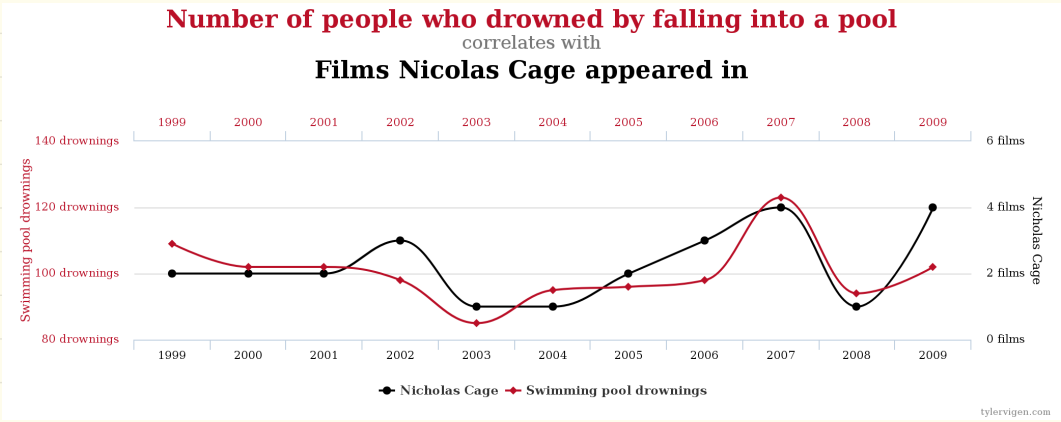
Multiple testing and variable detection

P. Reynaud-Bouret

Year 2017-2018

Chapter 0 : Introduction
A small case study



cf : http://tylervigen.com/spurious-correlations

I Correlation

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad \text{where} \quad Cov(X,Y) = E(XY) - E(X)E(Y)$$
$$\sigma_X = \sqrt{Var X}, \quad \sigma_Y = \sqrt{Var Y}$$

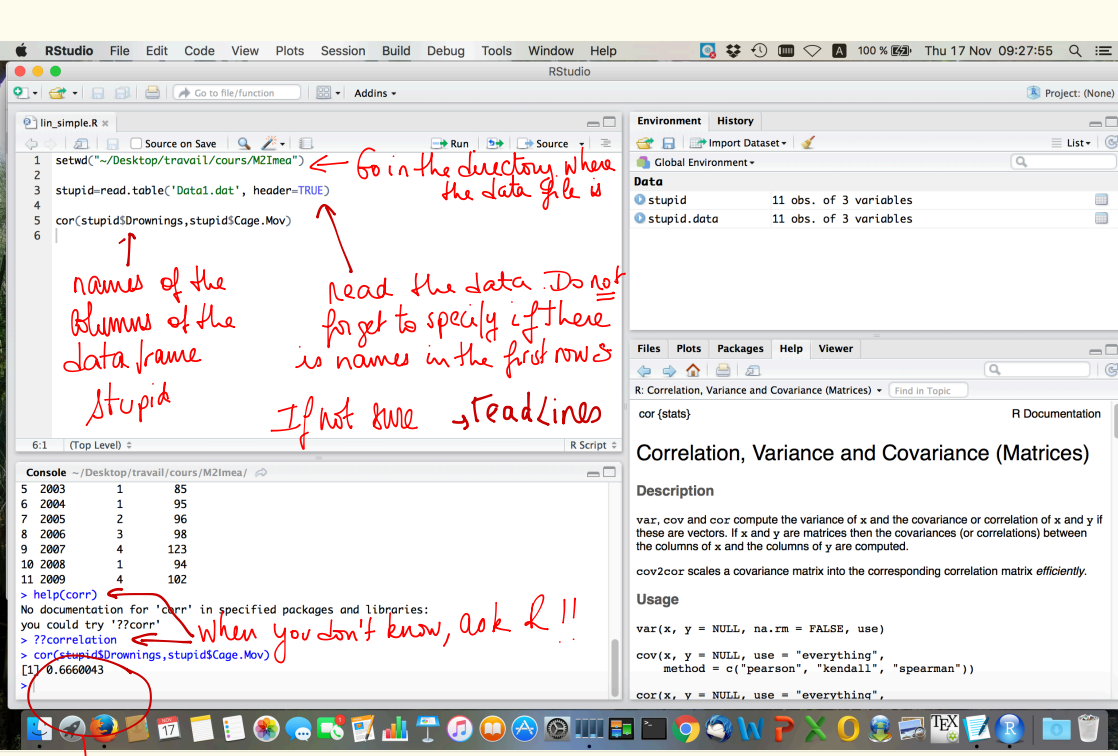
Empirical version

If $(X_1,Y_1), \dots, (X_n,Y_n)$ are iid with the same distribution as (X,Y), then r can be estimated by

$$\hat{r} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) - \bar{X}\bar{Y}}{\sqrt{\frac{1}{n}\sum_i X_i^2 - \bar{X}^2}\sqrt{\frac{1}{n}\sum_i Y_i^2 - \bar{Y}^2}}$$

NB : $\hat{r} \in [-1,1]$

"People" will rarely say that if $|\hat{r}| > 0{,}5$ then there is a strong dependency between the variables X and Y.



What do you think? What can be wrong with the analysis?

- "People" say 0.5 ↪ what does this threshold mean?
  Is there a way to define threshold and have more guarantee?
  Is which guarantee do we want?
- Tyler Vigen searched in a really large database ⇒ maybe this is like cheating!

⇒ The aim of this course is to understand the "guarantees" hidden behind statistical procedures (mainly tests) and to have a critical eye on what can be done!

II Correlation, independance and cause-effect link

1/ Can variables be dependent but not correlated?

2/ Can variables be "independent" in the common sense and still be correlated?

↳ Can you think to examples that we can simulate?

1/ If X and Y are not correlated then $E(XY) = E(X)E(Y)$
if n is really large then $\hat{r}$ should be close to $r = 0$, so below the threshold

⇒ need to find Y depending on X such that $E(XY) = EX\,EY$

⇒ take

$Y = X^2$ + eventually some noise

↳ compute a n sample and $\hat{r}$

2/ It is sufficient to depend on a third variable!

typically n = year and (see Tyler Vigen Web site)

depense in science } are both increasing with the year n.
suicide by hanging }

the main thing is that in this case, (X_i, Y_i) are not iid! since they depend on the year.

```
setwd("~/Desktop/travail/cours/M2Imea")

stupid=read.table('Data1.dat', header=TRUE)

cor(stupid$Drownings,stupid$Cage.Mov)

###### Verif par simu de variables dependantes non correlees

n= 20
sigma=0.01
x=seq(-1,1,length.out=n)
y=x^2+rnorm(n,sd=sigma)

plot(x,y)

cor(x,y)
```

Case 1

change σ, correlation will increase with noise

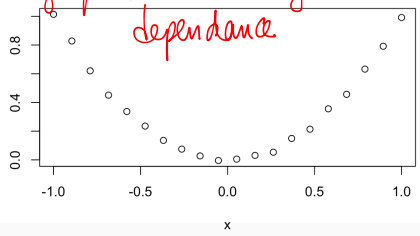
take also uniform variable to have (X_i, Y_i) iid

```
> plot(x,y)
>
> cor(x,y)
[1] -1.704156e-16
> n= 20
> sigma=0.01
> x=seq(-1,1,length.out=n)
> y=x^2+rnorm(n,sd=sigma)
>
> plot(x,y)
>
> cor(x,y)
[1] -0.009980623
>
```

$\sigma = 0$!!

amazingly small correlation

The graph shows clearly the dependance




```
year = 1999:2009
n=length(year)

sigma=0.1
a1= 0.3
a2=100

bruit1= rnorm(n,sd=sigma)
bruit2=rnorm(n,sd=sigma)
bruit1==bruit2

y1= 0.3 * year + bruit1
y2=100* year + bruit2

plot(year,y1,col='red',axes=FALSE,ylab='')
axis(side=4,col='red',ylab='y1')
par(new=TRUE)
plot(year,y2)
```

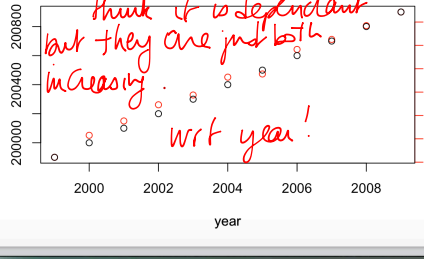
Case 2

the noises ("bruit") are random and !! not equal !!

```
> bruit1==bruit2
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
>
> y1= 0.3 * year + bruit1
> y2=100* year + bruit2
>
> plot(year,y1,col='red',axes=FALSE,ylab='')
> axis(side=4,col='red',ylab='y1')
> par(new=TRUE)
> plot(year,y2)
>
> cor(y1,y2)
[1] 0.9967143
>
```

huge correlation

The graph let us think it is dependant but they are just both increasing wrt year!



3/ Independance

↳ Do you know at least one test of independance?

Chi-square test of independance

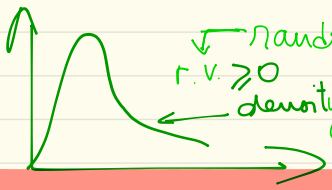
If (X_i, Y_i) iid → independent and identically distributed
Couples $i = 1, \dots, n$ with discrete values $1 \dots r$ for X, $1 \dots s$ for Y

If $X_i \perp\!\!\!\perp Y_i$, then ← independent

N_{jk} = number of couples (X_i, Y_i) with value (j,k)
$N_{j\cdot}$ = number of X_i's with value j
$N_{\cdot k}$ = number of Y_i's with value k

Then $$\sum_{j,k} \frac{\left(N_{jk} - N_{j\cdot} N_{\cdot k}/n\right)^2}{N_{j\cdot} N_{\cdot k}/n} \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2\left((r-1)(s-1)\right)$$

- N_{jk}: observed number
- $N_{j\cdot} N_{\cdot k}/n$: expected number if $\perp\!\!\!\perp$
- $\xrightarrow{\mathcal{L}}$: the cumulative distribution functions converge (c.d.f.)
- $\chi^2((r-1)(s-1))$: chi-square distribution with (here) $(r-1)\times(s-1)$ degrees of freedom.



