Statical tests

Anissa Guillemin

15 mars 2017

Historical generalities

- Hypothesis Tests
- Bayesian statistics

Historical generalities : ”Frequentists”

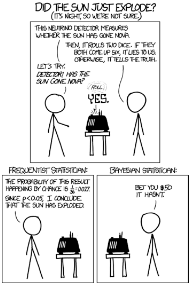
- Hypothesis Tests $\rightarrow$ we reject or not H_0 with a α risk

Historical generalities : ”Bayesian”

- Bayesian statistics

Bayes theorem :

posterior probability $=$ likelihood $\times$ prior probability



Statistical notion #1

- Descriptive statistics $\rightarrow$ data visualisation
- Inferential statistics $\rightarrow$ to draw conclusions about the entire population from samples

Statistical notion #2

A test is a rule to decide between H_0 or H_1 hypothesis. We compute a statistics and compare it to a decisional threshold ; if the value of statistics is $\leq$ threshold, observe such a value is too less likely considering the risk we are ready to take.

Statistical notion #2

- Hypothesis H_0 : null hypothesis

The observed differences are not different from random fluctuations
This is H_0 hypothesis that is controlled during the test

Statistical notion #2

- Hypothesis H_0 : null hypothesis

The observed differences are not different from random fluctuations
This is H_0 hypothesis that is controlled during the test

- Hypothesis H_1 : alternative hypothesis

Negation of H_0 hypothesis

Be careful with test conclusions : accept $H_0 \neq H_0$ is true. We can only reject or not H_0, never accept it !

Statistical notion #3

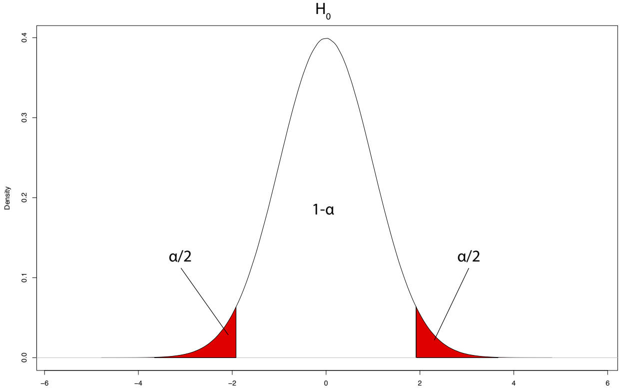
- α & β risks

		Decision	
		Reject H_0 -> H_1	«Accept» H_0 -> H_0
Actual	H_0 true	Type I error α-risk False positive	Correct decision Confidence interval = 1-α True negative
	H_1 true	Correct decision Power = 1-β True positive	Type II error β-risk False negative

Statistical notion #3

- α & β risks

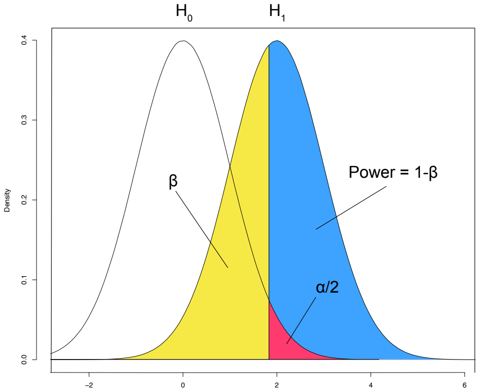
In the case of bilateral test where the statistic distribution is symetric



Statistical notion #3

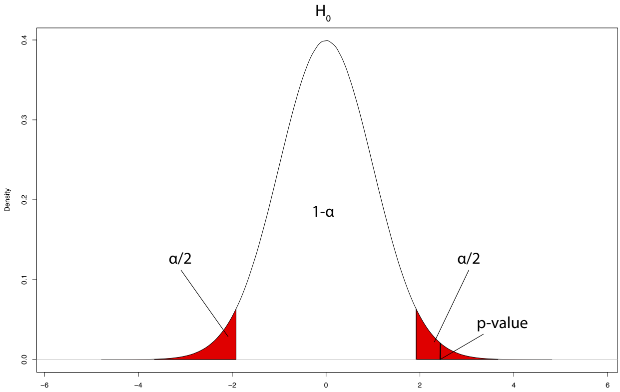
- α & β risks

In the case of bilateral test where the statistic distribution is symetric



Statistical notion #4

- P-value : level of significance.
  This is the probability the difference observed in population is the same than in the samples.



