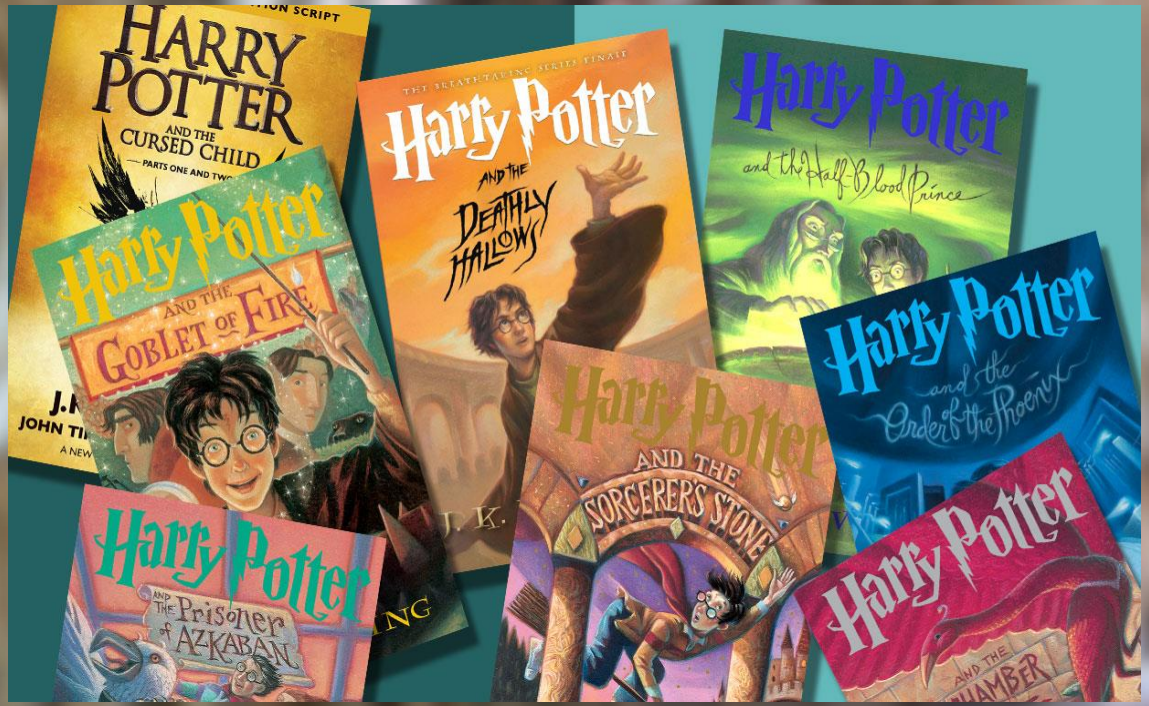
Hypothesis testing

Rita Skeeter

Ollivander

Rumor

Wizards have longer wands than witches


HARRY POTTER
AND THE CURSED CHILD
PARTS ONE AND TWO
Harry Potter
AND THE GOBLET OF FIRE
Harry Potter
AND THE PRISONER OF AZKABAN
THE DEATH-TAKING SERIES FINALE
Harry Potter
AND THE DEATHLY HALLOWS
J. K.
Harry Potter
and the Half-Blood Prince
Harry Potter
AND THE SORCERER'S STONE
Harry Potter
and the Order of the Phoenix
Harry Potter
AND THE

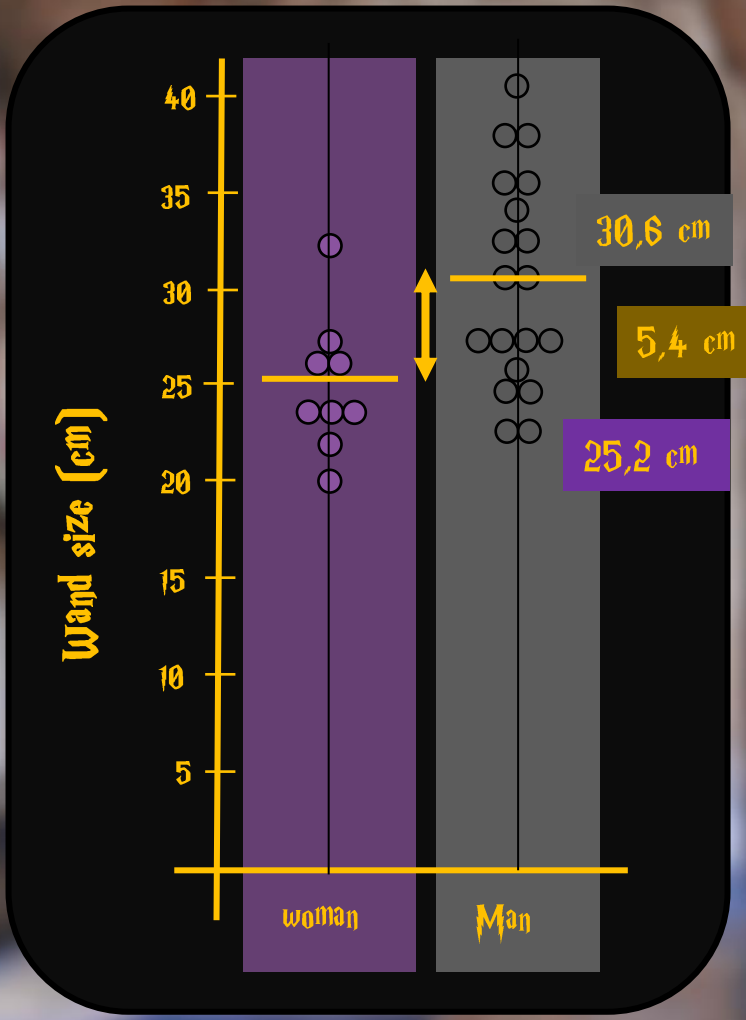
characters

Dataset

characters	Gender	Wand size (cm)
Harry Potter	Homme	27,94
Ronald Weasley	Homme	35,84
Hermione Granger	Femme	27,305
Albus Dumbledore	Homme	38,1
Rubeus Hagrid	Homme	40,64
Neville Longbottom	Homme	33,02
Fred Weasley	Homme	35,56
Lily Potter	Femme	26,035
James Potter	Homme	27,94
Remus Lupin	Homme	27,305
Peter Pettigrew	Homme	24,765
Charles Weasley	Homme	30,48
Minerva McGonagall	Femme	23,495
Draco Malfoy	Homme	25,4
Bellatrix Lestrange	Femme	32,385
Dolores Umbridge	Femme	20,32
Horace Slughorn	Homme	27,305
Tom Riddle	Homme	34,29
Celestina Warbeck	Femme	26,67
...	...	...

How to find out?

Look at the difference in averages



Question :

Real difference? Or a difference due to chance?

True on a given sample

Statistical test

Extrapolate to total population

Statistical test

Hypothesis: Wands come in different sizes

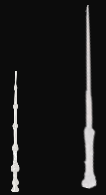
Hypothesis testing

H0 : women = men

Differences due to chance

H1 : women ≠ men

Differences not due to chance



Difference between the two hypotheses

Hypothesis testing

H0 Null Hypothesis

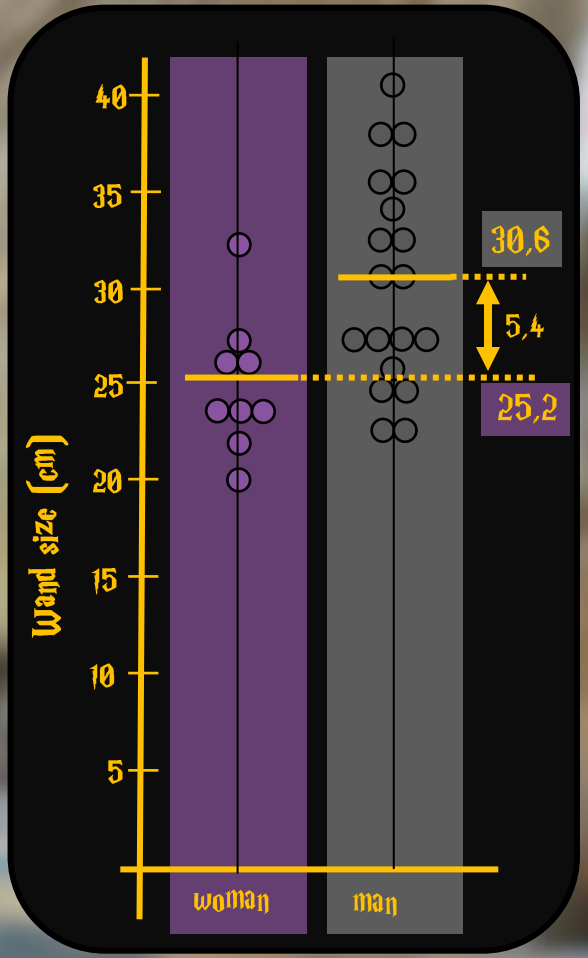
No difference in the population

The parameters (mean, variance, distributions) are the same.
This doesn't mean that there are no differences at all (sampling fluctuations).

H1 alternative hypothesis

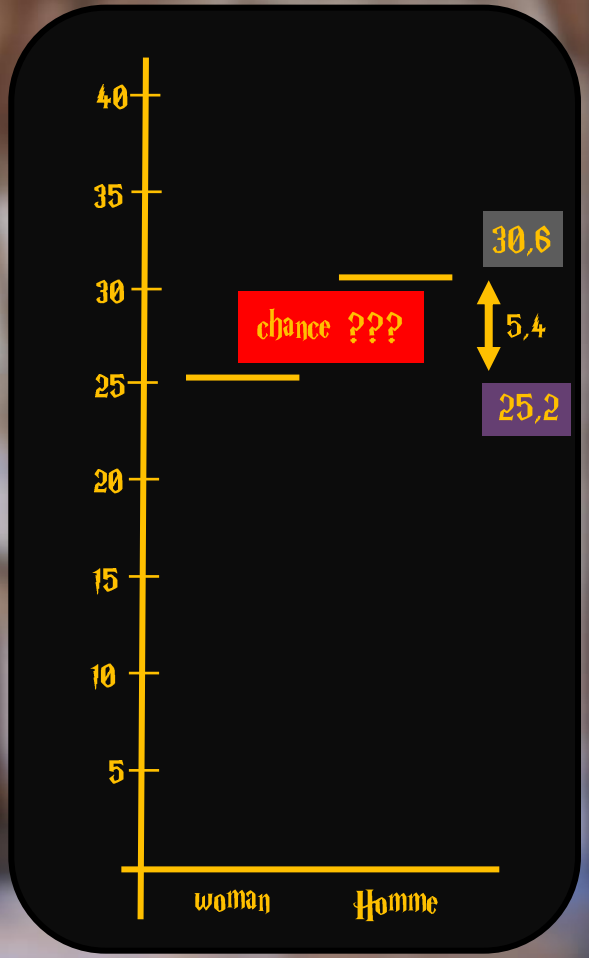
Difference in the population

The parameters (mean, variance, distributions) are different.
The differences observed are the consequence of two different populations.