In practice the approximation is valid as soon as all the expected numbers ≥ 5

```
data=stupid[,-1]
```
← removing the year column.

```
tab1=table(data)
```
← Contingency matrix

```
chi1=chisq.test(tab1) # pval : 15 %
```
Hence independance accepted but look at the warning

```
chi1$expected #
```
not ⩾5 ⟹ warning

```
data2=as.data.frame(cbind(data[,1],(data[,2]>100))) #
tab2=table(data2)
```
2 categories in "drowning"

```
chi2=chisq.test(tab2)
chi2$expected #
```
Still not ⩾5

```
data3 =as.data.frame(cbind(data[,1]>2,(data[,2]>100))) #
tab3=table(data3)
chi3=chisq.test(tab3)
chi3$expected #
```
2 categories for drowning and movie

Still not good → use the Yates correction mentionned by R

Once it is a 2x2 table, one can always compute the exact Fisher test

have a look on internet if you don't know what they are

```
fisher.test(tab3) #
```
Nothing is rejected pval = 1

NB: with large and larger categories, we are losing information

NB2: A test always prefers to accept its null hypothesis in case of doubt ⟹ no real information with large pvalues.

For parabolic data, → do your own category since it is not discrete!

```
n= 100 # increase the size of the simulated data set to ensure n exp ⩾5.
sigma=0.01
x=seq(-1,1,length.out=n)
y=x^2+rnorm(n,sd=sigma)

plot(x,y)
cor(x,y) #
```
data non correlated

```
data.para=as.data.frame(cbind((x>-0.5)+(x>0.5),(y>0.25)))
tab.para=table(data.para)
chi.para=chisq.test(tab.para) #
chi.para #
chi.para$expected #
```

rejection ($\Rightarrow$ dependent variable
Small pval)

$\leftarrow$ nexp ≥ 5 OK $\Rightarrow$ the χ^2 approx can be applied

NB: You can try on case 2 $\Rightarrow$ will say it's dependant

↳ because it is $\perp\!\!\!\perp$ conditionnally to a 3rd variable (year) that you don't observe

NB2: Variable detection can be done if one sees all the variables
If there are hidden variables and if you don't take this into account you may see absurd dependencies

4/ Cause - Effect link

Correlation does not mean cause → effect
To verify a cause - effect link,

→ need at least to have a temporal link: the cause happens b/fe the effect.

III What about modeling?

On the Anpid data, it is not because the independance test accepts that it means anything (A test that accepts, just does because H_0 is its daimy hypothesis)

↳ is there a model where $\hat{r}$ (the estimated correlation) has a meaning?

1/ Linear Regression Model

$$Y_i = a_0 + b_0 X_i + \varepsilon_i$$

with ε_i noise (ie $E(\varepsilon_i | X_i) = 0$)

a_0 and b_0 unknown

To guess a_0 and b_0, one can minimize the l_2 distance

↳ look for a and b minimizing $\sum_i (Y_i - (a + bX_i))^2$

Least-square contrast

↳ How do we do that?

$$\gamma(a,b) = \sum_i (Y_i - (a+bX_i))^2$$

$$= \sum_i \left(Y_i^2 - 2(a+bX_i)Y_i + (a+bX_i)^2\right]$$

$$= \sum_i Y_i^2 - 2a\sum Y_i - 2b\sum_i X_i Y_i + a^2 n + 2ab\sum X_i + b^2 \sum X_i^2$$

$$\frac{\partial}{\partial a} = -2\sum Y_i + 2an + 2b\sum X_i$$

$\Rightarrow a = \overline{Y} - b\overline{X}$ if (a,b) minimise the least square contrast

$$\gamma(\underset{a}{\bar{y}} - b\bar{x}, b)$$

$$= \sum y_i^2 - 2a\sum y_i - 2b\sum x_i y_i + a^2 n + 2ab\sum x_i + b^2 \sum x_i^2$$

$$= \sum y_i^2 - 2n(\bar{y} - b\bar{x})\bar{y} - 2b\,\overline{xy}\,n + (\bar{y} - b\bar{x})^2 n + 2(\bar{y} - b\bar{x})bn\bar{x} + b^2 n\overline{x^2}$$

$$= \sum y_i^2 - 2n\bar{y}^2 + \bar{y}^2 n + b(2n\bar{x}\bar{y} - 2n\,\overline{xy}) + b^2(n\overline{x^2} - n\bar{x}^2)$$

$$\frac{\partial}{\partial b} = 2n(\bar{x}\bar{y} - \overline{xy}) + 2bn(\overline{x^2} - \bar{x}^2)$$

$\Rightarrow$ min (verify monotony etc...)

$$\hat{b} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\mathrm{Cov}_n(x,y)}{\mathrm{var}_n(x)}$$

$$\hat{a} = \bar{y} - b\bar{x}$$

link with Correlation

$$\frac{\text{Var explained by the model}}{\text{Intrinsic variance}} = \frac{\sum(\bar{y} - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

where $\hat{y}_i$ = prediction of the model $= \hat{a} + \hat{b}x_i$

$$= \frac{\sum(\hat{b}\bar{x} - \hat{b}x_i)^2}{\sum(y_i - \bar{y})^2} = \frac{\mathrm{Var}_n x\,\hat{b}^2}{\mathrm{Var}_n y} = \frac{\mathrm{Cov}_n(x,y)^2}{\mathrm{Var}_n x\,\mathrm{Var}_n y} = \left(\hat{r}\right)^2$$

MOREOVER

If $r=0$, then $b=0$

Testing if $r=0$ is equivalent to testing $b=0$.

2/ Gaussian linear regression (the simple case)

To perform a test, we need the distribution of the test statistic under H_0 (or at least an approx)

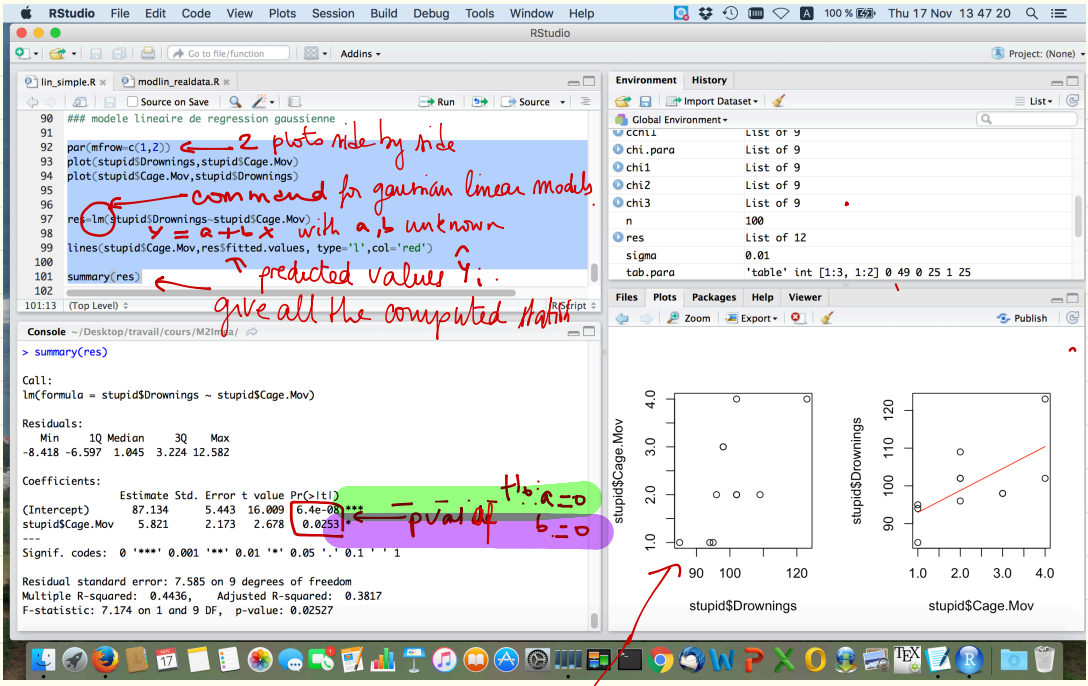
$\Rightarrow$ Here we assume the ε_i to be iid $\mathcal{N}(0, \sigma^2)$

$\Rightarrow$ Theory of Gaussian linear models $\Rightarrow$ derivation of tests

(in particular to see which variables are meaningful)

It's a lot of (big) linear algebra and probability

$\Rightarrow$ for the moment, trust R and see what it does.



Assume gaussian noise in

Cage.mov = a + b Drownings + noise

⇒ no sense Cage.mov takes 4 values !

but Drownings = a + b Cage.mov + gaussian noise

is more likely

NB : Note the pvalue code (*** ...)

Here the test at level 1% accepts, but at level 5% rejects.

of b=0, (i.e r=0)

⇒ There is a link ?!?!?