Statistical notion #5

- Degree of freedom : the number of values in the final calculation of a statistic that are free to vary.
  Without estimation, each value can take on any number $\rightarrow$ Each value is completely free to vary
- n = sample size

Remarks

- β risk
- Interpretations

Probability distributions - Plan

- Generalities
- Binomial & Bernoulli
- Poisson
- Exponential
- Normal & log-normal
- Gamma & Chi-squared
- Normal

Generalities

- Random variable $X : \Omega \to \mathbb{R}$
- Law of probability of a random variable : allow us to know occurrences of values of a variable X.

Bernoulli

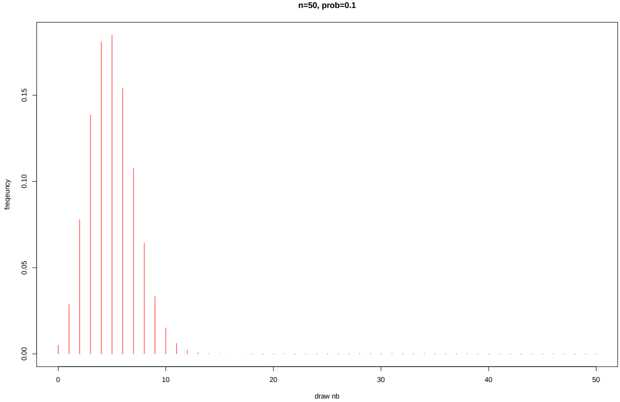
- Bernoulli distribution : a random draw

$X \sim$ Bern(p) with p the probability of success on n draws
Success or failure with p the probability of success for one draw.

Binomial

- Binomial distribution : n random draw

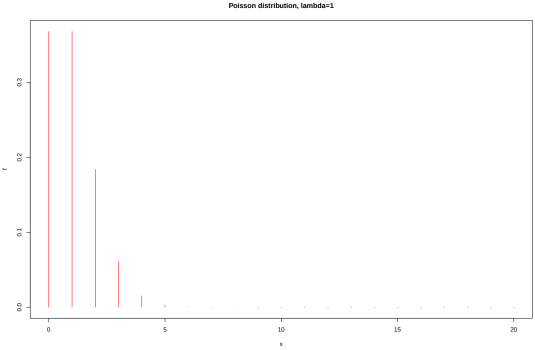
X $\sim$ Binom(p,n) with p the probability of success on n independent attempts
Success or failure with p the probability of success for n draw. This test represent a characteristic in a sample.



Poisson

- Poisson distribution : for rare events

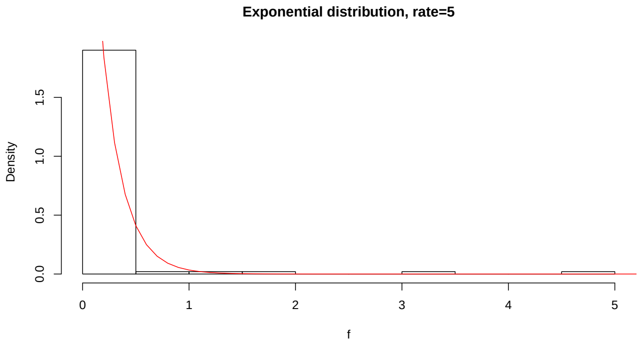
$X \sim$ Pois(λ) with λ = Mean = Variance
For discrete variable. This is the continue version of Bernoulli law.
Example : counting of UFC in some petri dish containing antibiotics