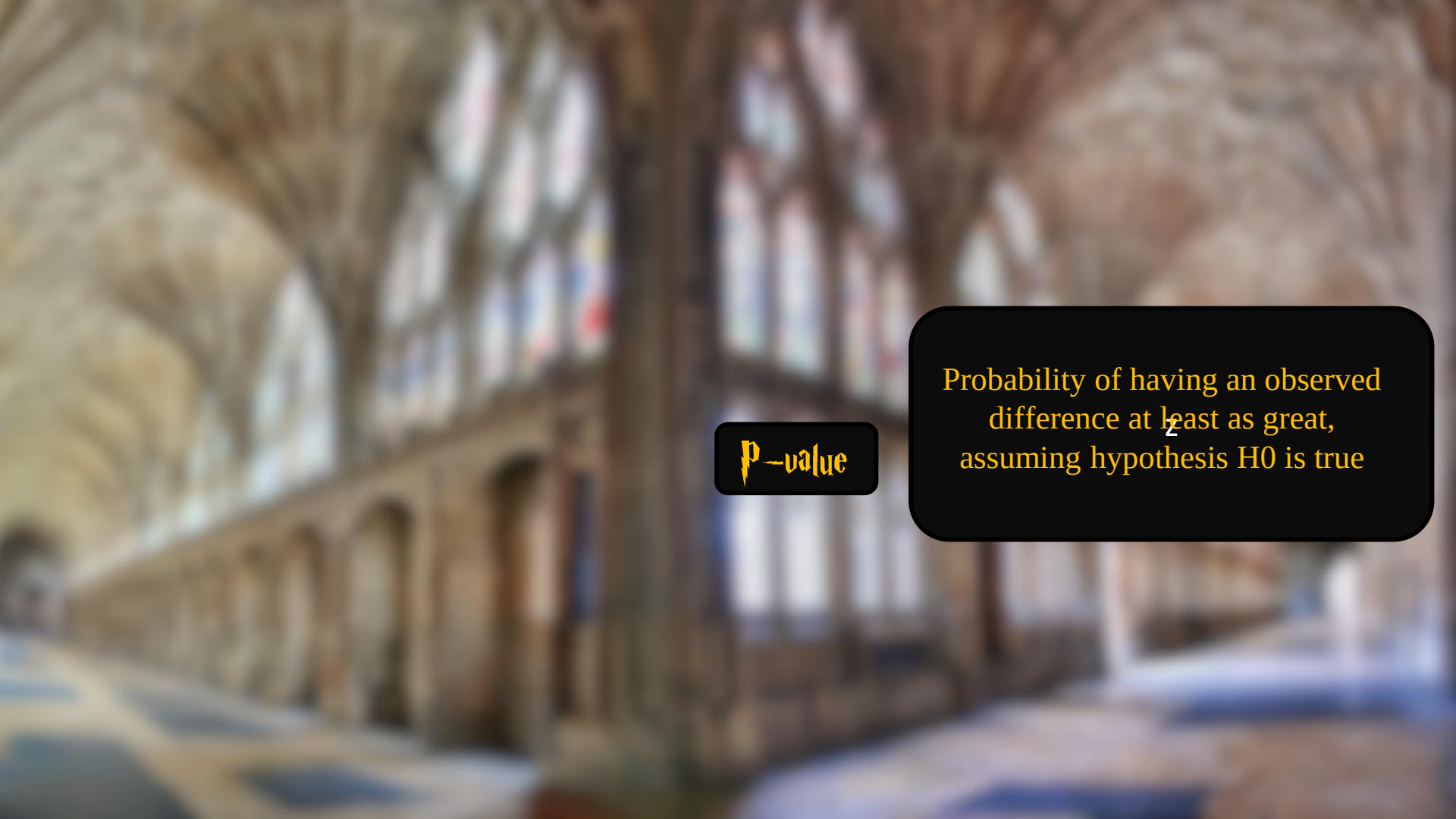


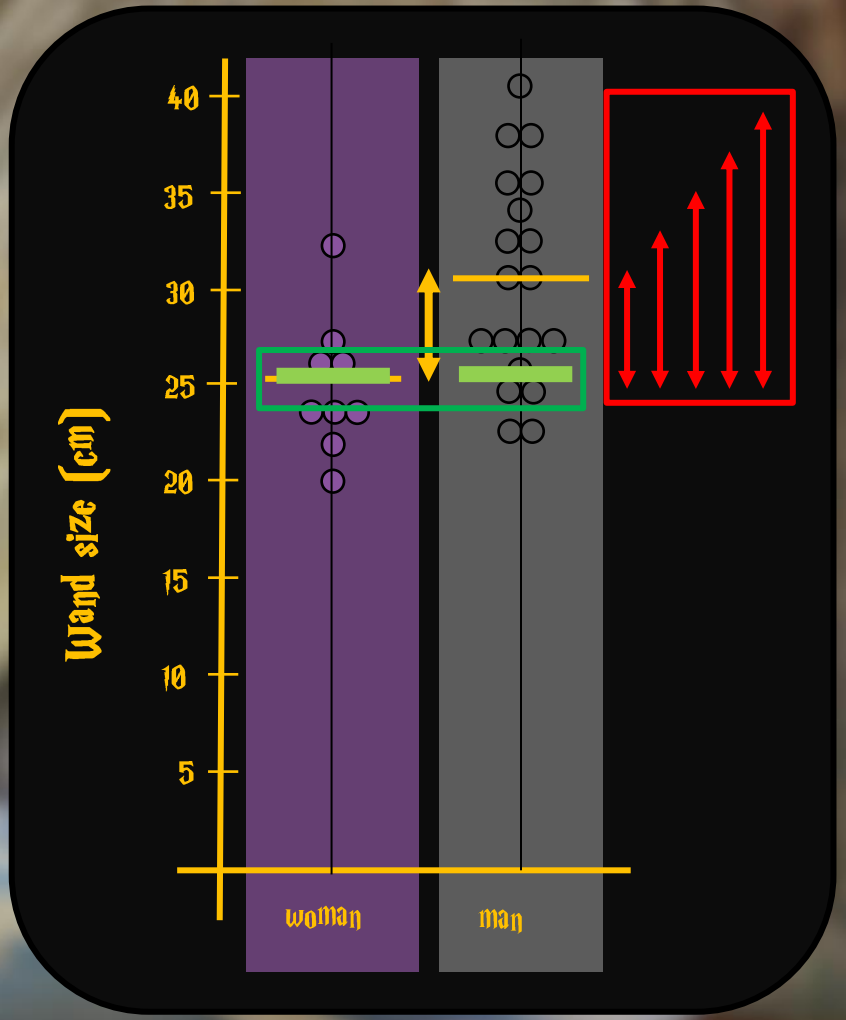
$\Delta = 5.4$

Is 5.4 far enough from 0 to reject H0 ?



We need a judge !


P-value
Probability of having an observed difference at least as great, assuming hypothesis H0 is true



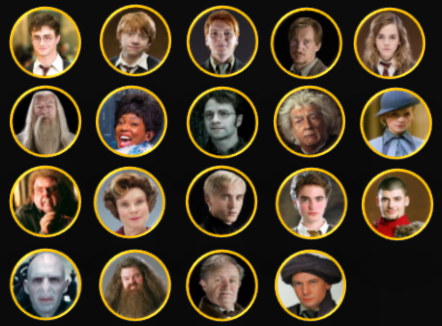
P-value

Probability of having

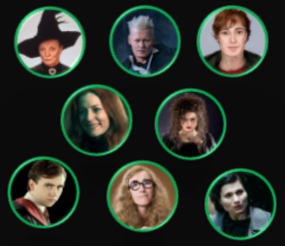
Probability of having an observed difference at least as great,

assuming hypothesis H0 is true

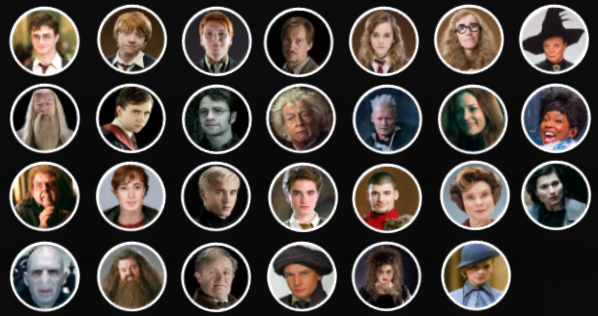
Yellow mean = 28,7 cm

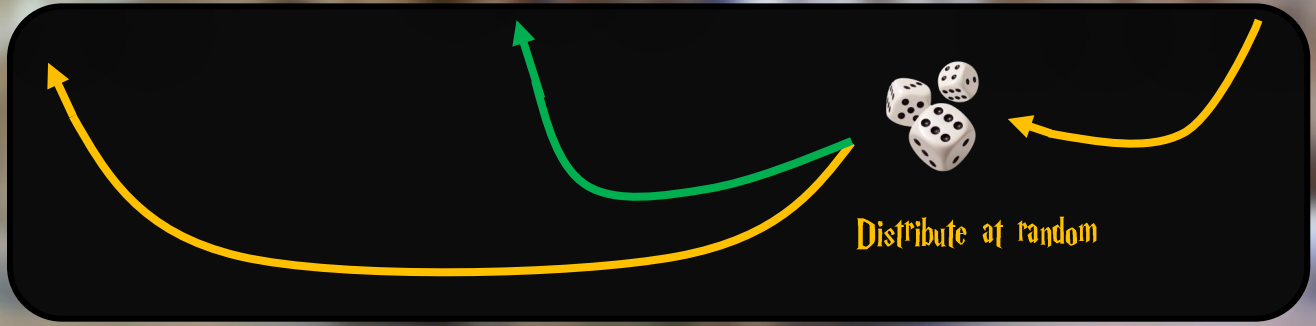


Green mean = 29,4 cm

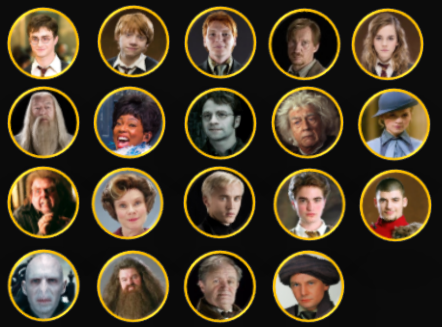


Dataset

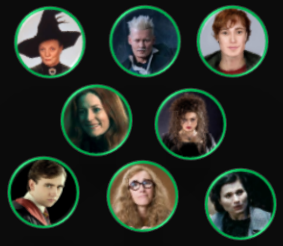


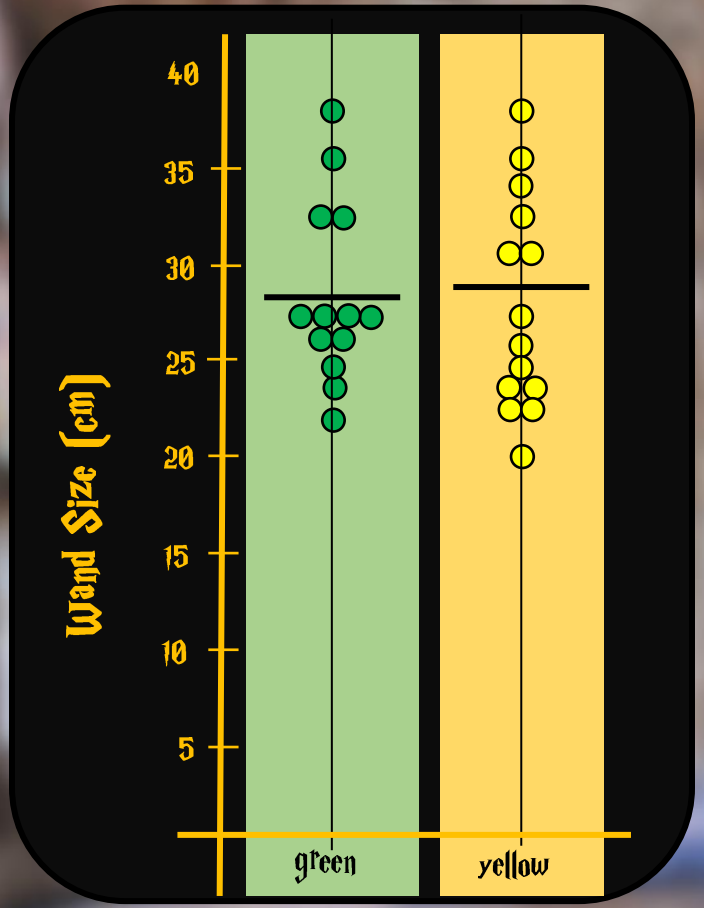


Yellow mean = 28,7 cm



Green mean = 29,4 cm



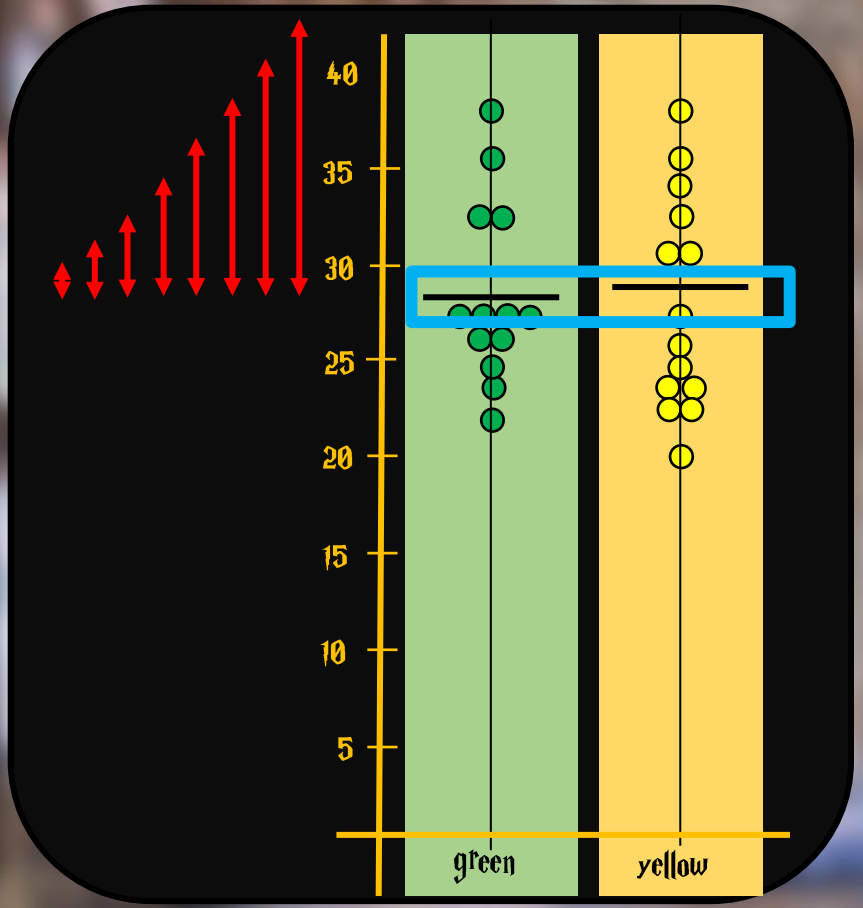


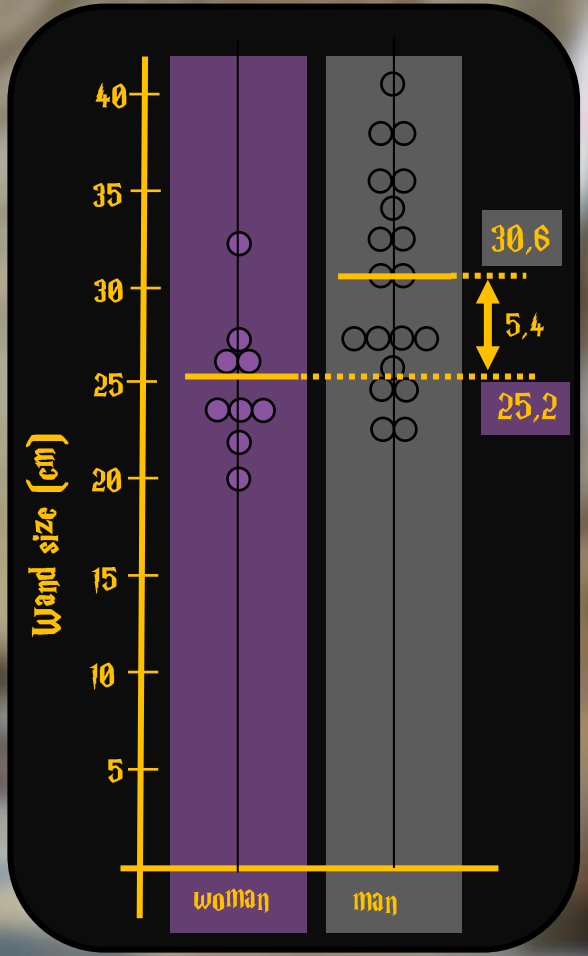
P-value

Probability of having

Probability of having an observed difference at least as great,

assuming hypothesis H0 is true



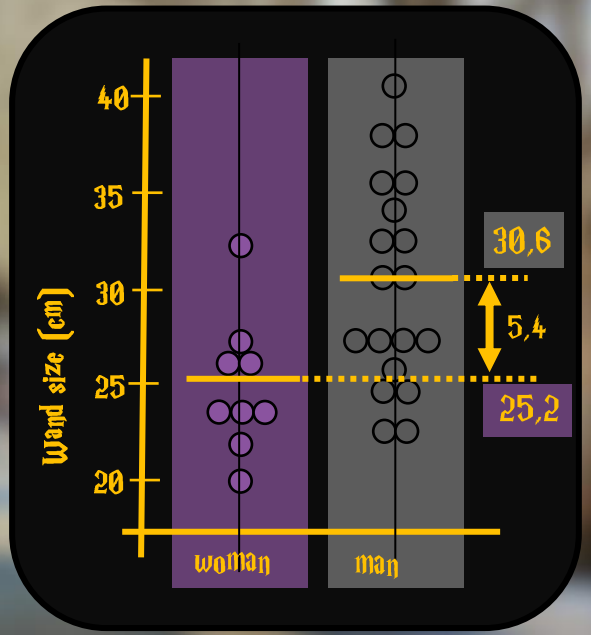


P-value ?