↳ Maybe it isn't Gaussian

⇒ Shapiro and Wilk test of gaussianity

↳ apply it on residuals

shapiro.test(res$residuals) → pval 26%

... Since Shapiro and Wilk is known to be the most powerful gaussianity test, probably Gaussian approx is not too bad

NB: of course, A number of drownings is an integer and cannot be gaussian, but approx reasonable (remember Binomial / Gaussian approx)

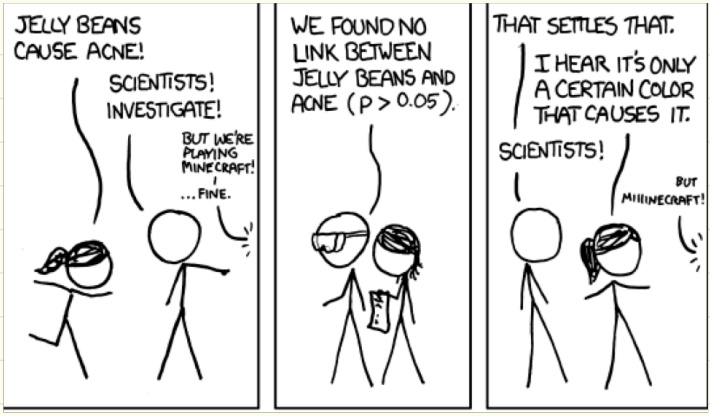
↳ So there is a link between drownings and Cage movies ??

⚠ pval are (usually) uniform under H_0

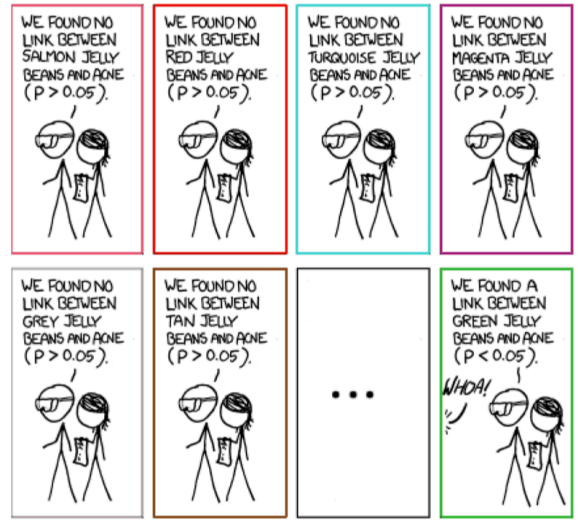
Hence $P_{H_0}(\text{pval} \leq \alpha) = \alpha$

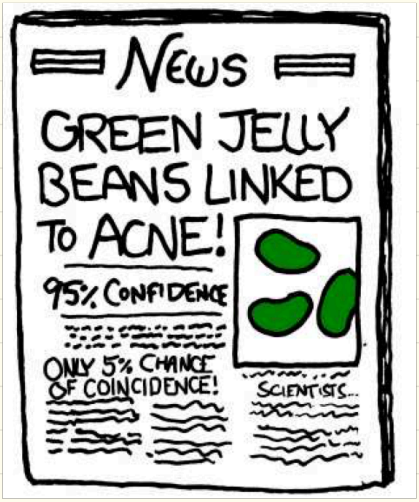
Hence deciding at level 5%, means you have 5% chance to reject whereas you shouldn't

⇒ Big problem if several tests have been performed but not shown!

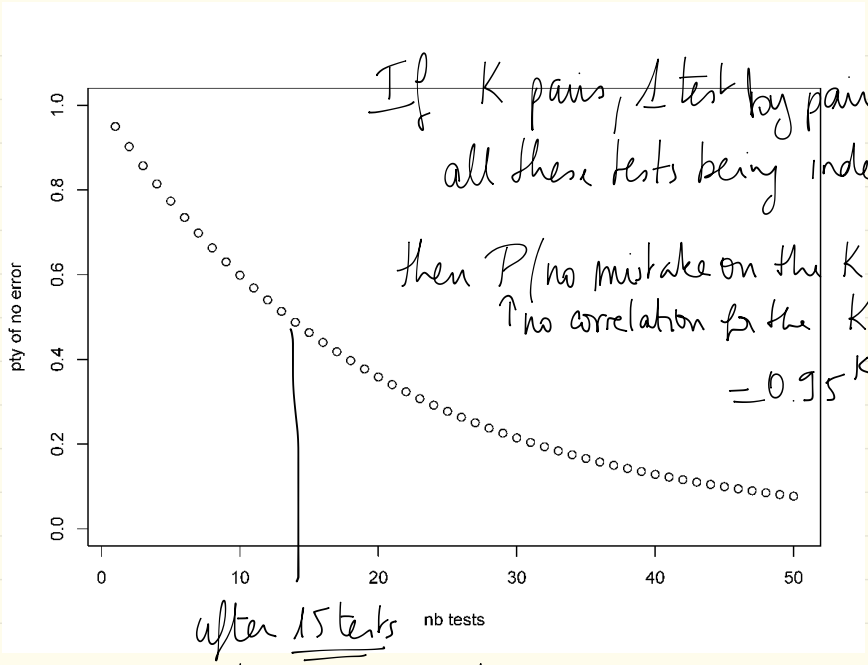


source : xkcd





Vigen probably looked at many stupid pairs before finding (Drownings / Cage movies)



If K pairs, 1 test by pair at level 5%, all these tests being independent,

then P(no mistake on the K tests) ↑ no correlation for the K pairs $= 0.95^K$

after 15 tests

1 chance over 2 to say something wrong

1 possible correction is Bonferroni (→ test at level α/K)

$$\underset{\uparrow \text{no corr}}{P(\text{one mistake})} \leq \sum_{\text{pair}} P(\text{test of this pair makes a mistake}) \quad \uparrow \text{no corr for this pair}$$

$$\leq K \frac{\alpha}{K} \leq \alpha$$

Remember that pval ~2% → with only 3 pairs → accept at level $\frac{\alpha}{K}$

⇒ "Vigen cheated by not saying how many pairs he took at first"

A toy example

HISTORICAL PERSPECTIVE

The basic Gaussian example

For $i = 1, \ldots, n$, we observe the X_i's such that

$$X_i = f_i + \epsilon_i,$$

with ϵ_i i.i.d. $\mathcal{N}(0, \sigma^2)$ and known σ.

Not : X and f corresponding vectors, P_f the distribution of X, $\mathcal{P} = \{P_f, f \in \mathbb{R}^n\}$.

The basic single test problem

Is $f = 0$?

Practical background

Statistical testing $\rightsquigarrow$ answer to practical YES/NO questions.

In practice

"YES" and "NO" are usually NOT equivalent.

NB: Usually "$f \neq 0$" means that the experiment works out. Typically a neuron responds to stimulus, a gene is expressed etc.

The null hypothesis H_0 is an hypothesis that is "never proved or established, but is possibly disproved, in the course of experimentation." (Sir R.A. Fisher, 35 / 20's)

Hence H_0:"$f = 0$" in the toy example.

p-values

Given the observation T_{obs} of a (test) statistics, T, whose distribution is known under H_0, one computes how likely it is to see such a large statistics under H_0

$$p = P_{H_0}(T \geq T_{obs}).$$

A small pvalue indicates H0 is unlikely

Fisher versus Neyman ?

- For Fisher, once p-values are given, their interpretation is left to the practitioner.
- Neyman (and Pearson) (33) = complete mathematical set-up.Alternatives, level and Type I and II errors are defined such that no room is left for interpretation. In certain cases, uniformly most powerful tests exist.

"Professor R. A. Fisher, in opening the discussion, said he had hoped that Dr. Neymans paper would be on a subject with which the author was fully acquainted, and on which he could speak with authority . . . . Since seeing the paper, he had come to the conclusion that Dr. Neyman had been somewhat unwise in his choice of topics. . . ."(discussion of Neyman's paper 1935 JRSSB)

"Dr Pearson said while he knew that there was a widespread belief in Professor Fishers infallibility, he must, in the first place, beg leave to question the wisdom of accusing a fellow-worker of incompetence without, at the same time, showing that he had succeeded in mastering his argument"

"[The unfounded presumptions] would scarcely have been possible without that insulation from all living contact with the natural sciences, which is a disconcerting feature of many mathematical departments" (Fisher, 1955, Phil. Trans. of RSLB)

Chapter 1 : Some basic goodness-of-fit tests

Goodness-of-fit = check if a model is good or not

H_0 : my model is true H_1 : my model is wrong

We you want not too small p values to say my model is plausible w.r.t data

I Chi-square tests

1 Vectorial central limit theorem

$X_1, \dots, X_n$ iid in $\mathbb{R}^d$ with mean $m \in \mathbb{R}^d$ and covariance matrix Σ ($d \times d$ matrix)

Then the CLT says that

$$\frac{X_1 + \dots + X_n - nm}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, \Sigma)$$

2 Multinomial distribution

X_i that can take only d values (different) $\quad p = \begin{pmatrix} P(X_1 = 1) \\ \vdots \\ P(X_1 = d) \end{pmatrix}$

$X_1 \dots X_n$ iid with the same dist.

$$N_j = \text{nb of } X_i / X_i = j = \sum_{i=1}^{n} \mathbb{1}_{X_i = j}$$

$\begin{pmatrix} N_1 \\ \vdots \\ N_d \end{pmatrix}$ is distributed as a Multinomial $\mathcal{M}(n, p)$

($d = 2 \rightarrow N_1 \sim \mathcal{B}(n, p)$)

So if $Y_i = \begin{pmatrix} \mathbb{1}_{X_i = 1} \\ \vdots \\ \mathbb{1}_{X_i = d} \end{pmatrix}$ then $\begin{pmatrix} N_1 \\ \vdots \\ N_d \end{pmatrix} = \sum_{i=1}^{n} Y_i$

$\mathbb{E}(Y_i) = p \quad \Sigma = \mathrm{Cov}(Y_i)$

because

$\Sigma_{jk} = \mathbb{E}\left(\mathbb{1}_{X_i=j}\mathbb{1}_{X_i=k}\right) - \mathbb{E}\left(\mathbb{1}_{X_i=j}\right)\mathbb{E}\left(\mathbb{1}_{X_i=l}\right)$

$\Rightarrow$ limit dist is not nice enough.