Exponential

- Exponential distribution : lifetime without aging

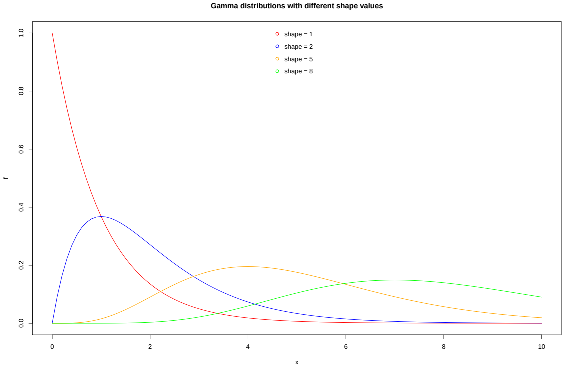
  X $\sim$ Exp(λ) with λ the mean nb of event per time or volume unit
  Memoryless
  Example : Radioactive disintegration



Gamma

- Gamma distribution : sum of exponential distributions

  X $\sim$ Gamma(α,λ) with α the nb of added variables
  Example : optimal staff in a call center

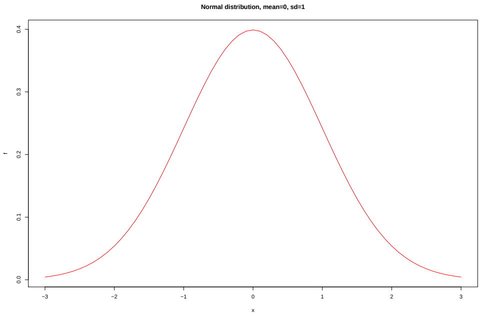


Normal

- Normal distribution : most famous in statistics

X $\sim$ N(μ,σ) with μ the mean and σ the standard deviation
Example : size
Useful when the distribution is **reduced centered** :

$$\text{Z score} = \frac{X - \mu}{\sigma}$$



- Comparison of two means
- Comparison of frequencies
- Linear correlation
- F-test of equality of variances
- Conditions
- Abuses

Comparison of two means

- Student test

Student law is a symmetric law with heavier tails than normal law for weak df
The decision variable t follow a Student law with $n_1 + n_2 - 2$ df.
X $\sim$ T(μ,σ,ν) with μ the mean, σ the variance and ν the degree of freedom
$H_0 : \mu_1 = \mu_2$
$H_1 : \exists$ a value $\Delta \neq 0$ for $\mu_1 - \mu_2 = \Delta$

Comparison of two means

- Student test

```
> t.test(dataBWT,a)
Welch Two Sample t-test
data: dataBWT and a
t = 0.36582, df = 375.05, p-value = 0.7147
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-123.1589 179.4599
sample estimates:
mean of x mean of y
2944.656 2916.506
```

Comparison of two means

- Student test

```
> t.test(dataBWT,dataBTW2)
Welch Two Sample t-test
data: dataBWT and dataBWT2
t = -13.878, df = 189.52, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-13231.050 -9937.866
sample estimates:
mean of x mean of y
2944.656 14529.114
```

Comparison of frequencies

- For proportion to a reference

$H_0 : p_0 = p_1$
$H_1 : p_1 \leq p_0$
condition : np_0 and n(1-p_0) $\geq$ 5

```
> prop.test(x, n, p, alternative = c("two.sided", "less", "greater"))
```

Ex : Test if the proportion of pregnant women $\leq$ 25 years and HIV+ is equal to 0.1%.
We would like to know if this prevalence is lower than the theoretical proportion p_0 = 0.1

HIV -	HIV +
137	10

Comparison of frequencies

- For proportion to a reference

```
> prop.test(10,147, p = 0.1,alternative = "less")
1-sample proportions test without continuity correction
data: 10 out of 147, null probability 0.1
X-squared = 1.6697, df = 1, p-value = 0.09815
alternative hypothesis: true p is less than 0.1
95 percent confidence interval:
0.000000 0.110572
sample estimates:
p
0.06802721
```