Get 5.4 cm or more

If H0 is true (the difference is due to chance)

P-value = 0,66



Difference of 5.4 cm or more

2 randomly assigned groups

Not exceptional → 66% random mixes

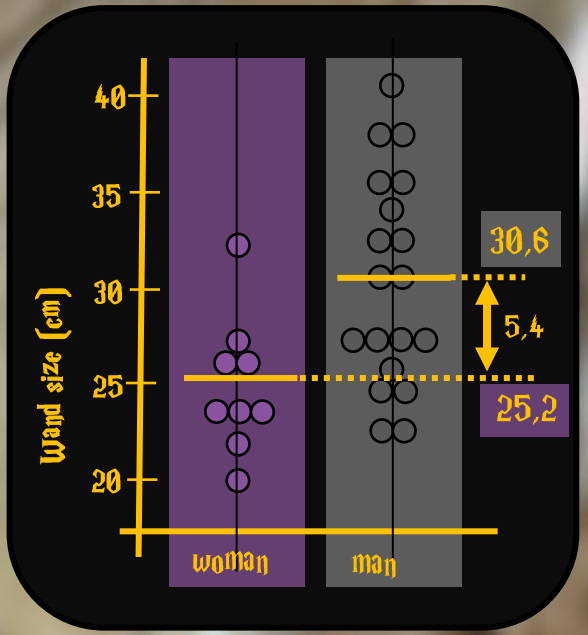
P-value = 0,66

H0 :

Differences due to chance

H1 :

Differences not due to chance



Difference of 5.4 cm or more

2 randomly assigned groups

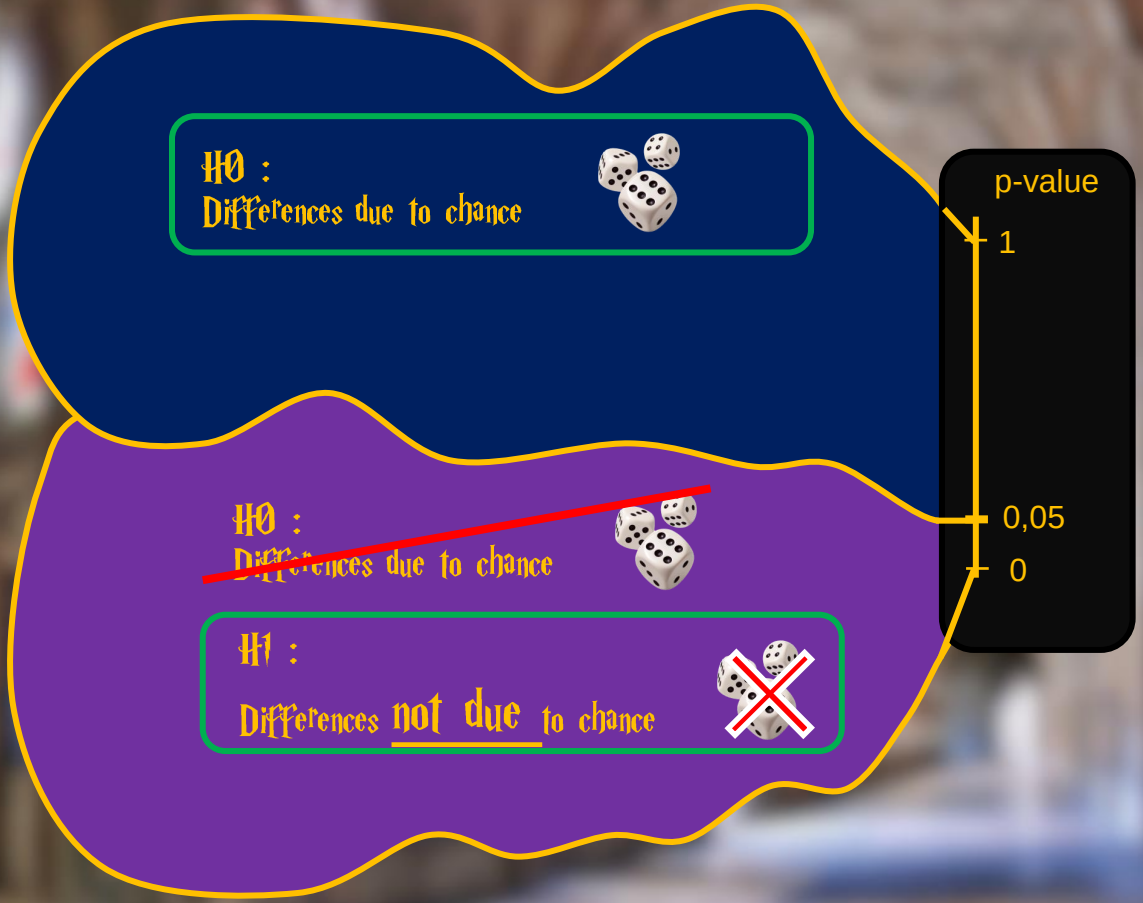
Really exceptional → 1 ‰ random mixes

P-value = 0,001

P-value = 0,001

H0 :
Differences due to chance

H1 :
Differences not due to chance


H0 :
Differences due to chance
H0 :
Differences due to chance
H1 :
Differences not due to chance
p-value
1
0,05
0

P-value definition

The rigorous definition: "Probability of obtaining an observed difference this large (or even larger) given H0 true."

The p-value: "the probability that the results are due to chance"?
Not really, it's a bit more subtle than that.

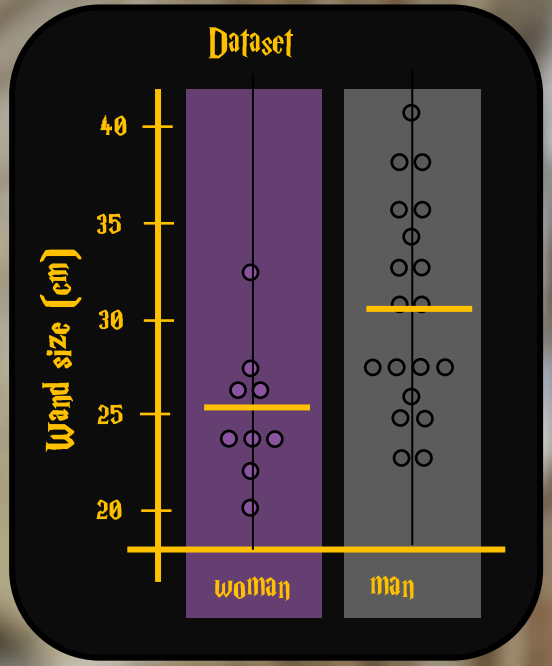
But does it really matter if we retain it? Not really, no.

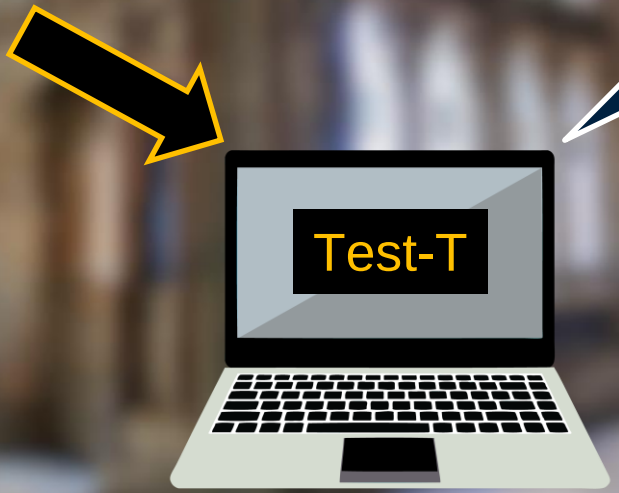
Threshold value for p-value

The usual threshold is set at $\alpha = 5\%$ (the maximum accepted). If $p < 5\%$ (0.05), we reject H0 and accept H1.

If $p \geq 5\%$ (0.05), we do not reject H0. We simply conclude that the observed difference is not significant.

WE DON'T SAY THERE IS NO DIFFERENCE !






```
> t.test(Taille_Baguette ~ Genre, data = HPW_MF)

        Welch Two Sample t-test

data:  Taille_Baguette by Genre
t = -3.2046, df = 23.25, p-value = 0.003899
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.949706 -1.930470
sample estimates:
mean in group Female   mean in group Male
            25.18833             30.62842
```

```
                Welch Two Sample t-test

data:  Wand_size by Gender
t = -2.4766, df = 12.385, p-value = 0.02859
alternative hypothesis: true difference in means between
group Female and group Male is not equal to 0
95 percent confidence interval:
 -10.1241806  -0.6650502
sample estimates:
mean in group Female   mean in group Male
            26.03500             31.42962
```