$\Rightarrow$

$$U = \begin{pmatrix} \frac{N_1 - np_1}{\sqrt{np_1}} \\ \vdots \\ \frac{N_d - np_d}{\sqrt{np_d}} \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \underline{I} - \underline{\sqrt{p}}\,\underline{\sqrt{p}}^T\right)$$

projection on a subspace of dimension $d-1$

$$\sum \sqrt{np_j}\, U_j = 0$$

Then $\|U\|^2 \xrightarrow{\mathcal{L}} \chi^2(d-1)$

2. Chi-square test of one dist p_0

I observe $X_1 \dots X_n$ iid with value in $\{1, \dots, d\}$ with pty p

I want to test $H_0 : p = p_0$ H_1 : not the case

$X_i \rightsquigarrow N_j$ = nb of X_i with value j

Under H_0, $\|U\|^2 = \sum_{j=1}^{d} \frac{(N_j - np_j^0)^2}{np_j^0} \xrightarrow{\mathcal{L}} \chi^2(d-1)$

I reject H_0 when $\sum_{j=1}^{d} \frac{(N_j - np_j^0)^2}{np_j^0} \geq q_{1-\alpha, d-1}$ (quantile of order $1-\alpha$ with $\chi^2(d-1)$)

obs

exp

Flowers.dat : pink flowers RW red RR white WW

Mendel : if gen 1 : I have only pink flowers

then at gen 2 : 1/4 red 1/2 pink 1/4 white

chisq.test

```
flower=read.table('Flower.dat',header=TRUE)
p=c(1/4,1/2,1/4) # mendel prediction
chi=chisq.test(flower,p=p)#
```

4. Chi square with estimation of parameter

$X_1, \dots X_n$ iid with value in $\{1, \dots J\}$ with pty p

$H_0 : p \in \{p_\theta, \theta \in \Theta\}$ H_1 : this is not the case

$\Theta \subset \mathbb{R}^d$ with $d < J$

↳ estimate θ by $\hat{\theta}$ (MLE)

↳ $$\sum_{j=1}^{J} \frac{\left(N_j - n p_{\hat{\theta}}(j)\right)^2}{n p_{\hat{\theta}}(j)} \xrightarrow[\text{Under } H_0]{\mathcal{L}} \chi^2(J-1-d)$$

NB for X^2 approx it holds as soon as the expected numbers ≥ 5

Calls.dat N_j - number of days with j calls (pty p)

$\Rightarrow H_0$: p is Poisson(λ) λ unknown H_1 : p is not Poisson

$X_1, \dots, X_n$ iid $\mathcal{P}(\lambda)$ $\quad P(X_1 = k) = \frac{\lambda^k}{k!} e^{-\lambda} = p_\lambda(k)$

Log likelihood: $\sum_{i=1}^{n} \log p_\lambda(X_i) = \ell_\lambda$

$\ell_\lambda = \sum_i X_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log X_i !$

$\frac{\partial \ell_\lambda}{\partial \lambda} = \frac{\sum_i X_i}{\lambda} - n \Rightarrow$ it's null in $\hat{\lambda} = \frac{\sum X_i}{n} = \frac{\sum_j j N_j}{n}$ MLE

We should compute:

$$\sum \frac{\left(N_j - n\, p_{\hat{\lambda}}(j)\right)^2}{n\, p_{\hat{\lambda}}(j)}$$

$n\, p_{\hat{\lambda}}(j)$ → should be bigger than 5

this is to have a vector p such that $\sum p_j = 1$

↳ the last category is $[11, +\infty[$

```
calls=read.table('Calls.dat',header=TRUE)

n= sum(calls$Days) # number of observation

Xbar= sum(calls$Calls*calls$Days)/n  ### the MLE is the empirical mean of the X_i's which can be computed with the contingency table

p=dpois(c(0:11),Xbar)  ### estimated probability , be careful for the last number it means [11,+\infty)

p[length(p)]=1-sum(p[1:(length(p)-1)])

nexp=n*p # last class too small -> regroup
```

0 1 2 3 ... 10 $[11, \infty[$ 12 categories

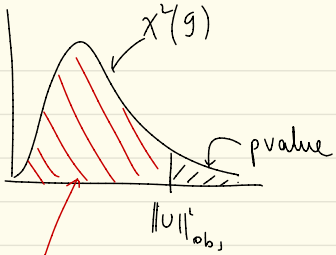
obs: N_0 ... N_{10} $N_{11} = N_{[11,\infty[}$

exp: $n p_{\hat{\lambda}}(0)$ $n p_{\hat{\lambda}}(10)$ $\left[1 - \sum_{j \le 10} p_{\hat{\lambda}}(j)\right] \times n$

obs $N_{10} + N_{11}$

exp $n\left[1 - \sum_{j \le 9} p_{\hat{\lambda}}(j)\right] = n\, p_{\hat{\lambda}}(10) + n\left(1 - \sum_{j \le 10} p_{\hat{\lambda}}(j)\right)$

↳ $\sum_{j=0}^{9} \frac{\left(N_j - n p_{\hat{\lambda}}(j)\right)^2}{n\, p_{\hat{\lambda}}(j)} + \frac{\left(N_{10} - n\,\text{exp}(10)\right)^2}{n\,\text{exp}(10)} \approx \chi^2(11 - 1 - 1)$



this is the cdf of χ^2

```
pbis=c(p[1:(length(p)-2)],p[length(p)-1]+p[length(p)])
# new probability vector
nexp_bis=n*pbis # new expected number

Obs_bis = c(calls$Days[1:(length(p)-2)], calls
$Days[length(p)-1]+calls$Days[length(p)]) # new
observed number

T=sum((Obs_bis-nexp_bis)^2/nexp_bis)

pval=1-pchisq(T,df=(length(pbis)-1-1)) ## large
pvalue
```

$$\text{Pvalue} = 1 - \underbrace{F_{\chi^2(g)}}_{\text{pchisq}}\left(\|U\|^2_{\text{stat}}\right)$$

Poisson dist° is plausible

5. χ^2 test of independance

$Z_i = (X_i, Y_i)$ $\quad X_i$ takes values in $\{1 \dots r\}$, Y_i ——— $\{1, \dots, s\}$

N_{jk} - number of $Z_i = (j,k)$ $\quad N_{j\cdot}$ - number of $X_i = j$ $\quad N_{\cdot k}$ - nb of $Y_i = k$

$$\sum \frac{\left(N_{jk} - \frac{N_{j\cdot} N_{\cdot k}}{n}\right)^2}{\frac{N_{\cdot} N_{\cdot k}}{n}} \xrightarrow[\text{under } H_0 : X \perp\!\!\!\perp Y]{\mathcal{L}} \chi^2\left((r-1)(s-1)\right)$$

(formula from previous course)

→ Reinterpretation

Z_i has $r \times s$ possibilities p $\quad \chi^2$ test of H_0 $p = p_0$ uses $\chi^2(rs-1)$

But here I want to test H_0: p satisfies $X \perp\!\!\!\perp Y$ in $Z = (X,Y)$

ie $p = p_X \otimes p_Y$

I need $r-1$ parameter ← I need $s-1$ parameter

So H_0 can be described by $r-1+s-1$ parameters

How to estimate p under H_0

$$\hat{P}_X(j) = \frac{N_{j\cdot}}{n} \qquad \hat{P}_Y(k) = \frac{N_{\cdot k}}{n} \qquad \Rightarrow \underset{\text{under } H_0}{\hat{P}} = \frac{N_{j\cdot} N_{\cdot k}}{n^2}$$

$$\Rightarrow \sum \frac{(N_{obs} - n_{exp})^2}{n_{exp}} \text{ still true} \qquad n_{exp} = \frac{N_{j\cdot} N_{\cdot k}}{n}$$

For the number of degrees of freedom: $rs - 1 - (r-1+s-1)$

$= rs - r - s + 1 = (r-1)(s-1)$

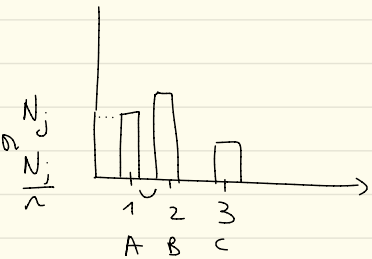
University dat $\longrightarrow$ test $\perp\!\!\!\perp$ between gender and field

Grades dat $\longrightarrow$ test if the grades in both groups are the same (homogeneity test)

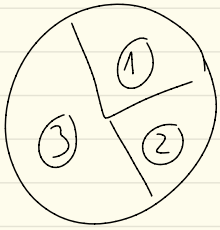
$(=)$ test $\perp\!\!\!\perp$ between grades and groups.