Comparison of frequencies

- For proportion to a reference

```
> prop.test(10,147, p = 0.1,alternative = "less")
1-sample proportions test without continuity correction
data: 10 out of 147, null probability 0.1
X-squared = 1.6697, df = 1, p-value = 0.09815
alternative hypothesis: true p is less than 0.1
95 percent confidence interval:
0.000000 0.110572
sample estimates:
p
0.06802721
```

Conclusion : we can't respond positively at the question with a chosen risk α.

We can make a binomial test for lower samples

```
> binom.test(10,147,0.1,alternative = "less")
```

In this case the result is not different

Comparison of frequencies

- Between two proportions

$H_0 : p_1 = p_2$
$H_1 : p_1 \neq p_2$
condition : n_1p, n_1(1-p), n_2p and n_2(1-p) ≥ 5
with

$$p = \frac{n1p1 + n2p2}{n1 + n2}$$

```
> prop.test(tableau)
```

Ex : Test if the mother treatment change the HIV status of the baby. To do that we compare with the test the proportions of baby HIV+ with a mother under treatment or not

Mother treated or not		Baby HIV+/-	
		HIV -	HIV +
	T -	139	59
	T +	152	41

Comparison of frequencies

- Between two proportions

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

```
> prop.test(table)
2-sample test for equality of proportions without continuity correction
data: table
X-squared = 3.7574, df = 1, p-value = 0.05257
alternative hypothesis: two.sided
95 percent confidence interval:
-0.0004122543 0.1715013839
sample estimates:
prop 1 prop 2
0.2979798 0.2124352
```

With the chosen α risk this is not possible to show that the treatment have an effect on the HIV status of babies

Comparison of frequencies

- Between two variables : X^2 of independence

H_0 : The two variables are independant
H_1 : The two variables are dependant
condition : All theoretical class size ≥ 5

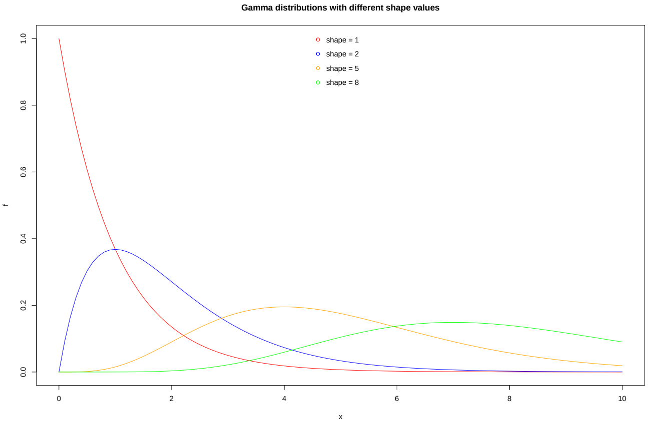
With the same example as before

```
> chisq.test(table)
Pearson's Chi-sqared test
data: table
X-squared = 3.7574, df = 1, p-value = 0.05257
```

The result is the same

Comparison of frequencies

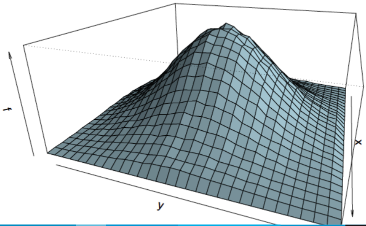
- X $\sim$ Gamma(k=ν/2,σ=2) equivalent to X $\sim$ $X^2(\nu)$ with ν the df.