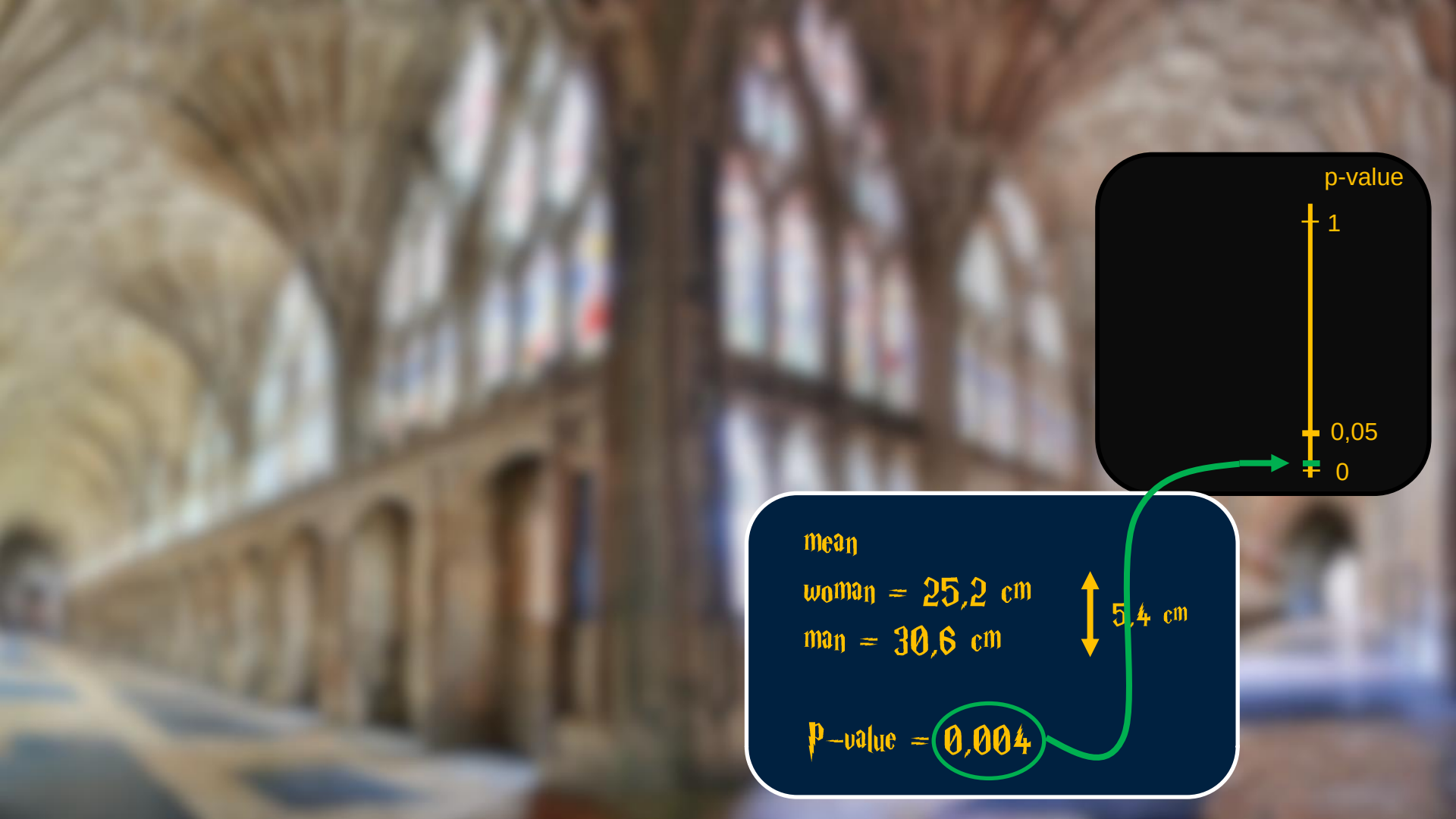
Means

woman = 25,2 cm

man = 30,6 cm

5,4 cm

P-value = 0,004


p-value
1
0,05
0
mean
woman = 25,2 cm
man = 30,6 cm
5,4 cm
p-value = 0,004

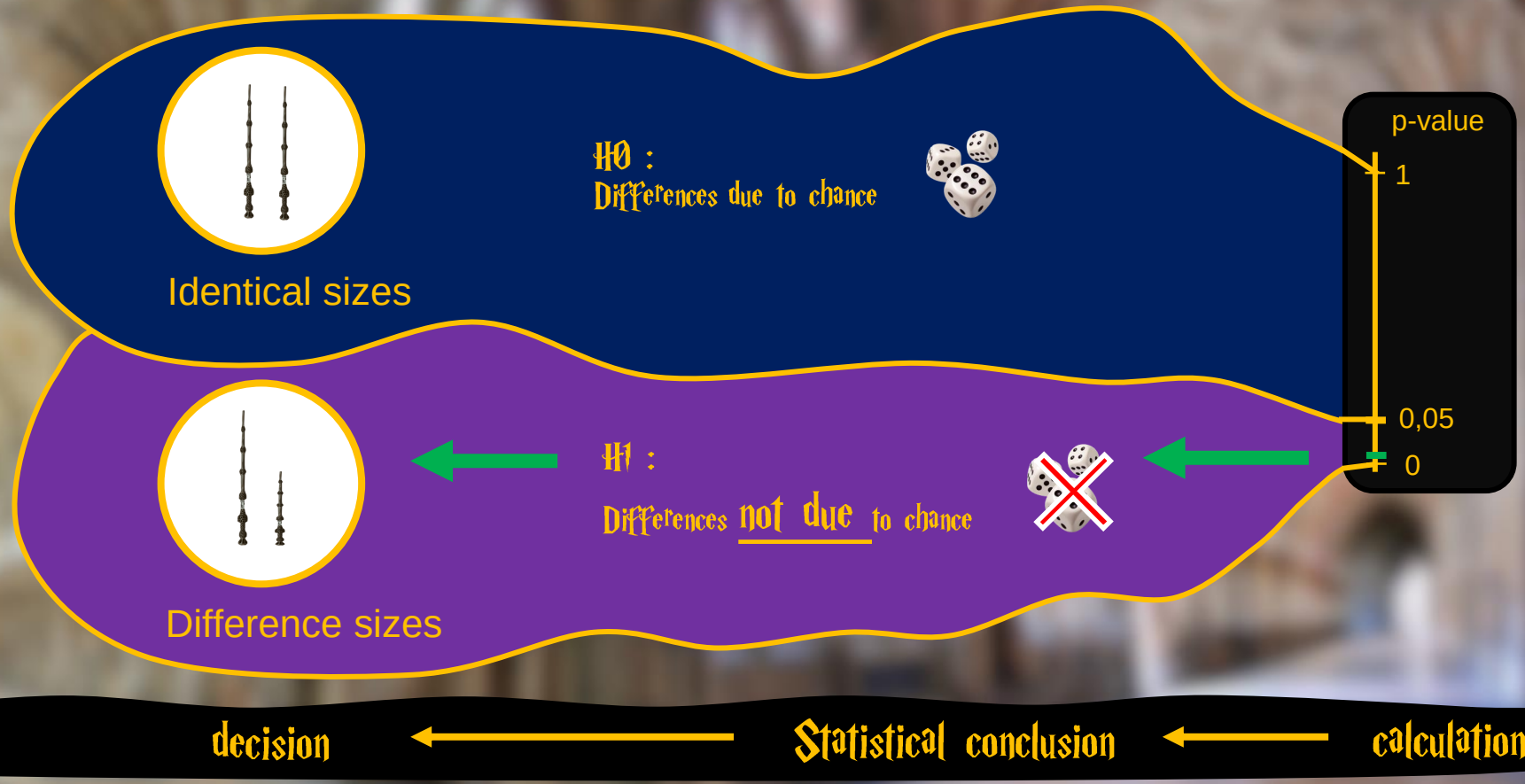
H0 :
Differences due to chance

H0 :
Différences dues au hasard

H1 :
Differences **not due** to chance

p-value
1
0,05
0

Statistical conclusion ← calculation


Identical sizes
H0 :
Differences due to chance
p-value
1
0,05
0
H1 :
Differences not due to chance
Difference sizes
decision
Statistical conclusion
calculation

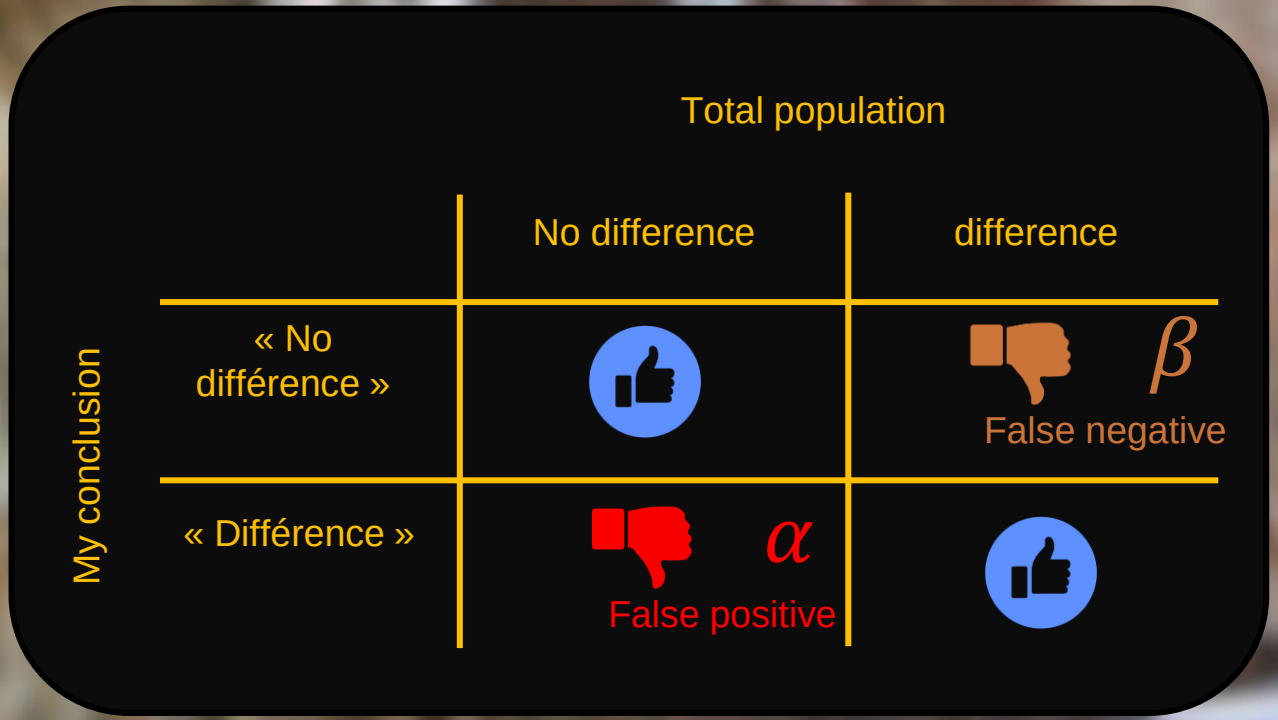
```
> t.test(Taille_Baguette ~ Genre, data = HPW_MF)

        Welch Two Sample t-test

data:  Taille_Baguette by Genre
t = -3.2046, df = 23.25, p-value = 0.003899
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.949706 -1.930470
sample estimates:
mean in group Female   mean in group Male
            25.18833             30.62842
```

Wizards and witches have significantly different wand sizes (means: 30.6 vs 25.2, t = -3.20, df=23.3,p = 0.004).

There is a significant effect of gender on wizard wand size (t = -3.20, df=23.3,p = 0.004).

Watch out for causality !

There's always the risk of making a mistake.


Total population
No difference
difference
My conclusion
« No différence »
« Différence »
β
False negative
α
False positive

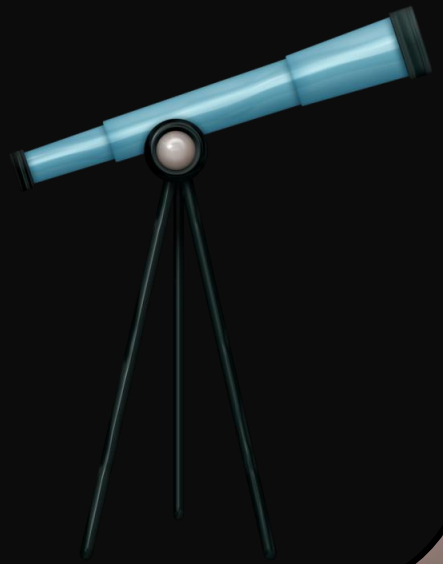
Type 1 error α (false positive)

Type 2 error β (false negative)

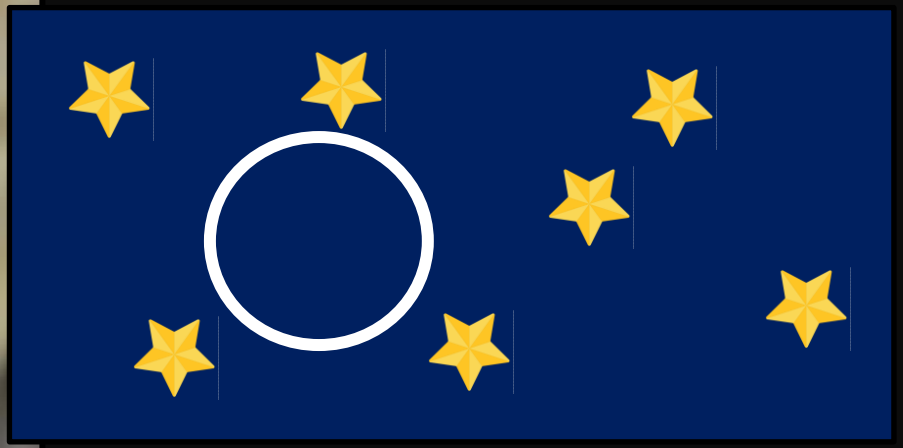
Power of a test $(1-\beta)$: let's go to the Astronomy Tower

The power of a test is the ability to reject H0 knowing that H1 is true, in other words, the ability not to miss a significant effect.

The power of a test is a bit like the resolution of a telescope.



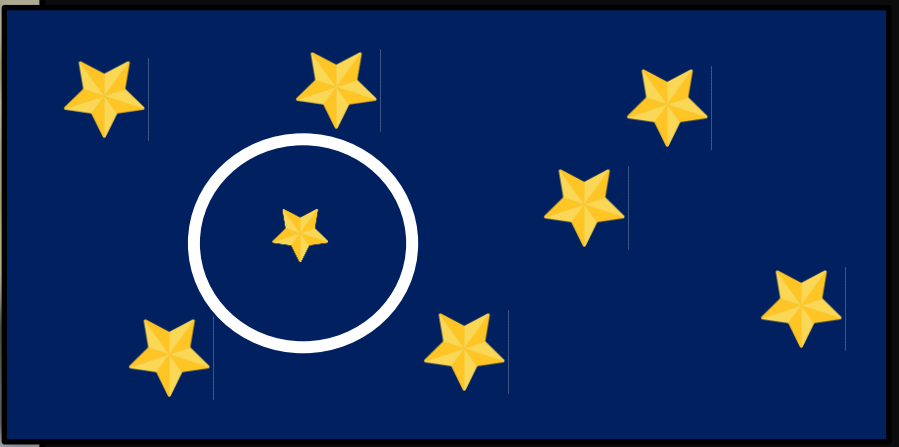
Power of a test: let's go to the Astronomy Tower



If the power of your test is low, you risk missing an effect.

The telescope's resolution is low, so you can't see any stars in the circled area.

Power of a test: let's go to the Astronomy Tower



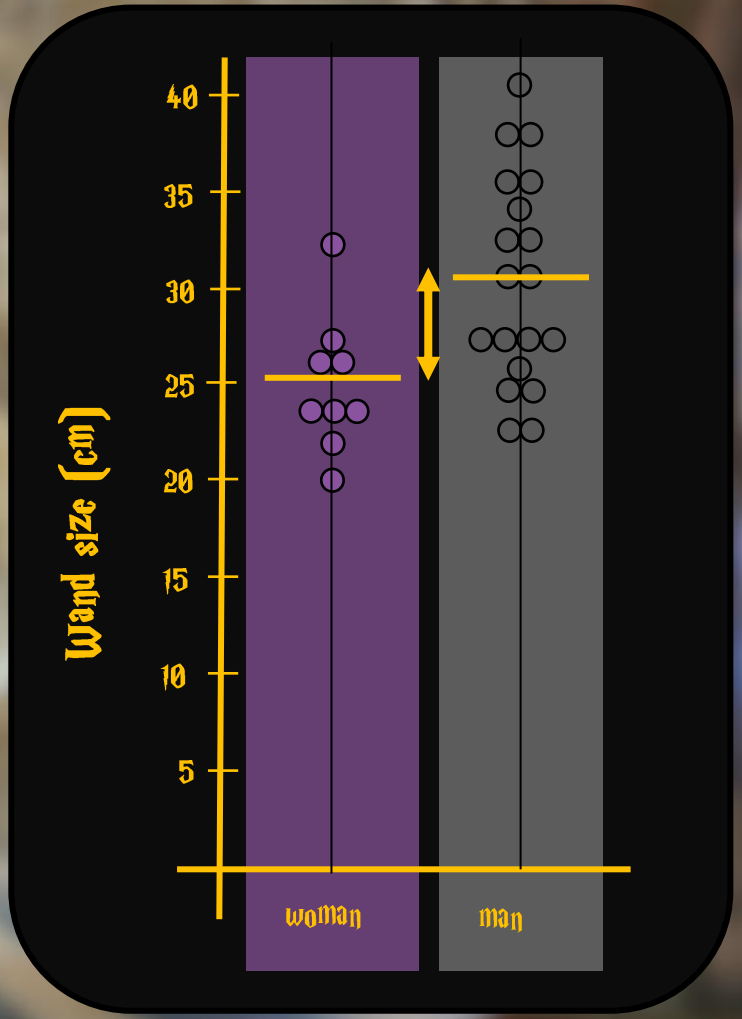
If the power of your test is high, you'll see more detail, more subtlety, more significance.