II Graphical representation

1. Qualitative data



bar plot



pie

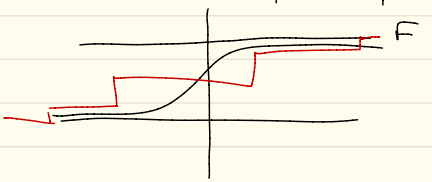
2. For continuous data, cumulative dist° function

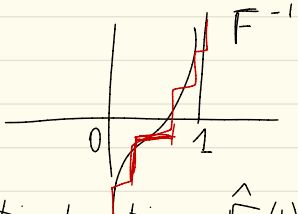
X has value in $\mathbb{R}$ (and even if there is no density or if its values are discrete) then one can characterized its dist° by: $F : t \mapsto P(X \leq t)$

(cumulative distribution function)

$F \nearrow$ $\lim_{-\infty} F = 0$ $\lim_{+\infty} F = 1$

Its inverse is the quantile function $F^{-1}(x) = \inf\{u / F(u) \geq x\}$





Estimate - empirical cumulative distribution function : $\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t}$

when $X_1, \dots, X_n$ iid are observed of cdf F

- unbiased $E(\hat{F}_n(t)) = F(t)$
- consistant - for fixed t $\hat{F}_n(t) \xrightarrow[n \to +\infty]{} F(t)$ (Law of large numbers)

- consistant in the space of all possible F : $\sup_t |\hat{F}_n(t) - F(t)| \xrightarrow[n \to +\infty]{a.s.} 0$.

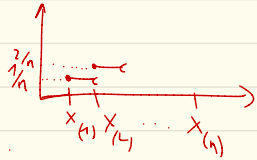
(Glivenko Cantelli theorem)

in R

→
- Simulate n X_i of $\mathcal{N}(0,1)$ iid
- plot the cdf (ecdf)
- compare it to the real cdf
- compare it to cdf of gaussian with different mean or variance
- How would you check visually that a distribution is gaussian without knowing its mean and its variance?
- Look how good or bad this gets with n

```
n=20 # 100
X=rnorm(n)
```

← generates $X_1 \dots X_n$ iid $\mathcal{N}(0,1)$




```
plot.ecdf(X)

abs=seq(-3,3,0.1)

lines(abs,pnorm(abs),col='red')

lines(abs,pnorm(abs,mean=0.5),col='green')
lines(abs,pnorm(abs,sd=0.5),col='blue')

lines(abs,pnorm(abs,mean=mean(X),sd=sd(X)),col='orange')

legend('topleft',c('empirical
cdf','N(0,1)','N(0.5,1)','N(0,0.5)','N(xhat,sdhat)'),col=c('black','red','green','blue','orange'),lty=c(1,1,1,1,1))
```

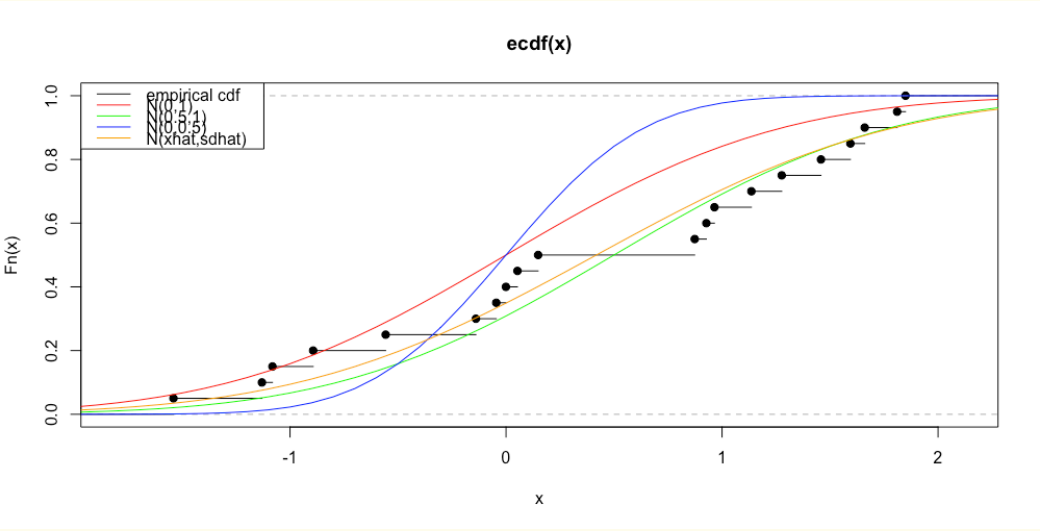
→ compute for $t \in$ abs

$\int_{-\infty}^{t} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx$

if I don't know mean and variance Estimate them

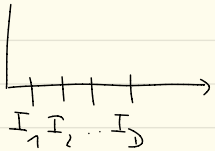
- If n small, one can guess a wrong distr°
- If n large, one can guess the true distr.

- If m and v unknown, the cdf with mean $\hat{m}$ variance $\hat{v}$ will always be closer to the ecdf than the true curve.



2 Density Estimation

- **histogramme** $X_1 \dots X_n$ iid with density f



indicator function

$$\text{length}(I_i) = h$$

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{D} \sum_{j=1}^{n} \mathbb{1}_{X_i \in I_j} \mathbb{1}_{x \in I_j}$$

$$= \sum_{j=1}^{D} \frac{N(I_j)}{nh} \mathbb{1}_{x \in I_j}$$

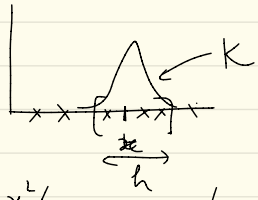
Choice of h Sturges rule $D = f(n)$
(what is implemented by default in hist)

in R (usually) R by default gives $\sum_{j=1}^{D} N(I_j) \mathbb{1}_{x \in I_j}$

which is not the density estimator

↳ hist(data, freq = FALSE)

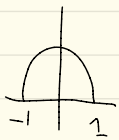
Kernel estimator



$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

$K \to$ gaussian $\frac{e^{-x^2/2}}{\sqrt{2\pi}}$ $\left(\int K = 1\right)$

$\to$ rectangle $\frac{1}{2} \mathbb{1}_{[-1,1]}$ $\to$ Epanechnikov



- density(data) in R is a kernel estimator

$\to$ choice of h (rule of thumb of Silverman)

It is NOT the density of the data

X = some gaussian variables $(n = 200)$ $\mathcal{N}(0,1)$

```
hist(X,freq=FALSE)
rug(X)
lines(density(X),col='cyan')
abs=seq(-3,3,0.1)
lines(abs,dnorm(abs),col='red')
lines(abs,dnorm(abs,mean=mean(X),sd=sd(X)),col='orange')
legend('topleft',c('histogramme','density','N(0,1)','N(xhat,sdhat)'),col=c('black','cyan','red','orange'),lty=c(1,1,1,1))
```

use what happens if you forget this part

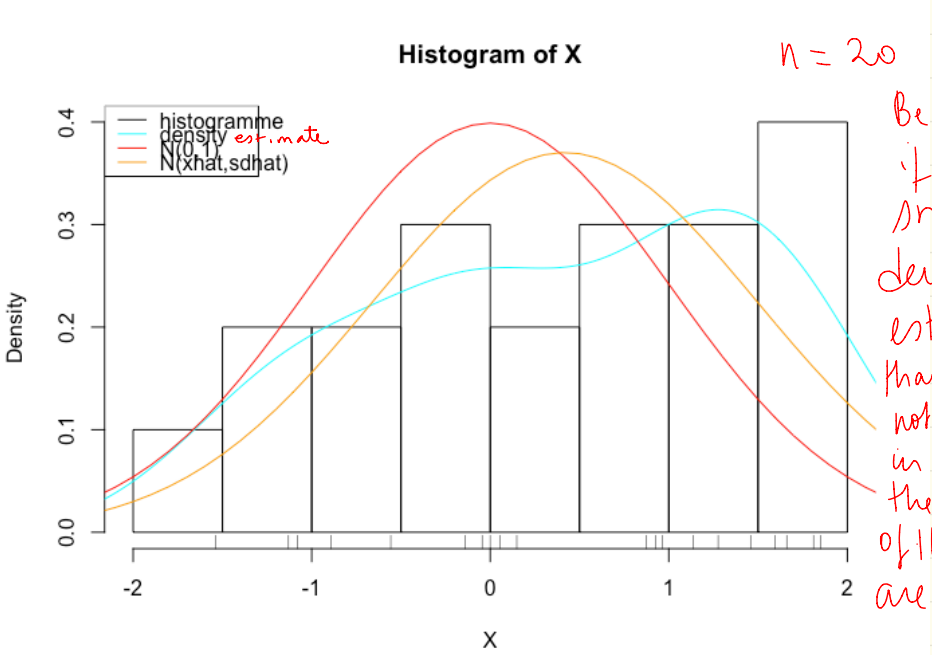
to see the data on your plot

kernel estimate of the density

the true density

estimate the density by $\mathcal{N}(\hat{m}, \hat{v})$ if I know it's gaussian

estimate ↑ true density



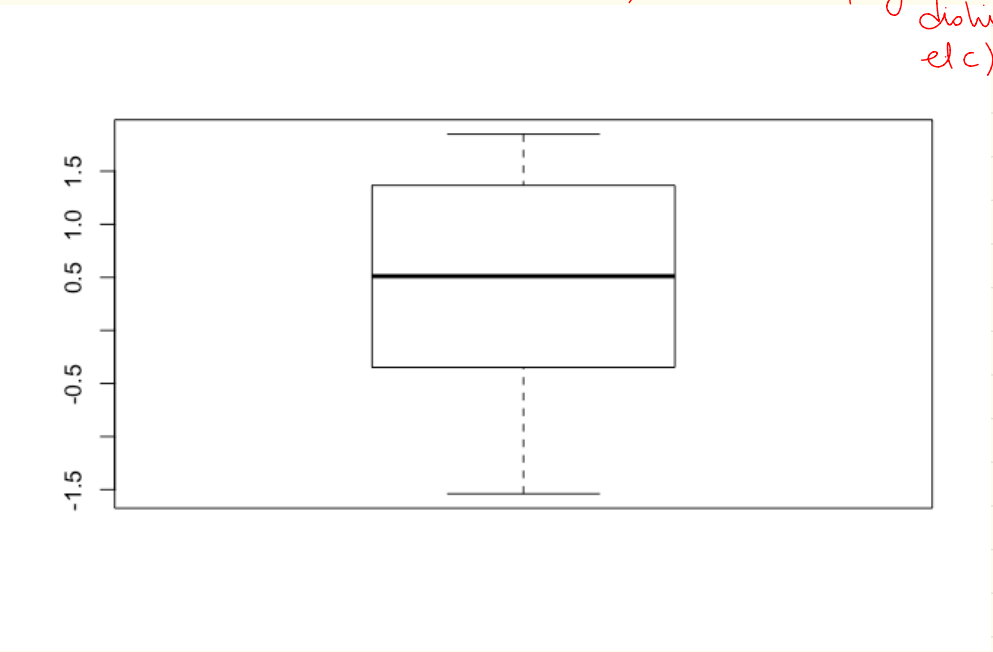
$n = 20$

Be careful if n too small density estimates that do not know in advance the shape of the density are bad.

