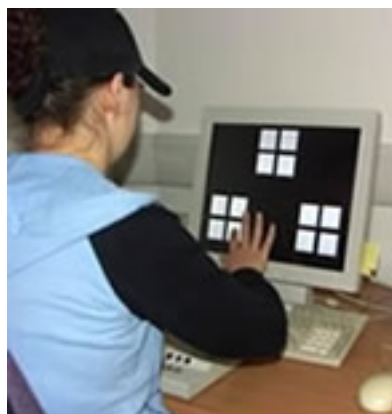
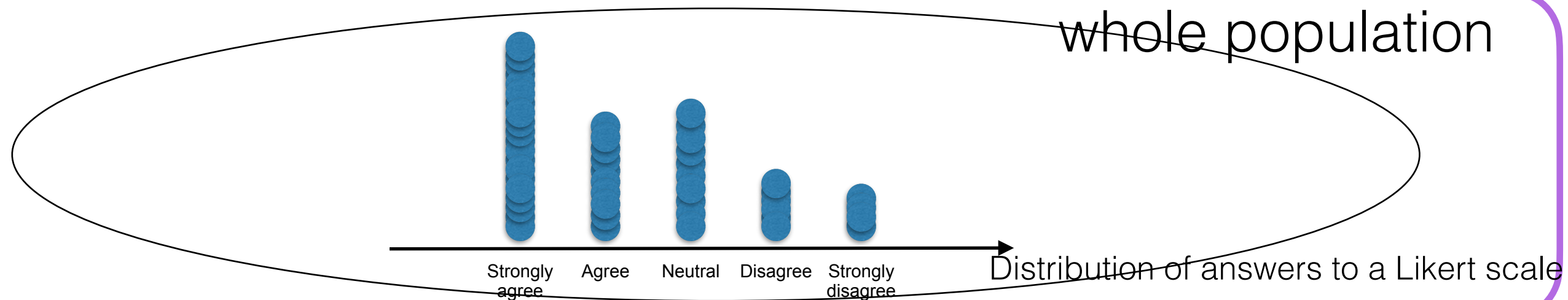


# Parametric and non-parametric statistics

	<b>Descriptive</b> describe, show or summarize a <u>data sample</u>	<b>Inferential</b> make generalization about the <u>whole population</u> based on one data sample
<b>Parametric</b> make <u>assumption</u> about the <u>data</u> <u>distribution</u> for the whole population	Descriptive parametric statistics	Inferential parametric statistics
<b>Non parametric</b> <u>no assumption</u> about the <u>data</u> <u>distribution</u>	Descriptive non-parametric statistics	Inferential non-parametric statistics

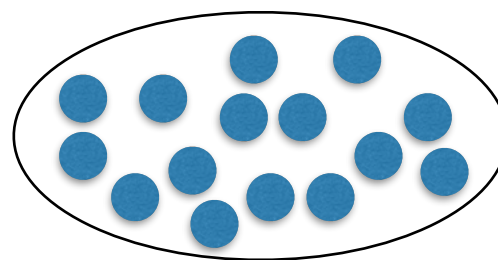
**no assumption about the data distribution for the population**

# Experiment & non-parametric statistics



Experiment

data sample



we assume nothing about the population

# Non-parametric: when?

In many cases, you can assume the normality for the population and use parametric tests

When should you go for non-parametric then?

Case 1: If the type of measure is not ratio (continuous),

Example: results from Likert-scale questions

Case 2: An assumption of a parametric test is violated

largely unbalanced data (number of observations largely varies among groups), non normal distribution, ...

*Note that many parametric tests remain fairly robust against the non-normality assumption so Case 1 is the most common.*

# Chi-square test

## When should we use a Chi-square test?

When comparing two groups with regard to a categorical (nominal measure)

Example: We want to test whether young and old people differ in the system they use. We have one factor age (i.e., two groups {young, old}) and the measure is the answer to question “which system do you use?”