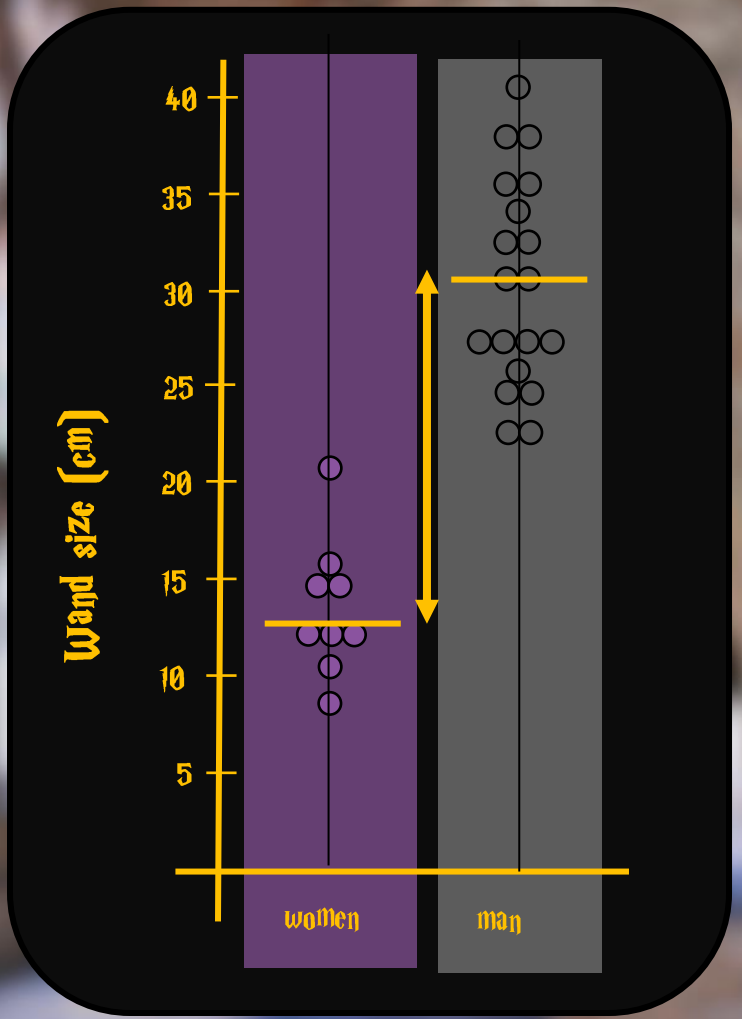
If the telescope's resolution is higher, you'll notice the small star.

Power of a test: important ?

The idea is to avoid wasting time and resources if there's little chance of showing anything.

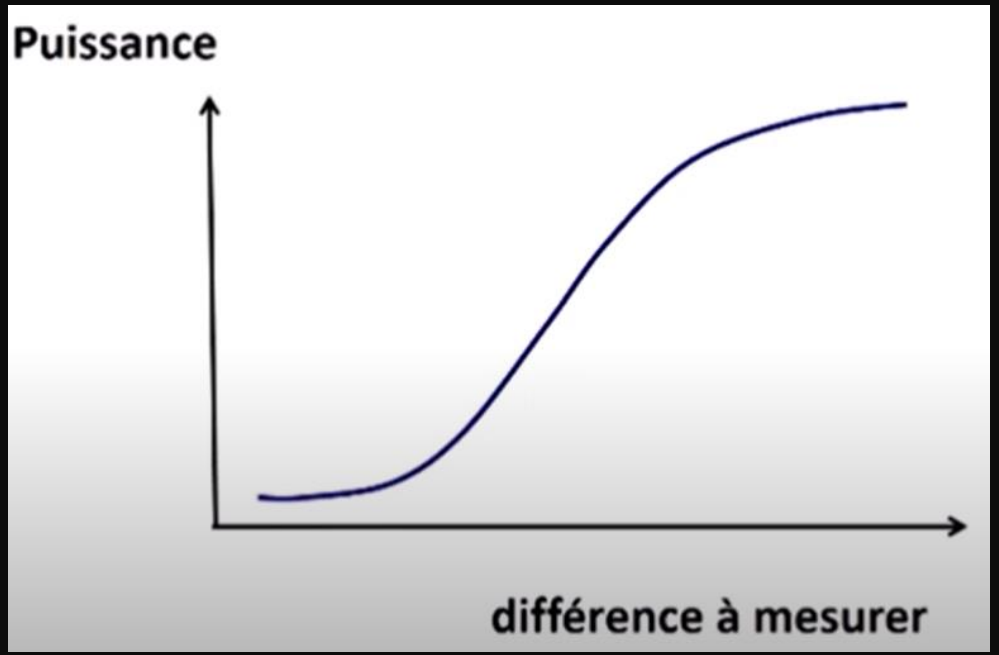
What does power depend on?


Wand size (cm)
40
35
30
25
20
15
10
5
woman
man


Wand size (cm)
40
35
30
25
20
15
10
5
women
man

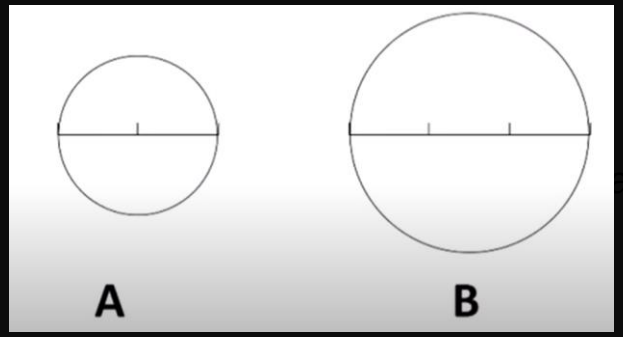
What does power depend on?

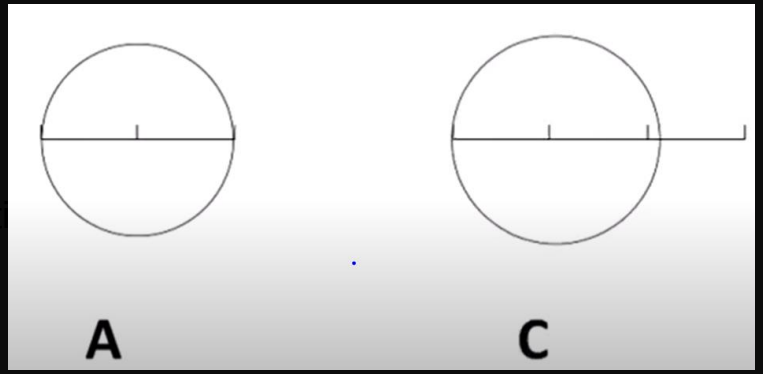
- effect size

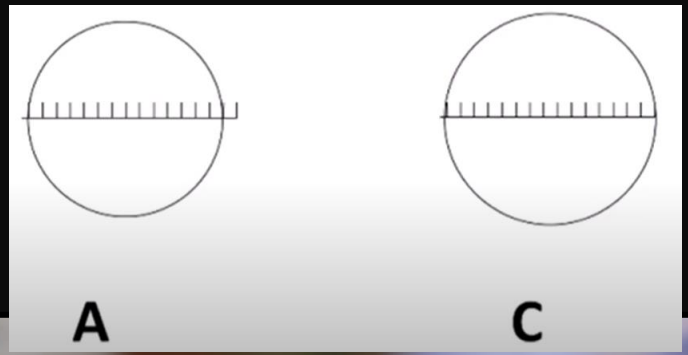


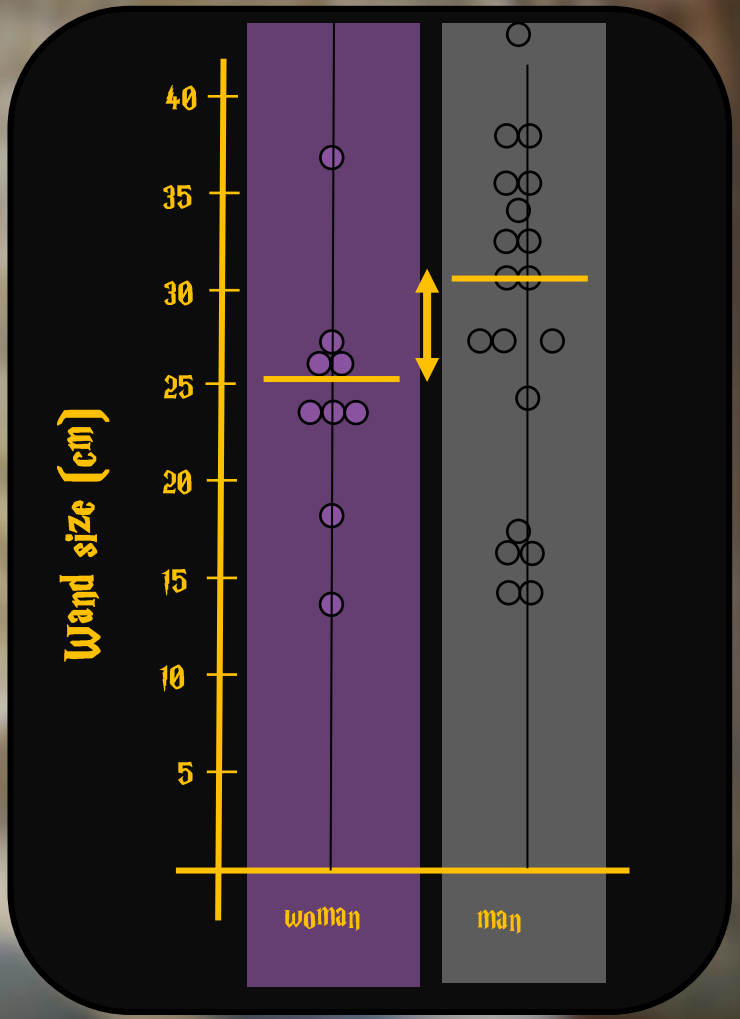
What does power depend on?