Linear correlation

- Pearson test

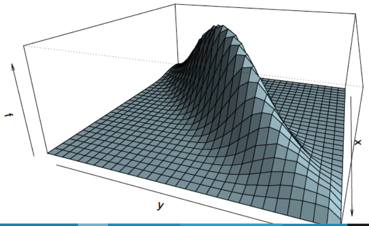
H_0 : r = 0
H_1 : r $\neq$ 0 with r the correlation coefficient between two quantitative variables X and Y
This value is an indicator of the point cloud elongation. The more $|r|$ is near from 1, the more points are line up.
Condition : X and Y have to follow a Normal bivariate law (elliptical point cloud)



Linear correlation

- Pearson test

H_0 : r = 0
H_1 : r $\neq$ 0 with r the correlation coefficient between two quantitative variables X and Y
This value is an indicator of the point cloud elongation. The more $|r|$ is near from 1, the more points are line up.
Condition : X and Y have to follow a Normal bivariate law (elliptical point cloud)



Linear correlation

- Pearson test

H_0 : r = 0
H_1 : r $\neq$ 0 with r the correlation coefficient between two quantitative variables X and Y
This value is an indicator of the point cloud elongation. The more $|r|$ is near from 1, the more points are line up.
Condition : X and Y have to follow a Normal bivariate law (elliptical point cloud)

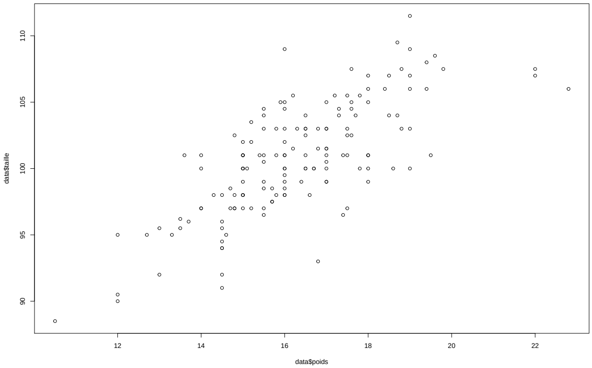
```
> cor.test(x, y, method = c("pearson", "kendall", "spearman"))
```

Here we would like to know if a linear correlation exists between size and weight of children.

Linear correlation

- Pearson test

H_0 : r = 0
H_1 : r $\neq$ 0



Linear correlation

- Pearson test

H_0 : r = 0
H_1 : r $\neq$ 0

```
cor.test(datapoids,datataille,method="pearson")
Pearson's product-moment correlation
data: datapoids and datataille
t = 13.433, df = 150, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.6570527 0.8036174
sample estimates:
cor
0.7389562
```

Linear correlation

- Pearson test

R calculated r=0.7389562 with a weak p-value : 2.2e-16. With a chosen α risk (0.01 for exemple), we can conclude that a linear association between these two variable exists.

Linear correlation

- Be careful, a correlation between two observed variables does not necessarily a cause and effect relationship !
- Pearson is highly influenced by extreme values.
- You have to see the plot between the two variables before starting the test.

F-test of equality of variances

- Fisher-Snedecor

Fisher test is the ratio between the two corrected variances following a FS law at (n_1-1,n_2-1) df

$H_0 : \sigma_1 = \sigma_2$

$H_1 : \exists$ a value $\Delta \neq 0$ for $\sigma_1 - \sigma_2 = \Delta$

```
> var.test(dataBWT,dataBWT2)
F test to compare two variances
data: dataBWTanddataBWT2
F = 0.004052, num df = 188, denom df = 188, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.003041773 0.005397621
sample estimates:
ratio of variances
0.004051955
```

F-test of equality of variances

- Fisher-Snedecor

Fisher test is the ratio between the two corrected variances following a FS law at (n_1-1,n_2-1) df

$H_0 : \sigma_1 = \sigma_2$

$H_1 : \exists$ a value $\Delta \neq 0$ for $\sigma_1 - \sigma_2 = \Delta$

```
> var.test(dataBWT,a)
F test to compare two variances
data: dataBWT and a
F = 0.90436, num df = 188, denom df = 188, p-value = 0.4914
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.6788951 1.2046983
sample estimates:
ratio of variances
0.9043582
```