boxplot (boite a moustache) (n small)

boxplot(X)

(look again at the precise definition of each bar, x (outliers for gaussian distrib) etc)



4) Qq plot

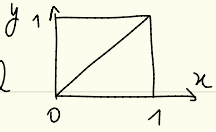
↳ our brain is able to see very quickly if there is a straight line or not ⟹ change the picture to make a line appear

a) qq plot of $X_1 \dots X_n$ against a fixed distribution F_0

One point (x, y) of the qq plot is given by

$$\exists i \;/\; y = F_n^{-1}(t) \qquad x = F_0^{-1}(t)$$

if $X_1 \dots X_n$ are iid with cdf F_0 then the qq plot should be close

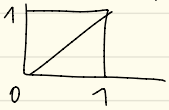


to the diagonal

b) Qq plots between two samples

$X_1 \dots X_n$ iid F_0 $\quad Y_1 \dots Y_m$ iid G_0 question is $F_0 = G_0$?
F_0, G_0 unknown
ecdf of X's $\quad$ ecdf of Y's
draw the qq plot given by $(x,y) \;/\; \exists t,\; x = F_n^{-1}(t),\; y = G_m^{-1}(t)$

If $F_0 = G_0$, you should see something close to



c) Qq plot for gaussian distribution

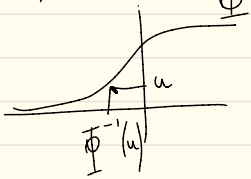
if $X \sim \mathcal{N}(m, \sigma^2)$ then $\frac{X-m}{\sigma} \sim \mathcal{N}(0,1)$
Let Φ be the cdf of $\mathcal{N}(0,1)$

Let's compute the cdf of X

$$P(X \le t) = P\left(\frac{X-m}{\sigma} \le \frac{t-m}{\sigma}\right) = \Phi\left(\frac{t-m}{\sigma}\right)$$

If $F^{-1}(u)$ is the u quantile of X

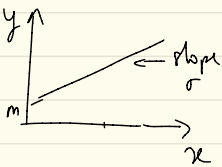
then $F(F^{-1}(u)) = u = \Phi\left(\frac{F^{-1}(u)-m}{\sigma}\right)$



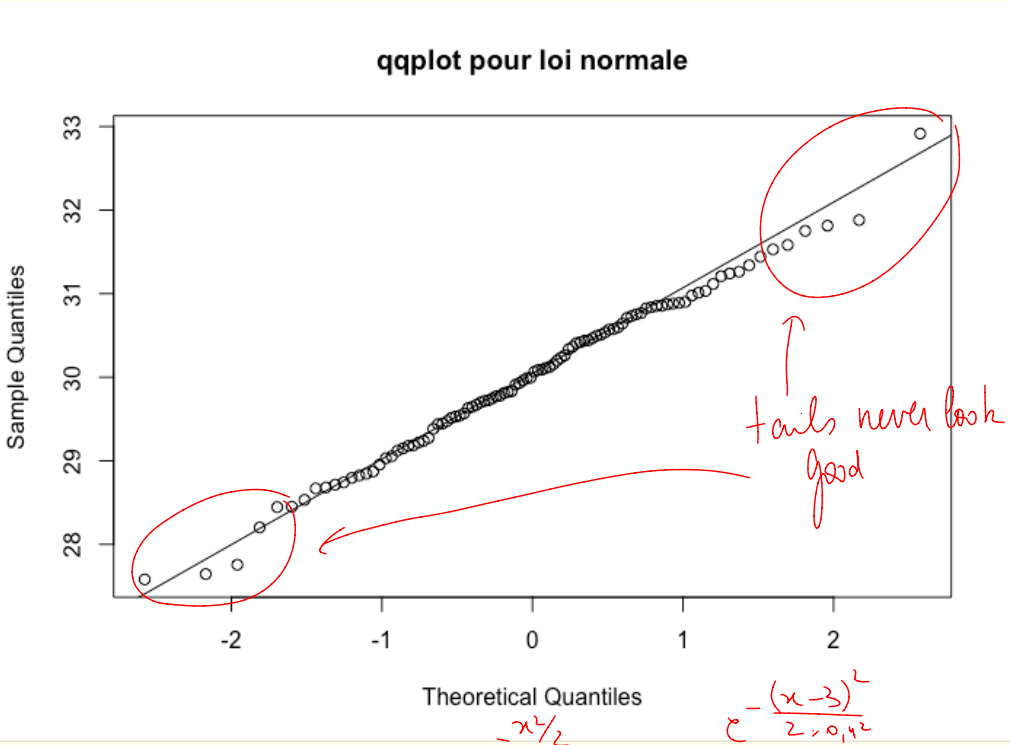
$$F^{-1}(u) = \Phi^{-1}(u) \times \sigma + m \quad \Longleftarrow \quad \Phi^{-1}(u)$$

So if $X_1 \dots X_n$ iid and if they are $\mathcal{N}(m, \sigma^2)$
then the qq plot given by $(x,y) \;/\; \exists t \;/\; y = F_n^{-1}(t),\; x = \Phi^{-1}(t)$

ecdf of X's should look like




```
### qq plot pour la loi normale
X=rnorm(100,mean=30)

qqnorm(X,main='qqplot pour loi normale')
qqline(X)
```



$$\frac{e^{-x^2/2}}{\sqrt{2\pi}} \qquad \frac{e^{-\frac{(x-3)^2}{2\cdot 0{,}4^2}}}{\sqrt{2\pi}\,0.4^2}$$

$$f(x) = 0{,}75 \cdot \mathcal{N}(0,1) + 0{,}25 \cdot \mathcal{N}(3,\ 0.4^2)$$

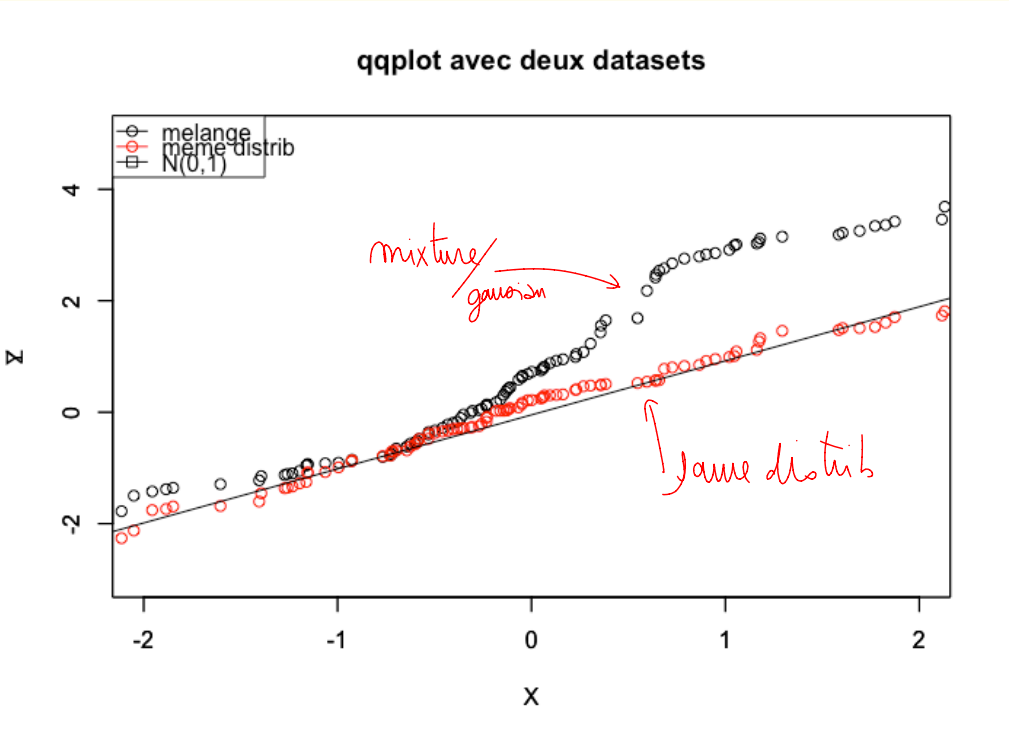
(mixture of gaussian distributions)

} what is this distribution?

```
### qq plot de deux data sets

X=rnorm(100)

nb=rbinom(200,0.25)
Y=c(rnorm(nb,mean=3,sd=0.4),rnorm(200-nb))

Z=rnorm(100)

qqplot(X,Y,xlim=c(-2,2),ylim=c(-3,5),main='qqplot avec deux datasets')
par(new=TRUE)
qqplot(X,Z,xlim=c(-2,2),ylim=c(-3,5), col='red')
qqline(X,distribution=function(p) qnorm(p,mean=0,sd=1))

legend('topleft',c('melange','meme distrib','N(0,1)'),pch=c(1,1,0),lty=c(1,1,1), col=c('black','red','black'))
```

← 2 plots in 1

← draw a line between the 1st quantile (= 0,25) and 3rd quantile



III Kolmogorov-Smirnov tests and its variant

1) KS test for one sample

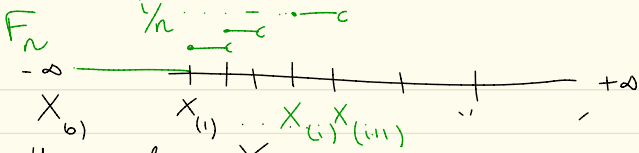
We observe $X_1, \dots, X_n$ iid with cdf F and we want to test $H_0 : F = F_0$ against $H_1 : F \neq F_0$

We will reject if a distance (KS distance) between F_n and F_0 is too large

$$D_n^{F_0} = \sup_{t \in \mathbb{R}} |F_n(t) - F_0(t)|$$

(F_n: ecdf of X's)

We reject when $D_n^{F_0}$ is larger than the $1-\alpha$ quantile of the distrib



of $D_n^{F_0}$ under H_0.