We organize observations into a contingency table (count of answers per category)

	windows	mac	linux
young	16	11	3
old	21	8	1

# Chi-square test / McNemar's test

It computes the statistics  $\chi^2$  (and watches where it lies in the theoretical  $\chi^2$  distribution)

Effect size ( $\phi$  or Cramer's  $V$ ):

$$\phi = \sqrt{\frac{\chi^2}{N(k-1)}}$$

where  $N$  is the number of observation

and  $k$  is the smallest number of rows  $r$  or of columns  $c$

If groups are paired (i.e., the factor is tested according to a within-subject design), use a McNemar's test

# Friedman test

## When should we use a Friedman test?

When comparing more than two groups with regard to an ordinal measure

This is a case that typically occurs when we collect qualitative appreciations using Likert scales. For example, we want to test whether people find some lenses easier to use than others. We have one factor Lens (i.e., five groups for five types of lens that we had in our experiment) and we ask participants:

How much do you agree (from -2 to 2) with the following statement:

**This Lens is easy to use**

(-2: Strongly Disagree, -1: Disagree, 0: Neutral, 1: Agree, 2: Strongly Agree)

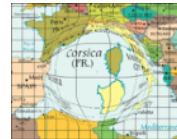
We want to compare five groups (the five lenses) with regard to their score on ease of use (ordinal measure).

# Friedman test with `pingouin`

Five different lenses:



**ML** (Manhattan Lens) ,



**FL** (Fisheye Lens) ,

**SCF** , **SCB** ,



**BL** (Blending Lens)

We believe that, because of their different design properties, users will give different scores for ease of use.

H: User scores differ depending on the type of Lens.


Null hypothesis: User scores are the same regardless of Lens.

We test one nominal factor (Lens) according to a within-subject design (paired groups) on an ordinal measure (score), we thus run a Friedman test using the `pingouin` library.

# Friedman test with pingouin

The Friedman test works with a **wide-format** dataframe where there is one column per factor level. Each line contains the measure value for each of the factor level. (In a **long-format** dataframe, there are columns for participant, factor(s), and measure and each line is a trial that contains the value for all these columns).

```
data_likert_full = pd.read_csv('lens_experiment/easy_scores.csv', sep=';')
data_likert_full
```



	Participant	Easy_ML	Easy_FL	Easy_BL	Easy_SCF	Easy_SCB	Unnamed: 6
0	1	-2	0	-1	1	0	NaN
1	2	-1	-2	0	1	1	NaN
2	3	-2	-1	0	0	1	NaN
3	4	-1	0	-1	1	2	NaN
4	5	-2	-2	-1	1	0	NaN
5	6	-2	0	0	1	2	NaN
6	7	-1	-2	-2	0	0	NaN
7	8	-2	0	0	1	2	NaN
8	9	-1	0	0	2	1	NaN
9	10	-1	-1	-1	1	2	NaN

Factor Lens has 5 levels:  
{ML, FL, BL, SCF, SCB}



```
import pingouin as pg
```

# Friedman test with pingouin

```
data_likert = data_likert_full.drop('Unnamed: 6',axis = 1)
data_likert = data_likert.drop('Participant',axis = 1)
data_likert = data_likert.rename(columns={'Easy_ML': 'ML', 'Easy_FL': 'FL', 'Easy_BL': 'BL', 'Easy_SCF': 'SCF', 'Easy_SCB': 'SCB'})
data_likert
```

	ML	FL	BL	SCF	SCB
0	-2	0	-1	1	0
1	-1	-2	0	1	1
2	-2	-1	0	0	1
3	-1	0	-1	1	2
4	-2	-2	-1	1	0
5	-2	0	0	1	2
6	-1	-2	-2	0	0
7	-2	0	0	1	2
8	-1	0	0	2	1
9	-1	-1	-1	1	2

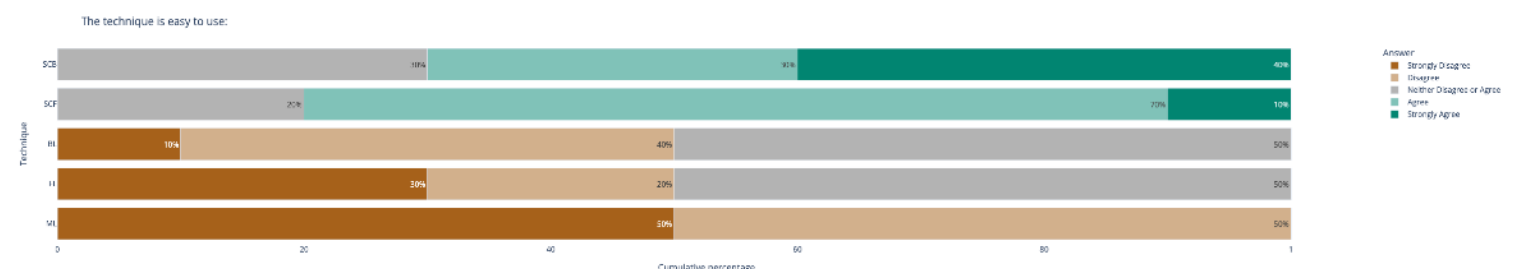
```
pg.friedman(data_likert)
```

	Source	W	ddof1	Q	p-unc
Friedman	Within	0.803763	4	32.150538	0.000002

## Final report:

We found a significant effect of factor Lens on Easy scores ( $Q(4) = 32, p < 0.001$ ).