Conditions

These tests are the strongest tests (high 1-β) only if data follows certain conditions

- Normality : data follows normal law
- Homoscedasticity : equality of variances (F-test)

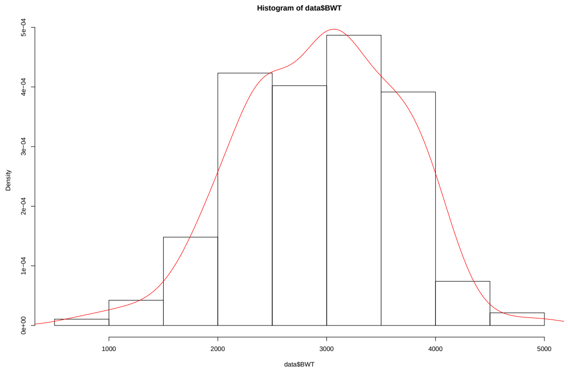
Control the normality of your data

- Histogram
- Boxplot
- qqplot
- Skewness & Kustosis
- Shapiro test

Control the normality of your data

- Histogram

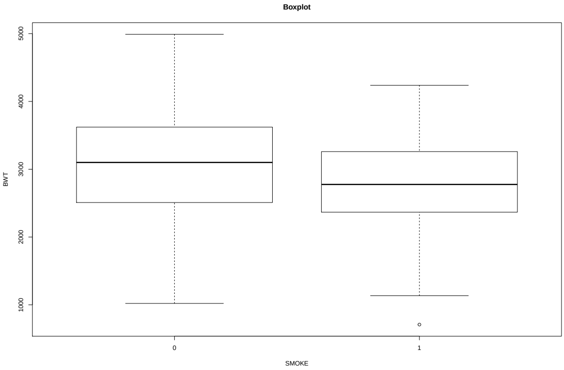
```
> hist(dataBWT,freq=F)
> den<-density(dataBWT)
> lines(den,col="red")
```



Control the normality of your data

- Boxplot

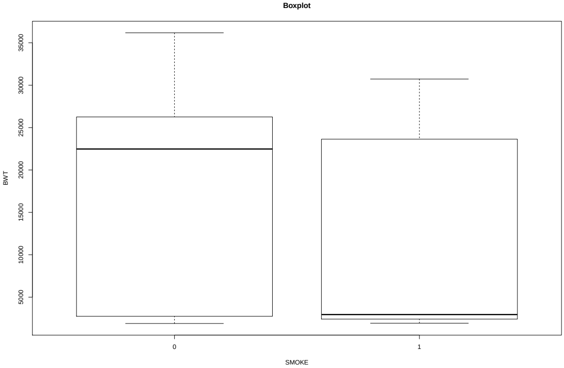
```
> boxplot(dataBWT ~ dataSMOKE,xlab="SMOKE",ylab="BWT",main="Boxplot")
```



Control the normality of your data

- Boxplot

```
> boxplot(dataBWT2 ~
dataSMOKE,xlab="SMOKE",ylab="BWT2",main="Boxplot2")
```

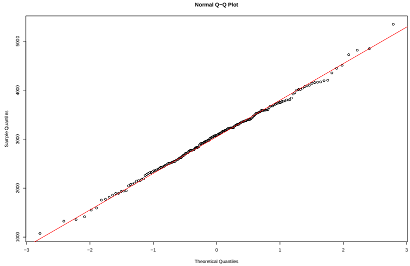


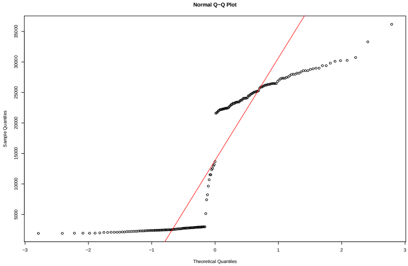
Control the normality of your data

- QQ-plot

”Droite de Henry” is represented in red = a line to a ?theoretical ?, by default normal, quantile-quantile plot which passes through the probs quantiles, by default the first and third quartiles.

```
> qqnorm(dataBWT)
> qqline(dataBWT,col="red")
```