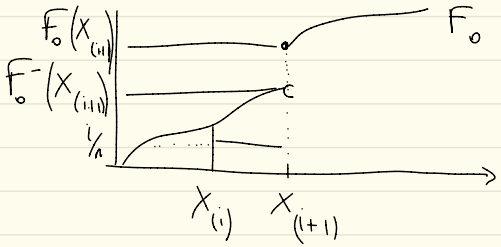
Computation of $D_n^{F_0}$ Let's order the sample $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$

$X_{(0)} = -\infty \quad X_{(n+1)} = +\infty$

$$D_n^{F_0} = \sup_{i=0,\dots,n} \sup_{t \in [X_{(i)}, X_{(i+1)}[} \left| \underbrace{\hat{F}_n(t)}_{\frac{i}{n}} - F_0(t) \right|$$

F_0 is nondecreasing so the sup is either achieved in $X_{(i)}$ or $X_{(i+1)}^-$

$$= \sup_{i=0\dots n} \max \left(\left| \frac{i}{n} - F_0(X_{(i)}) \right| ; \left| \frac{i}{n} - F_0^-(X_{(i+1)}) \right| \right)$$



$2n$ values to compute at most

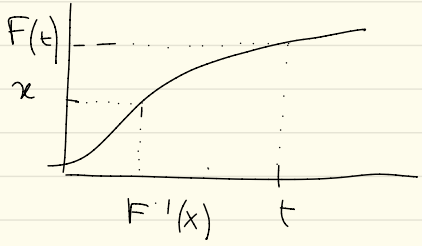
↳ take the max

↳ one can compute it

The distribution of $D_n^{F_0}$ does not depend on F_0 if F_0 is continuous (and we H_0)

Why?

$$F^{-1}(x) \leq t \iff x \leq F(t)$$



$$D_n^{F_0} = \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t} - F_0(t) \right|$$

But one can simulate X_i by taking $U_i \sim \mathcal{U}(0,1)$ iid and taking $X_i = F_0^{-1}(U_i)$

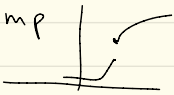
$D_n^{F_0}$ has the same distribution as

$$\sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{F_0^{-1}(U_i) \leq t} - F_0(t) \right|$$

$$= \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{U_i \leq F_0(t)} - F_0(t) \right| = \sup_{x \in F_0(\mathbb{R})} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{U_i \leq x} - x \right|$$

If F_0 is continuous $F_0(\mathbb{R}) = [0,1]$

NB: $F_0(\mathbb{R}) \subset [0,1]$ and if it is not = it means that F_0 has a jump



If F_0 is C^0, $D_n^{F_0} \sim D_n^{\mathcal{U}[0,1]}$

in R: ks.test (be careful with the warnings in case of F_0 not continuous)

Informally $F_n \xrightarrow[n \to +\infty]{} F_0$ under H_0

Vt fixed $\sqrt{n}(F_n(t) - F_0(t)) \longrightarrow \mathcal{N}(0, F_0(t)(1 - F_0(t)))$

$\rightsquigarrow$ when $n \to +\infty$ $\sqrt{n} D_n^{F_0}$ converges towards a limit dist°

Kolmogorov (33) shows that $P(\sqrt{n} D_n \geq x) \xrightarrow{n \to +\infty} 2 \sum_{k \geq 1} (-1)^{k+1} e^{-2k^2x^2}$

Smirnov (41) ——— $P(\sqrt{n} D_n^+ \geq x) \xrightarrow{n \to +\infty} e^{-2x^2}$ (Gaussian tail)

$D_n^+ = \sup_{t \in \mathbb{R}} \left[F_n(t) - F_0(t) \right]$ (one sided KS)

↳ to test $H_0: \forall t \; F(t) \leq F_0(t)$ $H_1: \exists t \; F(t) > F_0(t)$

if $\forall t \; F(t) \leq F_0(t) \iff \forall x \; F^{-1}(t) \geq F_0^{-1}(t)$

So if I want to simulate $X \sim F$ and $Y \sim F_0$ with the same $U \sim \mathcal{U}(0,1)$

(think life times) $X = F^{-1}(U) \geq Y = F_0^{-1}(U)$

Under H_1 (two-sided KS test)

$\exists t \,/\, F(t)$ (real cdf of X's) $\neq F_0(t)$

$$\left(D_n^{F_0} = \sup_{t \in \mathbb{R}} \left| F_n(t) - F_0(t) \right| \right) + \underbrace{\sup_{t \in \mathbb{R}} \left| F_n(t) - F(t) \right|}_{\xrightarrow[n \to +\infty]{as} 0}$$

$$\geqslant \sup_{t \in \mathbb{R}} \left| F_0(t) - F(t) \right| > 0$$

$$\geqslant \eta > 0$$

So for n large enough, $D_n^{F_0} \geqslant \frac{\eta}{2}$ as

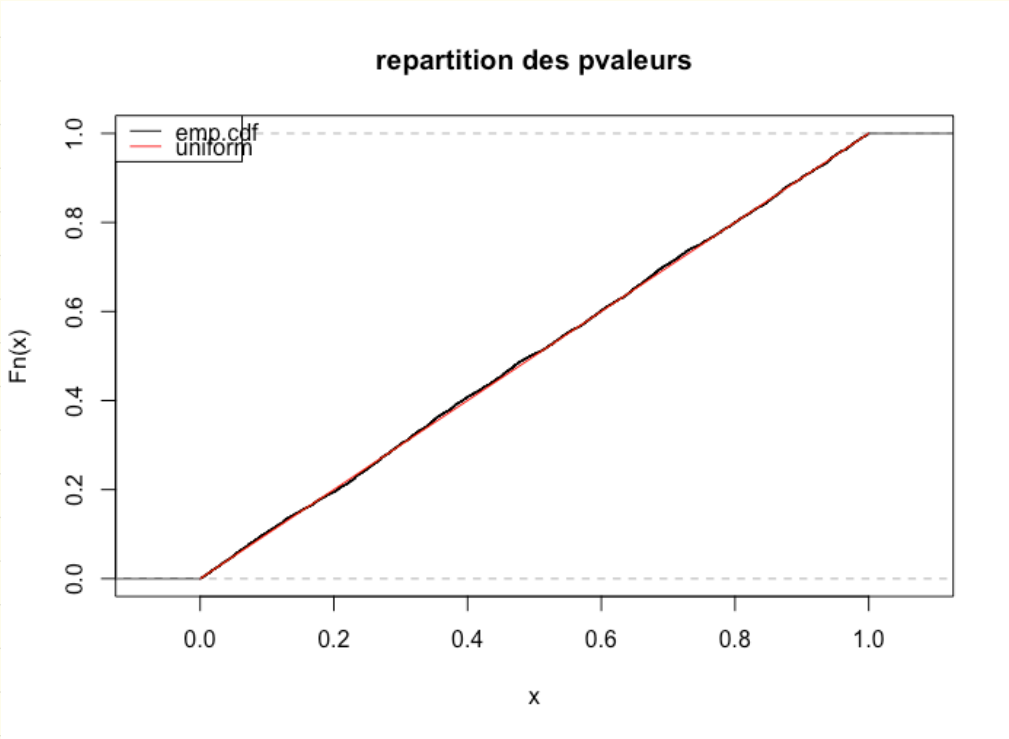
On the other hand one rejects if $D_n^{F_0} \geqslant q_{1-\alpha}^n$ (of the order $\frac{1}{\sqrt{n}}$ by Kolmogorov result)

So for n large enough, as $D_n^{F_0} \geqslant q_{1-\alpha}^n$ and the test rejects with probability $\to 1$

(The power of KS tends to 1)

Check that the pvalues of ks test in $\mathbb{R}$ are unif on $[0,1]$ (under H_0)

```
#

Nsimu=5000
n=30 #

pval_cont=rep(0,Nsimu) #


for(i in 1:Nsimu)
{
  X=rnorm(n) #

  a= ks.test(X,'pnorm')

  pval_cont[i]=a$p.value

}

plot.ecdf(pval_cont,main='repartition des pvaleurs')
lines(c(0,1),c(0,1),col='red')
legend('topleft',c('emp.cdf','uniform'),col=c('black','red'),lty=c(1,1))
```



```
alpha=0.05 #

niveau_emp=mean(pval_cont<alpha) #

#

#

erreur_adm= sqrt(0.05*0.95/Nsimu)*qnorm(0.975)

abs(alpha-niveau_emp)>erreur_adm #

#

ks.test(pval_cont,'punif')
```