- + post-hoc tests
- + chart (see notebook)



# Inferential statistics

Choosing the right test

# Which test when?

This class does not cover all possible statistical tests

In order to choose the right test, consider:

- the experiment design:

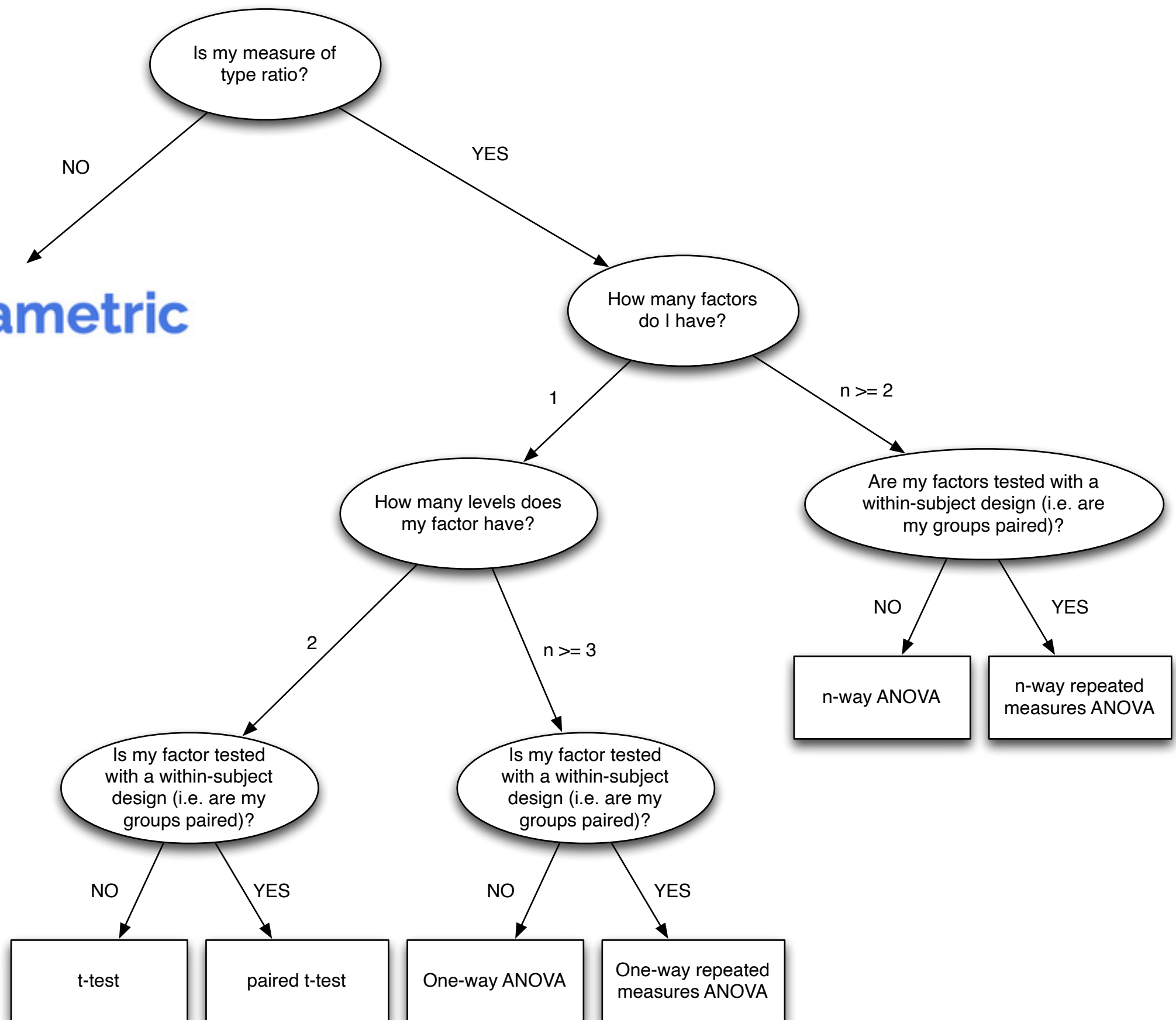
  - within-subject / between-subject

- the type of your variables:

  - number and types of independent variables (factors) and  
type of dependent variable (measure)

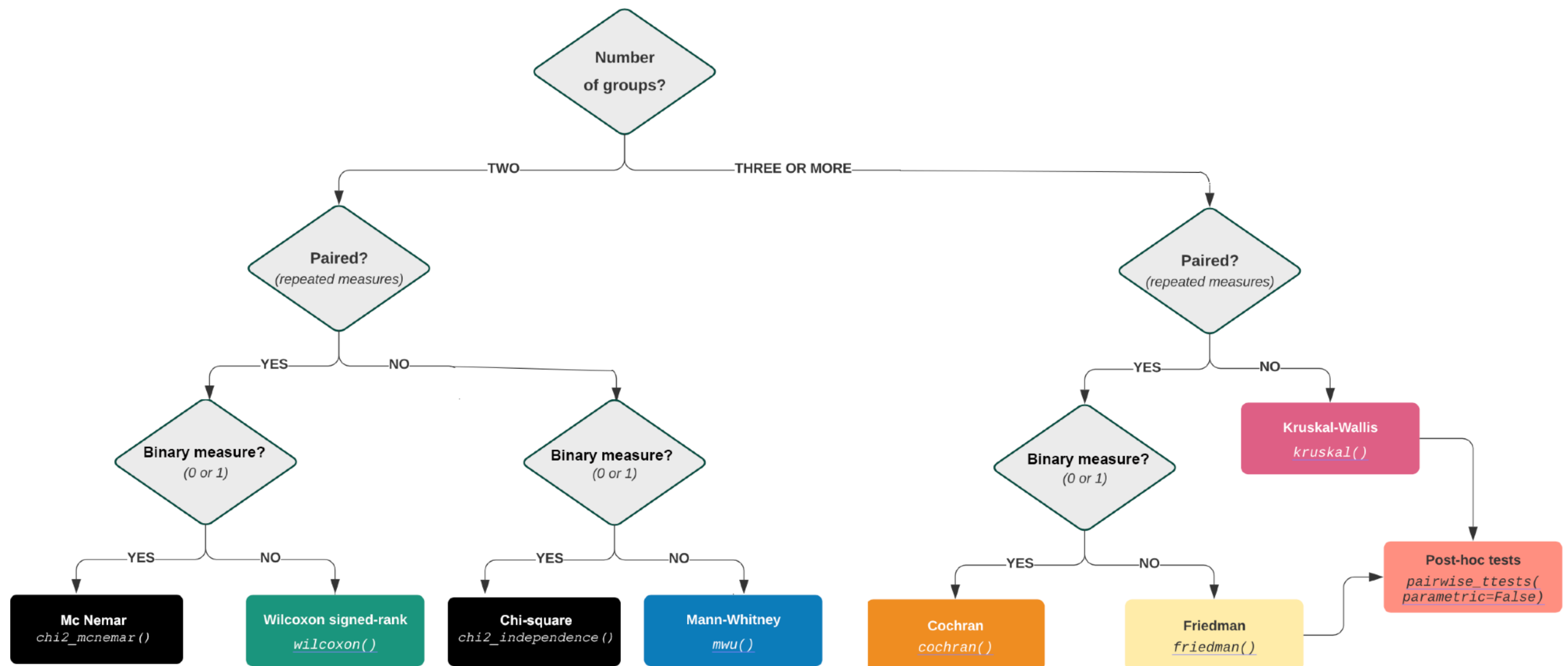
# A decision tree to help choose the right statistical test

**Non-parametric**



# A decision tree to help choose the right statistical test

## Non-parametric



Adapted from:

[https://www.fabriziomusacchio.com/teaching/python\\_course\\_neuropractical/01\\_statistical\\_data\\_analysis\\_with\\_pandas\\_and\\_pingouin](https://www.fabriziomusacchio.com/teaching/python_course_neuropractical/01_statistical_data_analysis_with_pandas_and_pingouin)