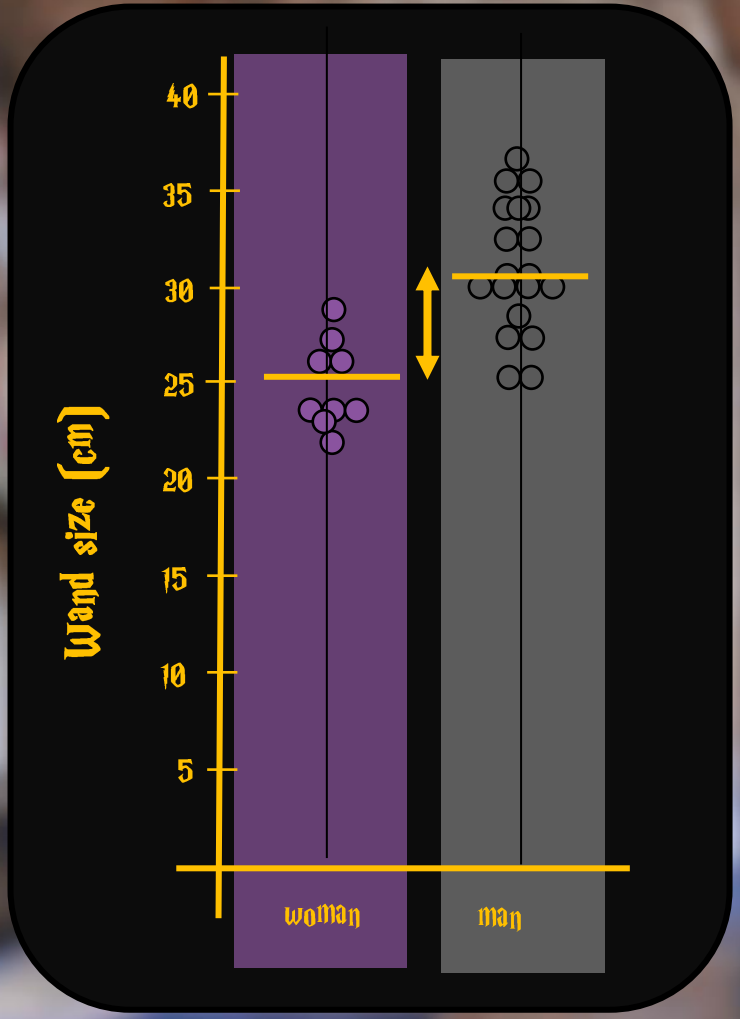
- measure precision





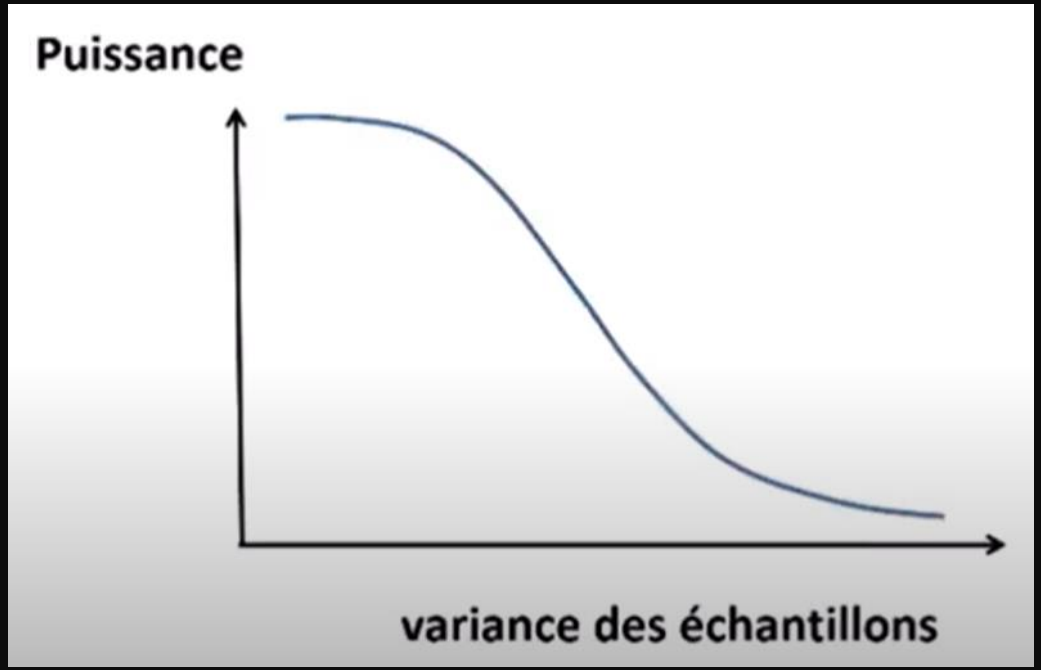


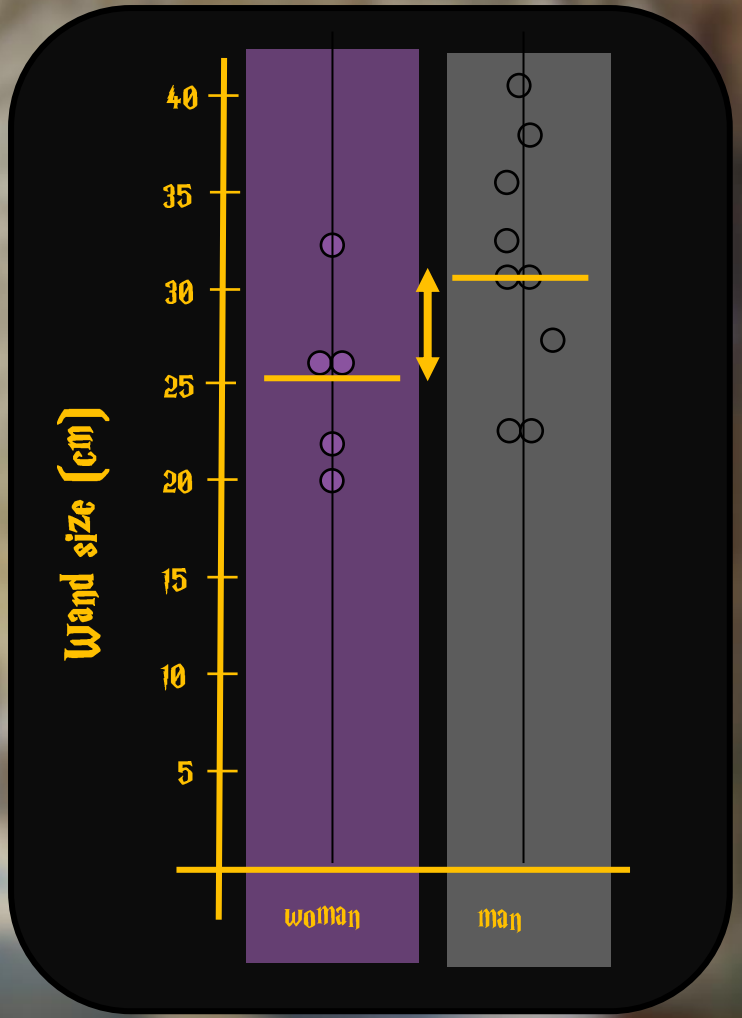

Wand size (cm)
40
35
30
25
20
15
10
5
woman
man

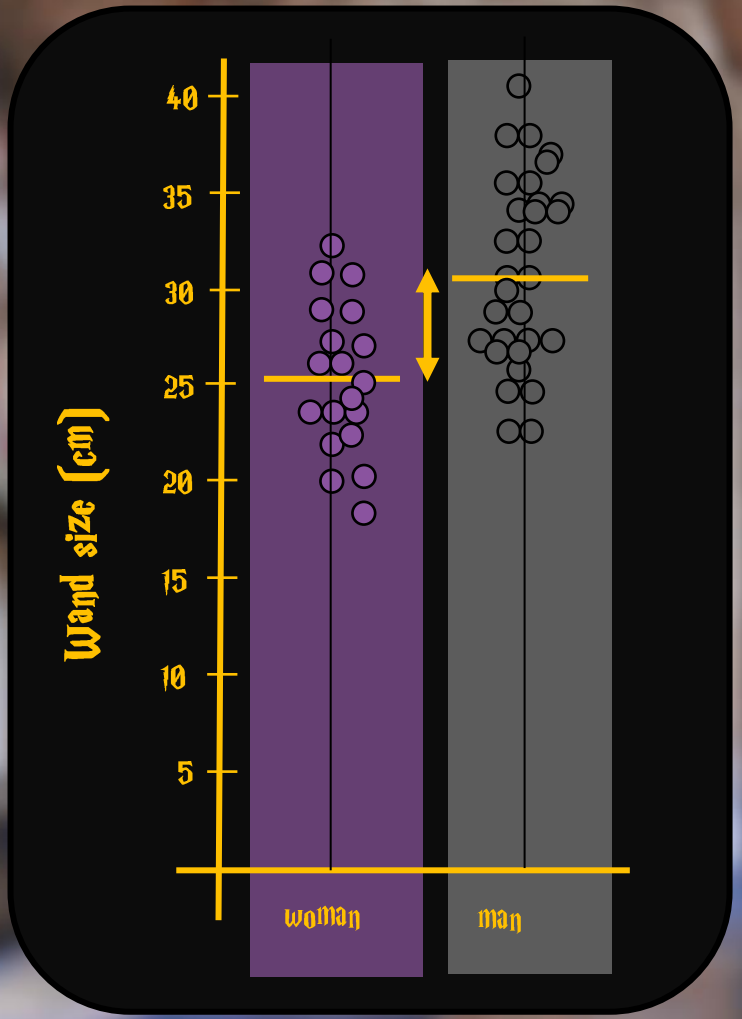

Wand size (cm)
40
35
30
25
20
15
10
5
woman
man

What does power depend on?

- dispersion

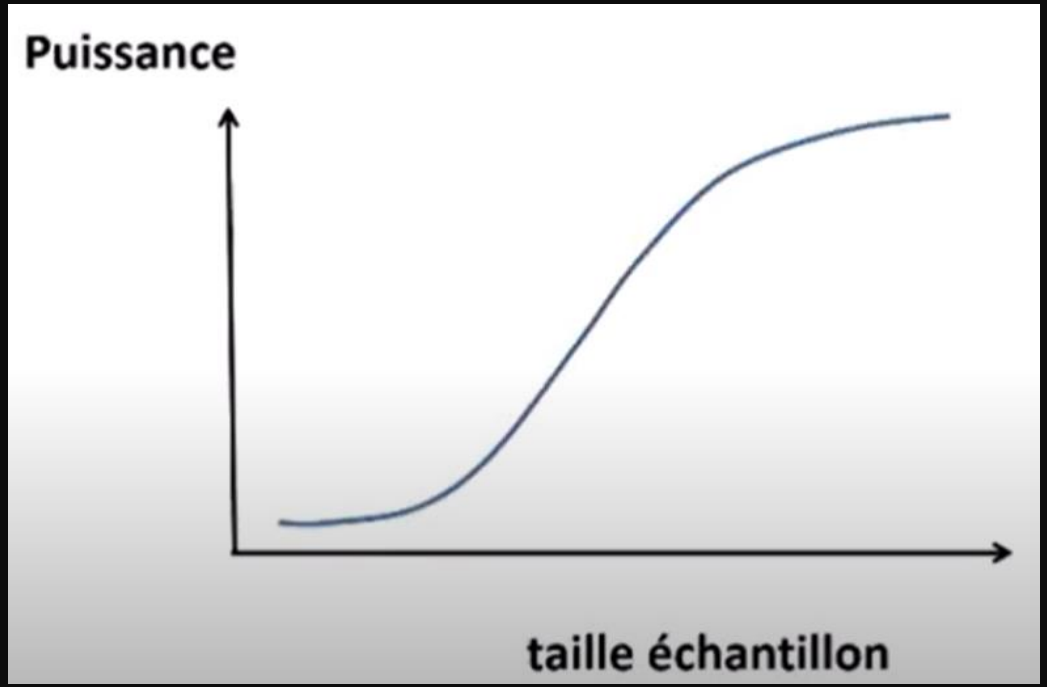



Wand size (cm)
40
35
30
25
20
15
10
5
woman
man


Wand size (cm)
40
35
30
25
20
15
10
5
woman
man

What does power depend on?

- sampling size



What does power depend on?

- Effect size
- Measurement accuracy
- Sampling size
- Dispersion
- Alpha risk

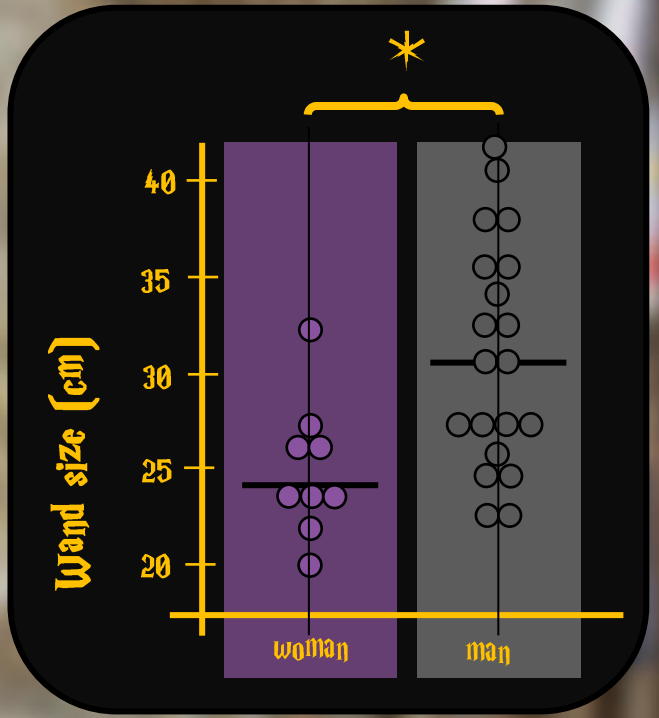
How to increase the power of a test?

1)reduce noise (control of other factors, laboratory experiments)

La taille d'échantillonnage

2) amplify the magnitude of the effect (dose amplification, stimulus amplification, etc.)

3) increase sampling (with more and more measurements, randomness becomes negligible)



Help to decision

- p-value
- Size effect

"Absence of significance indicates small effect size".

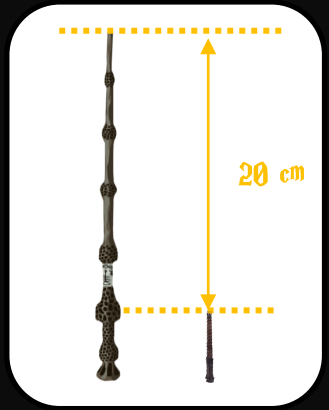
"My p-value is very small so the effect is big".


"Absen ance
indicates sh ct
size".


"M alue is
ve e
effec is big".

P-value = 0,10

Mean difference = 20 cm



P-value = 0,004

Mean difference = 5 cm

P-value = 0,004

Mean difference = 0,5 cm

Lack of data ?

Offense ?

Not a big deal ?

Size matters...

Mean difference

cohen's D

d = 0,2 (small effect)

d = 0,5 (moderate effect)

d = 0,8 (strong effect)

Paired t-test

$$d = \frac{mean_D}{SD_D}$$

Welch t-test

$$d = \frac{m_A - m_B}{\sqrt{(Var_1 + Var_2)/2}}$$

Independant t-test

$$d = \frac{m_A - m_B}{SD_{pooled}}$$

$$SD_{pooled} = \sqrt{\frac{(n_1-1)SD_1^2 + (n_2-1)SD_2^2}{n_1+n_2-2}}$$

Statistical significance

p-value :

- does not measure effect size

Statistical tests are a decision-making tool; they cannot replace knowledge of the field, of what is or isn't "scientifically" or "biologically" important.

T-test : application condition

The two samples must be independent.

Data must be continuous.

Data must be normally distributed.

The variances of the two samples must be equal or very similar. If the variances are very different, the t-test may give erroneous results.

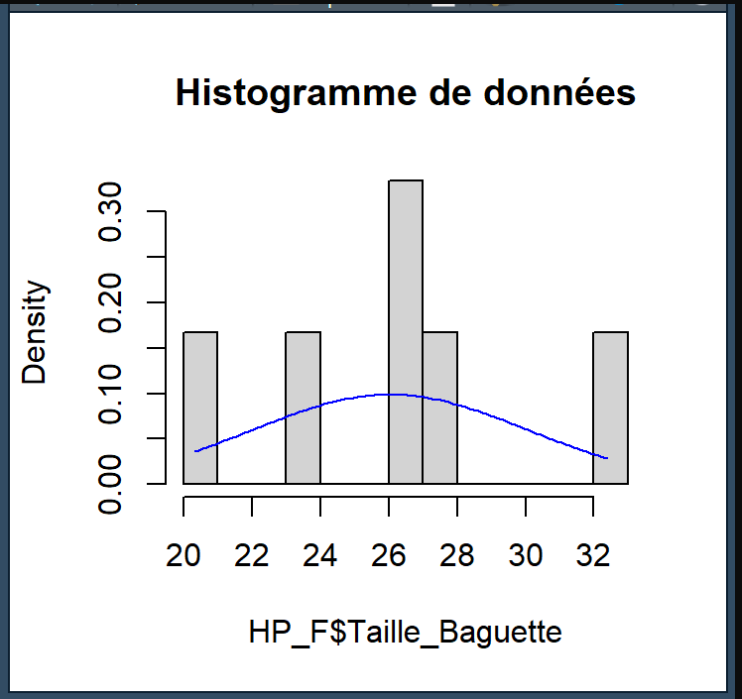
Normality hypothesis

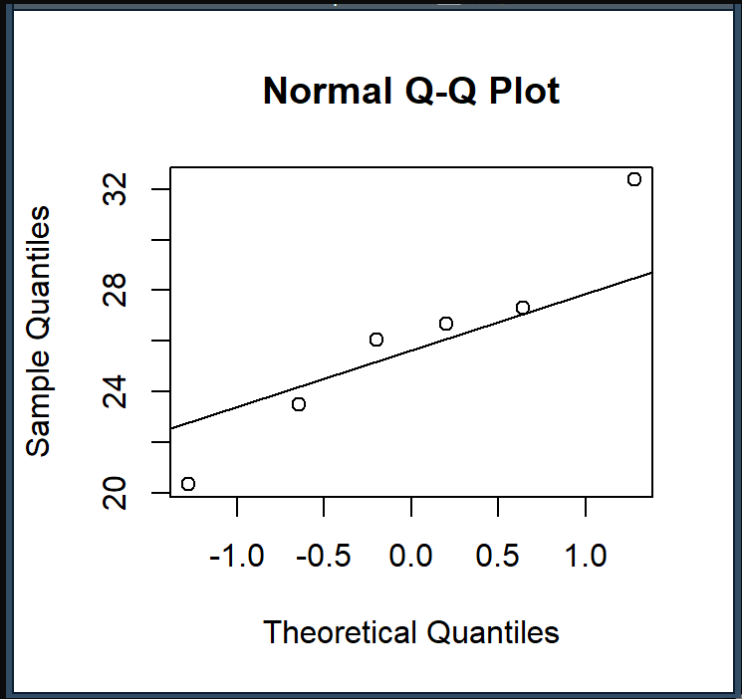
3 ways to check:

- -Shapiro test
- -Graphically with a histogram
- -Graphically with a qqplot

T-test : Normality hypothesis

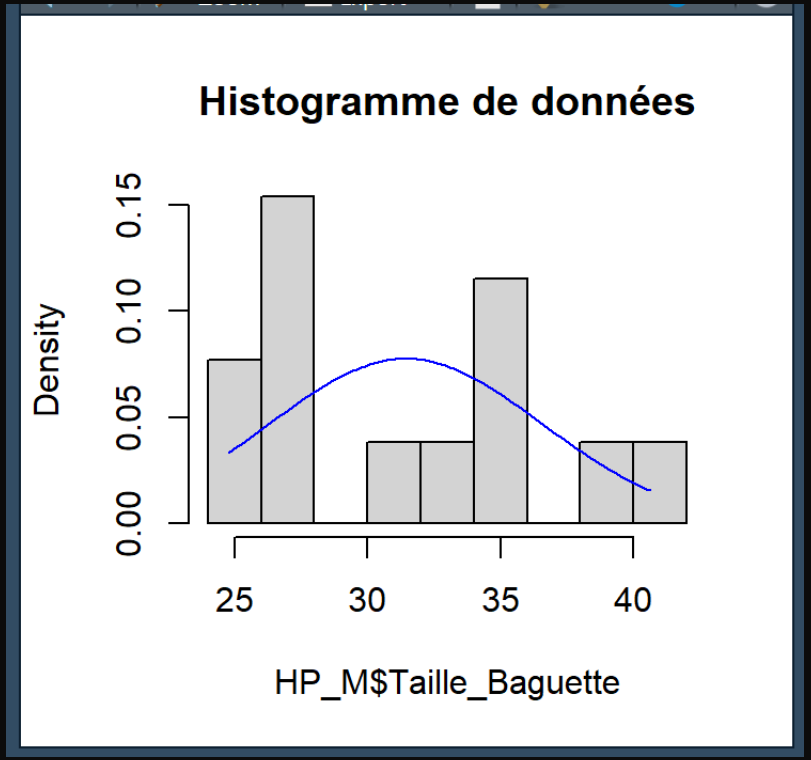
```
> shapiro.test(HP_F$Wand_size)

        Shapiro-Wilk normality test

data:  HP_F$Wand_size
W = 0.96877, p-value = 0.8841
```

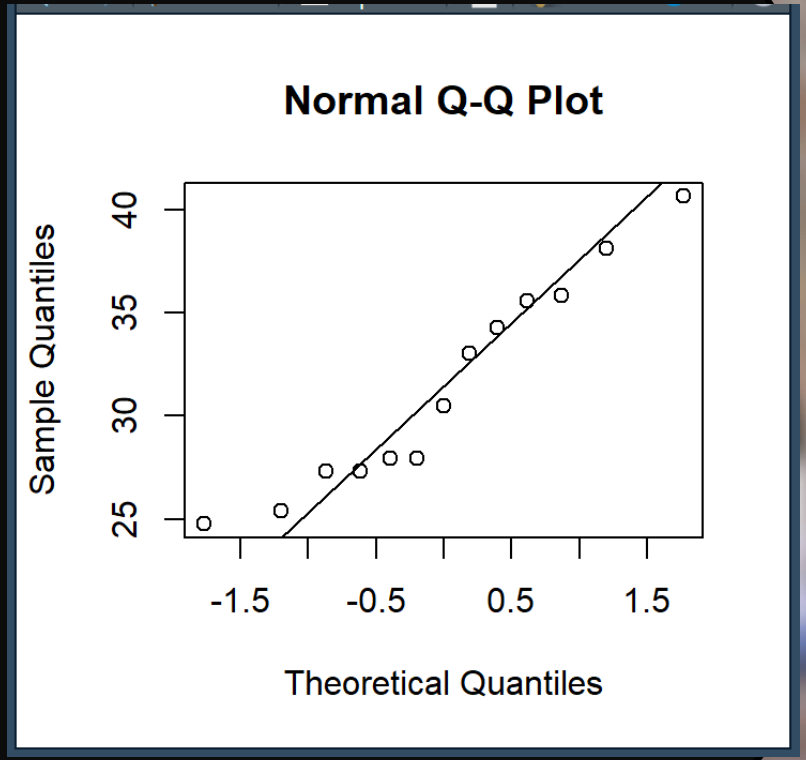




T-test : Normality hypothesis

```
> shapiro.test(HP_M$Wand_size)

        Shapiro-Wilk normality test

data:  HP_M$Wand_size
W = 0.93123, p-value = 0.3537
```





T-test : homoscedasticity check

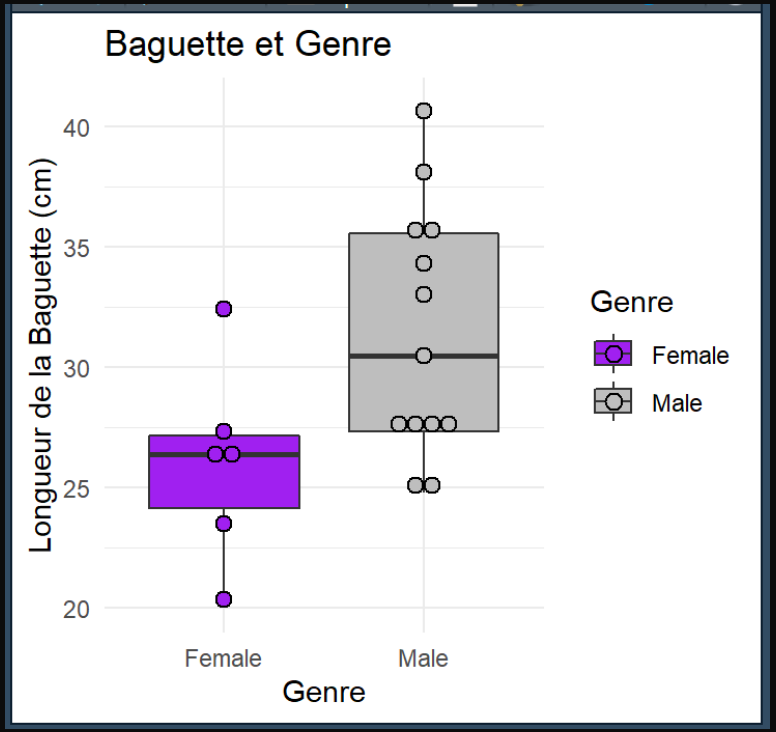
Homoscedasticity = Homogeneity of variances

Comparison of 2 variances ?

T-test : homoscedasticity check

```
# A tibble: 2 x 2
  Genre  variance
  <chr>     <dbl>
1 Female     16.3
2 Male       26.4
```

```
> 16.3/26.4
[1] 0.6174242
```



T-test : homoscedasticity check

```
> var.test(Wand_size ~ Gender, data = HP)

        F test to compare two variances

data:  Wand_size by Gender
F = 0.61744, num df = 5, denom df = 12, p-value = 0.6214
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1586789 4.0285231
sample estimates:
ratio of variances
         0.6174408
```

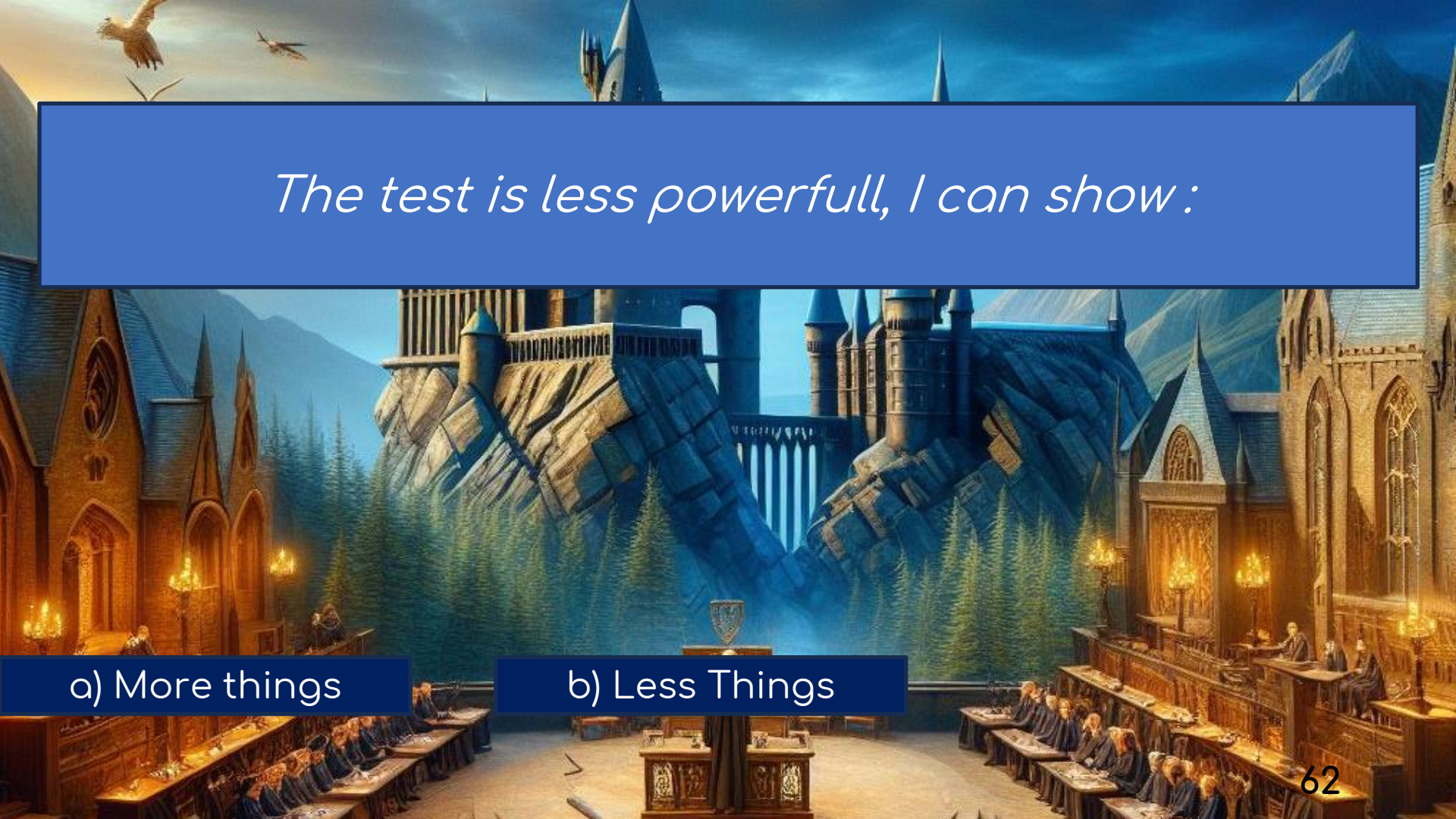
Parametric vs non-parametric tests

Parametric tests impose conditions on the distribution (normality/homoscedasticity) of the data series tested.

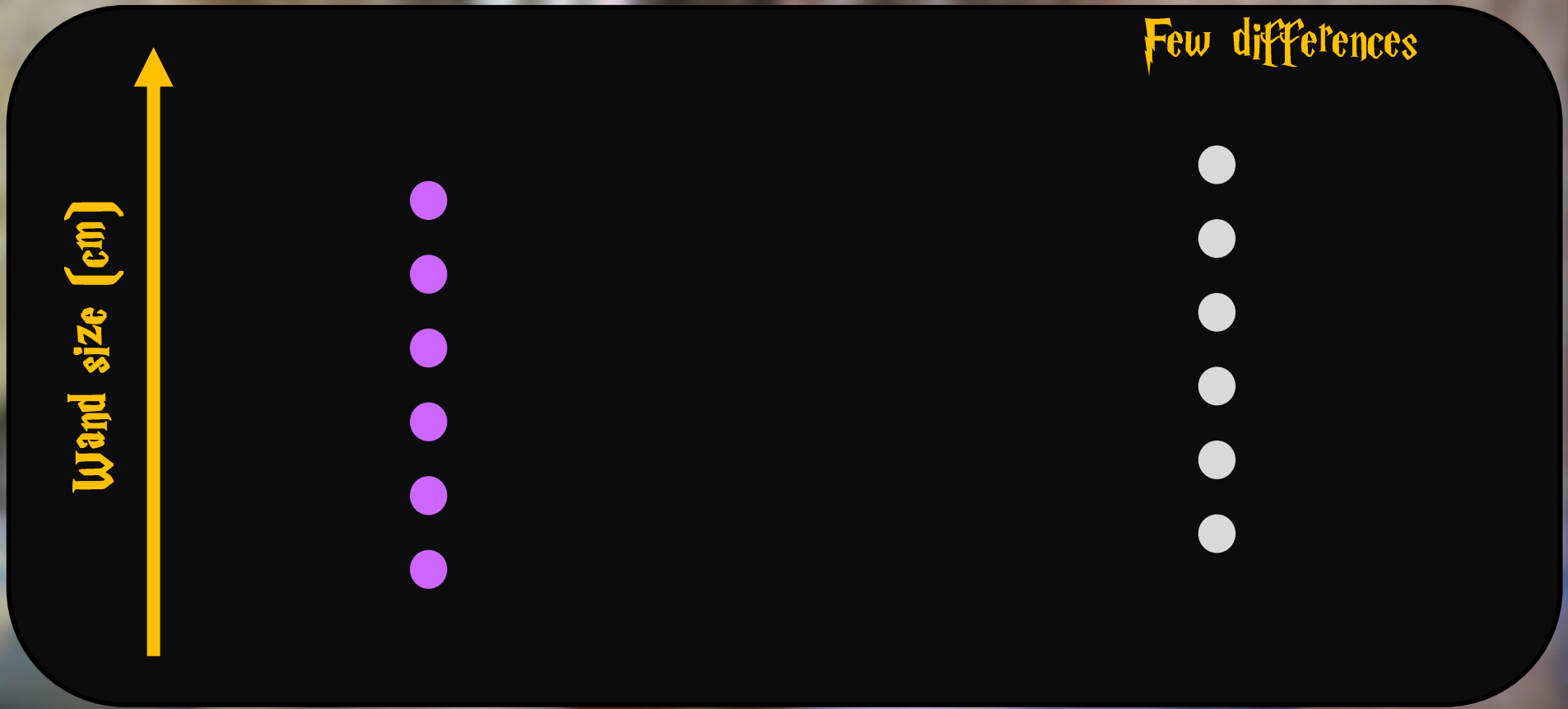
Conditions not met : non-parametric tests.

Based on a notion of ranks.