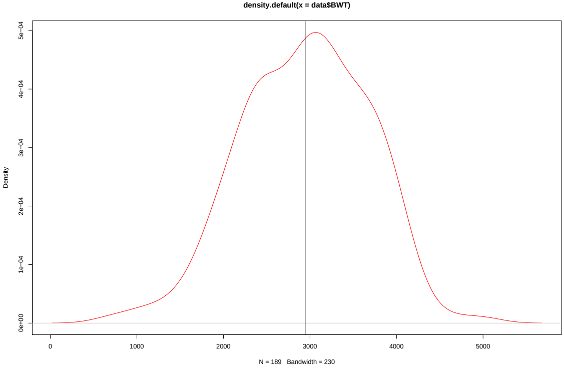




Control the normality of your data

- Skewness & Kurtosis

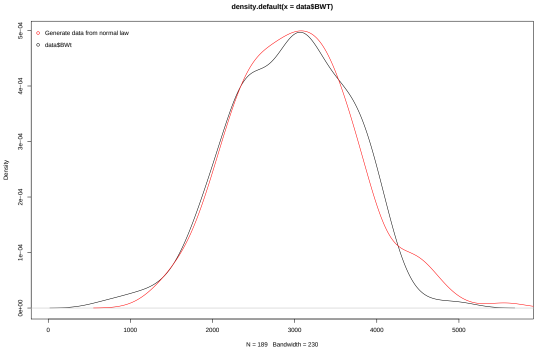
```
> library(e1071)
> skewness(dataBWT) -0.2068467
> kurtosis(dataBWT) -0.1413488
```



Control the normality of your data

- Shapiro-Wilk

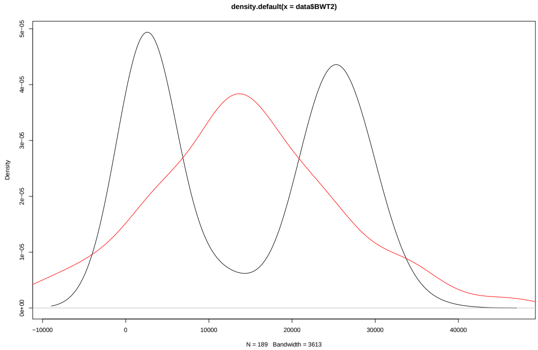
```
> shapiro.test(dataBWT)
Shapiro-Wilk normality test
data: dataBWT
W = 0.99247, p-value = 0.4383
```



Control the normality of your data

- Example : other distribution

```
> shapiro.test(dataBWT2)
Shapiro-Wilk normality test
data: dataBWT2
W = 0.78418, p-value = 2.038e-15
```



Non-parametric tests

If data doesn't follow these two conditions (Normality &/or Homoscedasticity) you have to use other tests, less strong. Non-parametric tests make no assumptions about probability distributions and are based on the ranks of observations.

- To compare two means : Wilcoxon test also called Mann-Whitney test

```
> wilcox.test(x, y)
```

- To test a linear correlation : Spearman test

```
> cor.test(x,y,method="spearman")
```

Abuses

- When we performed a multiple test, we have to make a correction of α risk (Bonferroni, Tukey...)
- Main mistakes are due to a wrong choice of the statistical test. You have to think about the biological question, hypothesis and check if your data satisfy the conditions of tests.
- Don't abuse of p-value : observe your data before starting a test, the p-value is not very informative. The confidence intervalle is often more informative.
- The more tests we performed, the more we expect significant results by chance.

Hypothesis test is like a safeguard that prevent the biologist to early conclude without evidences.

Thank you !