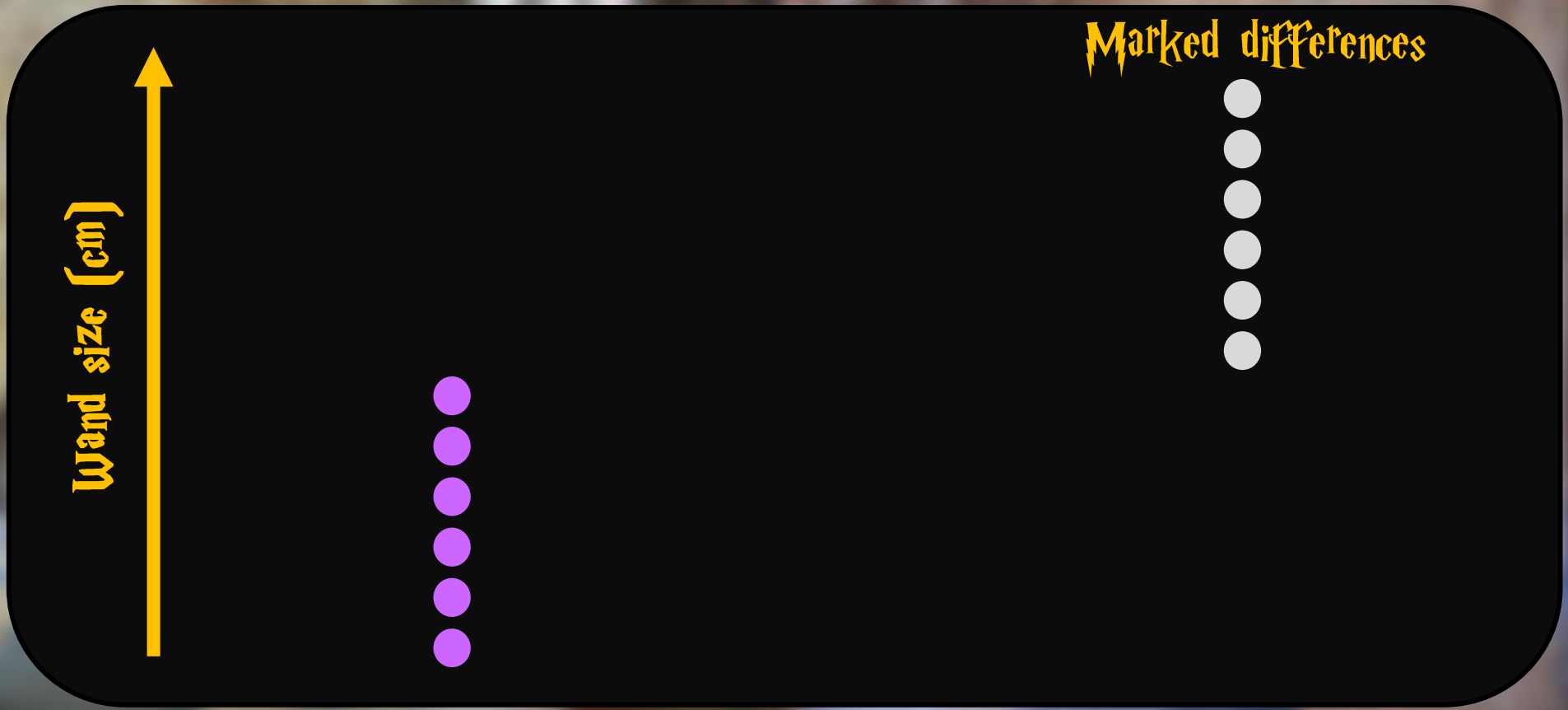
The disadvantage: they are less powerful.


The test is less powerfull, I can show :
a) More things
b) Less Things

Mann-Whitney test : the logic of rank tests



Mann-Whitney test : the logic



Mann–Whitney test

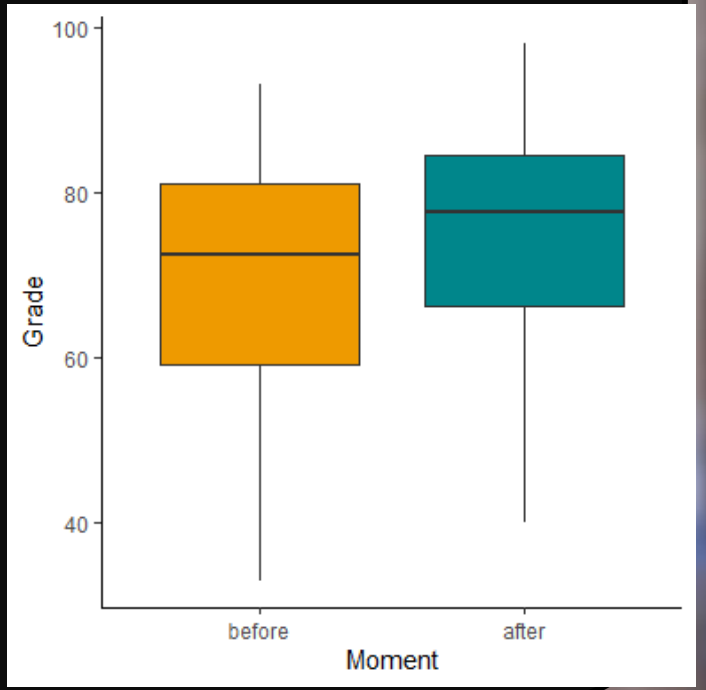
```
> wilcox.test(HP$Wand.Length ~ HP$Gender)

        Wilcoxon rank sum test with continuity correction

data:  HP$Wand.Length by HP$Gender
W = 7.5, p-value = 0.00405
alternative hypothesis: true location shift is not equal to 0
```

Beware of the independence assumption !!!!!

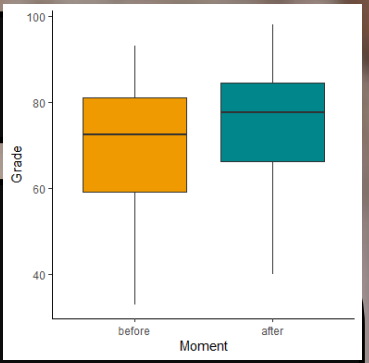
One individual should appear once (and only once) in the dataset

Paired t-test

character	house	before	after
Harry Potter	Gryffondor	68	75
Hermione Granger	Gryffondor	93	98
Ron Weasley	Gryffondor	33	40
Draco Malfoy	Serpentard	77	80



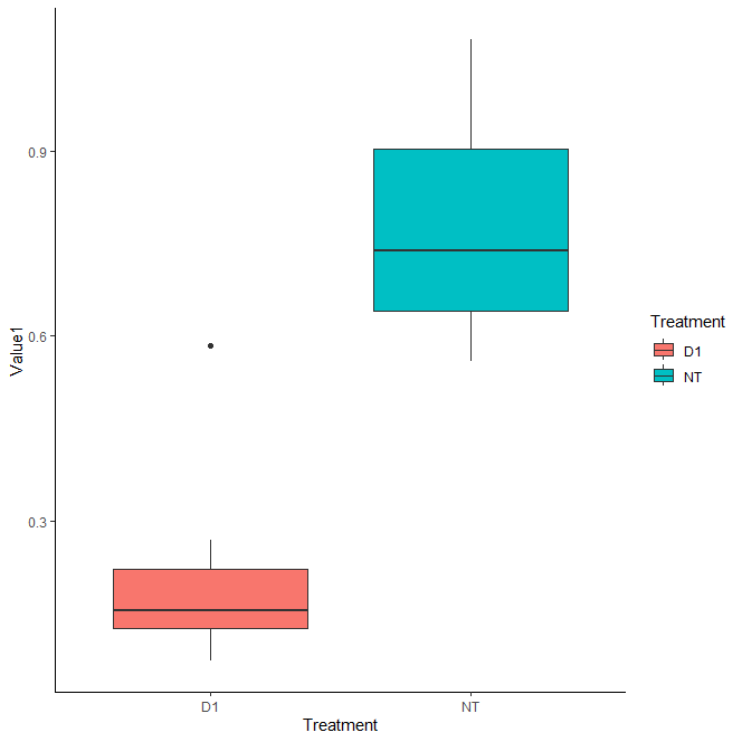
Paired t-test



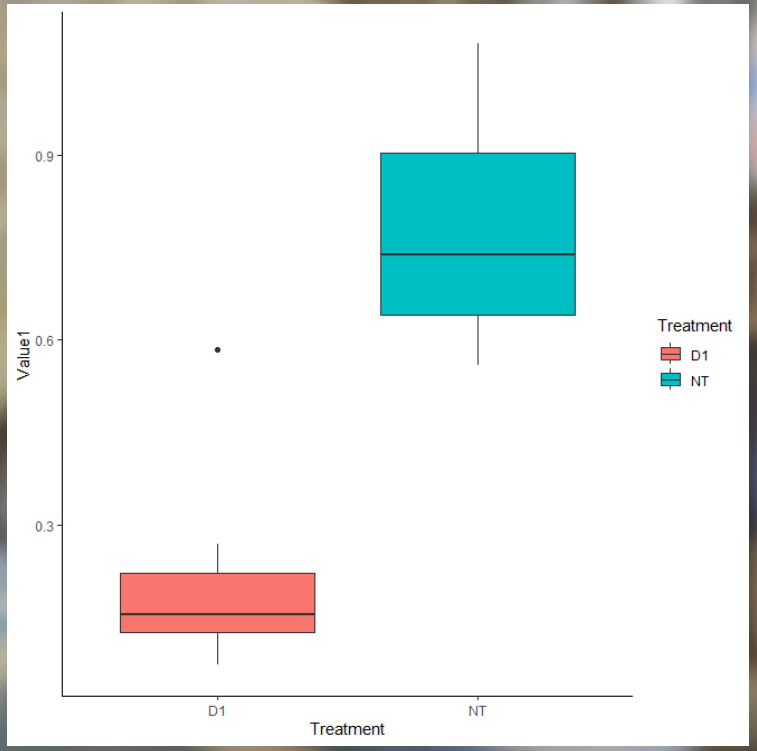
R

```
> t.test(arithmancy$before, arithmancy$after, paired = TRUE)

        Paired t-test

data:  arithmancy$before and arithmancy$after
t = -5.7446, df = 3, p-value = 0.01048
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -8.54696 -2.45304
sample estimates:
mean difference
           -5.5
```

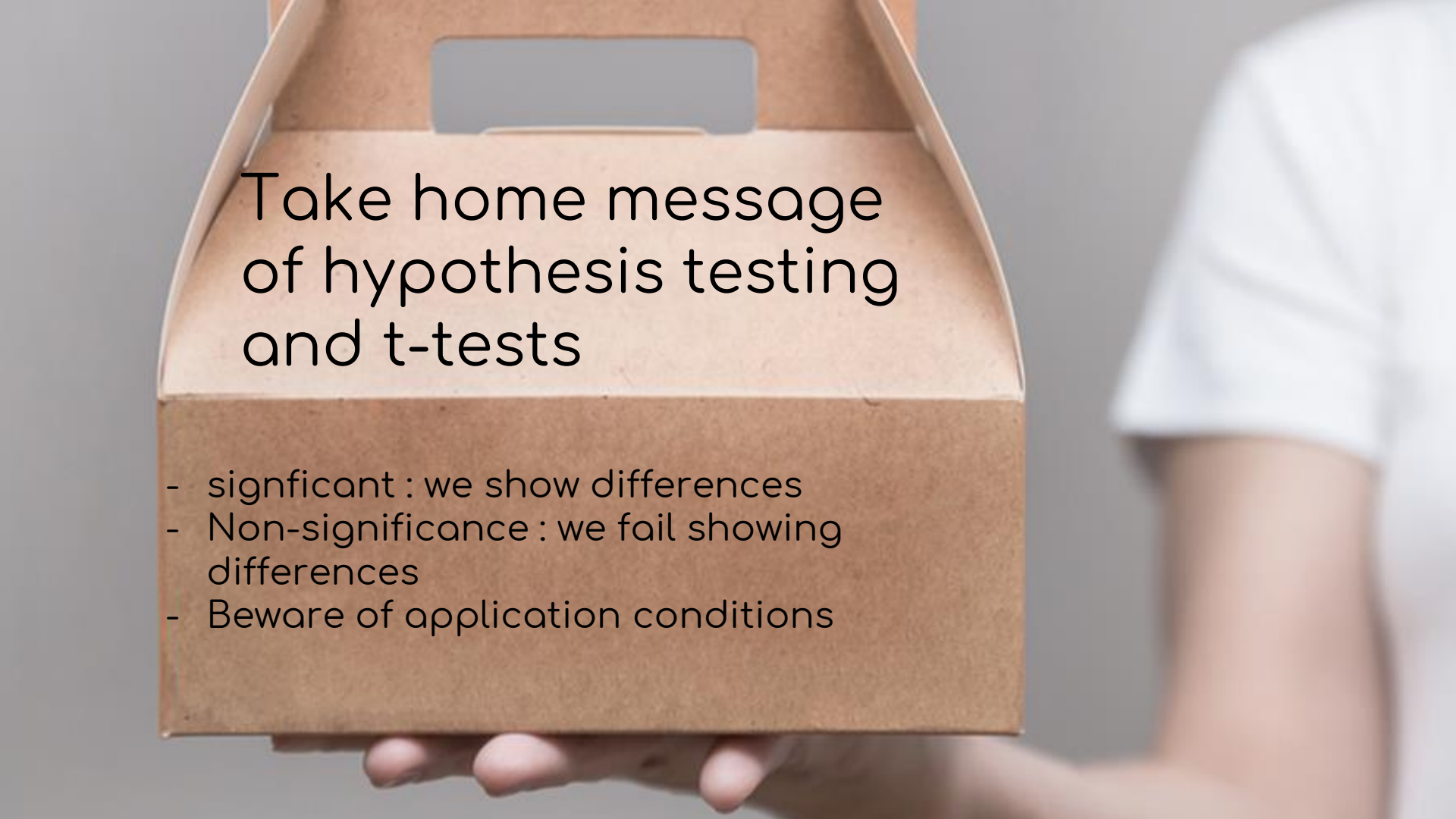
Your Field



Your Field




```
        Welch Two Sample t-test

data:  Value1 by Treatment
t = -6.2493, df = 13.542, p-value = 2.475e-05
alternative hypothesis: true difference in means between group D1
 and group NT is not equal to 0
95 percent confidence interval:
 -0.7691222 -0.3751528
sample estimates:
mean in group D1 mean in group NT
       0.2074375        0.7795750
```



Take home message of hypothesis testing and t-tests

- signficant : we show differences
- Non-significance : we fail showing differences
- Beware of application conditions