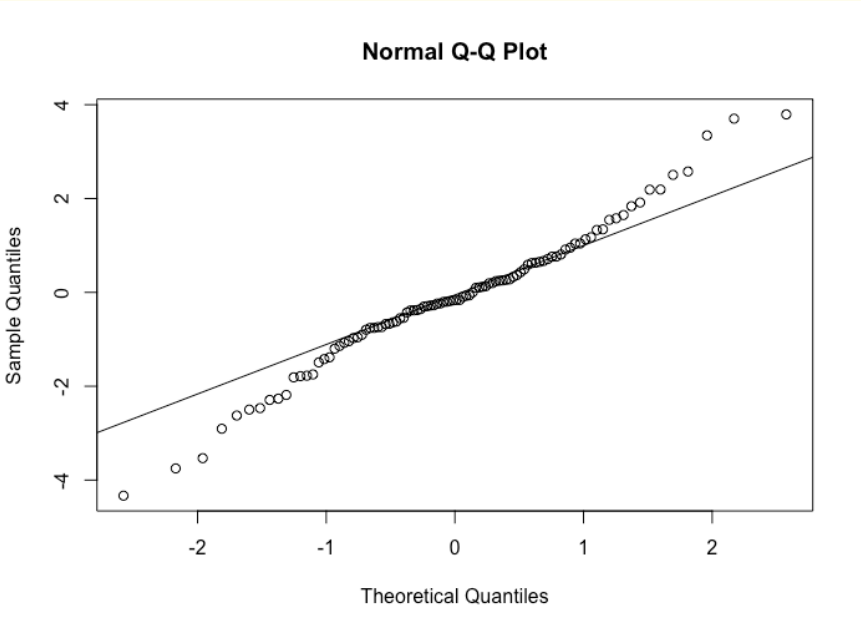
To check that test at level 5%

$\sqrt{\text{variance}}\, z_{1-\alpha}$

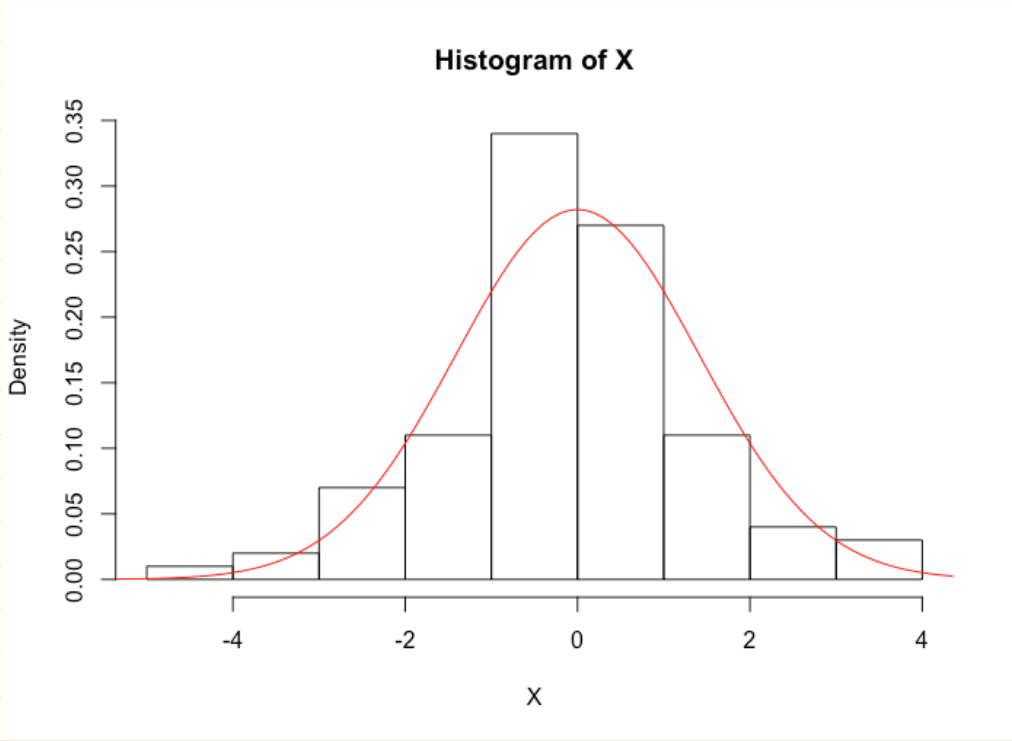
← accept or not that

```
##      H1

n=100 #

X=(2*rbinom(n,1,0.5)-1)*rexp(n,1)  #                 N(0,2) mais c'est p

qqnorm(X)
qqline(X)
```

Laplace



```
hist(X,freq=FALSE)
abs=seq(-6,6,0.1)
lines(abs,dnorm(abs, sd=sqrt(2)),col='red')
```




```
ks.test(X,'pnorm',sd=sqrt(2))


for(i in 1:Nsimu)
{
  X=(2*rbinom(n,1,0.5)-1)*rexp(n,1) #

  a= ks.test(X,'pnorm',sd=sqrt(2))

  pval_cont[i]=a$p.value

}

plot.ecdf(pval_cont,main='repartition des pvaleurs sous H1')
lines(c(0,1),c(0,1),col='red')
```

Anderson Darling (n) $\int \frac{(F_n(t) - F_0(t))^2}{F_0(t)(1 - F_0(t))} dF_0(t)$

3. KS with two samples

$X_1 \dots X_n$ iid $\sim F$ $\quad Y_1 \dots Y_m$ iid $\sim G$

$H_0 : F = G$ against $H_1 : F \neq G$

$$D_{n,m} = \sup_{t \in \mathbb{R}} |F_n(t) - G_m(t)|$$

test statistic

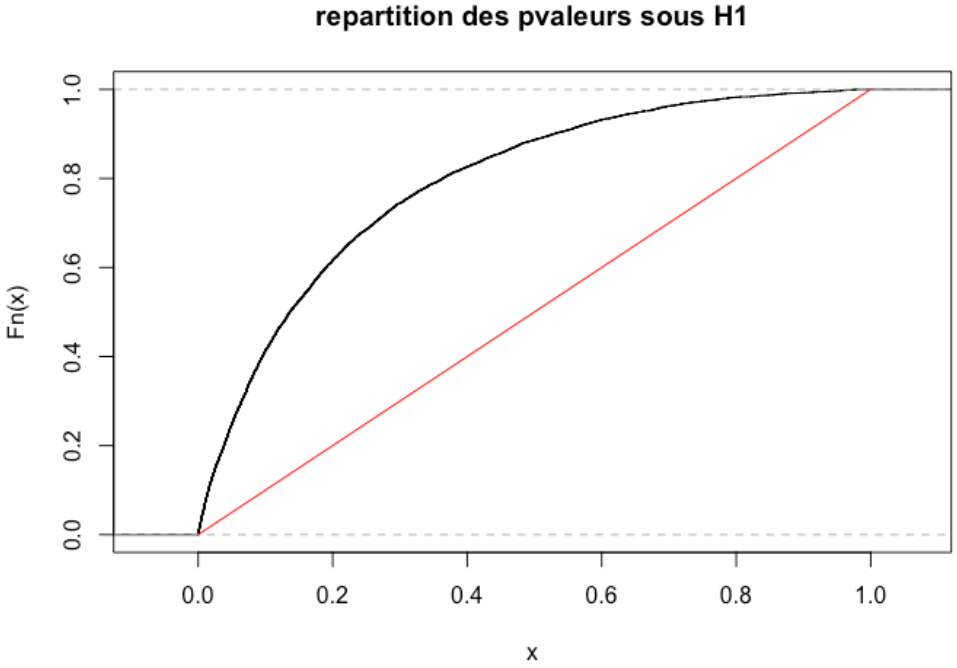
F_n: ecdf of X's, G_m: ecdf of Y's

$\Rightarrow$ if $F = G$ is continuous then the dist° of $D_{n,m}$ is independent of the precise value of F and G

So one can reject H_0 when $D_{n,m} > q_{1-\alpha}^{KS,n,m}$ (quantile of order $1-\alpha$ of this distribution)

$\rightarrow$ try it in R `ks.test(X,Y)`

$\Rightarrow$ if F and G are not c° $\rightsquigarrow$ X and Y are discrete $\Rightarrow$ χ^2-test



2. Variante

Cramer von Mises Test statistics : $(n) \int \left(\hat{F}_n(t) - F_0(t)\right)^2 dF_0(t)$

IV Rank tests of Wilcoxon (1945)

1. Signed rank test

Let $X_1 \dots X_n$ be iid rv, one wants to test

H_0: The distribution of X is symmetric around 0
against H_1: this is not the case

Idea: if X symmetric then X has the same distribution as $\text{sign}\, X \cdot |X|$ (Rademacher)

ε is Rademacher if $\varepsilon \in \{-1, 1\}$ and $P(\varepsilon = 1) = 1/2 = P(\varepsilon = -1)$

$\Rightarrow$ To make the distribution of the test statistic free of the underlying distrib of $|X|$, one uses the rank

$|X_1| \dots |X_n| \longrightarrow$ rank then R_i is the position of $|X_i|$ in the list

If $|X_i| \neq 0$ as then $\boxed{W = \sum_{i=1}^{n} \text{sign}(X_i)\, R_i}$

$\Rightarrow$ Wilcox test (in R)

Main Interest

for each trial i two measures (X_i, Y_i)
$\hookrightarrow Z_i = X_i - Y_i$ should be symmetric around 0 if X and Y have the same distribution

If one tests that H_0: the distribution of Z is symmetric by the Wilcoxon signed rank test

$\Leftrightarrow$ testing that there is as many X_i's below their Y_i's than Y_i's ——— X_i's

↳ wilcox.test(X, Y, alternative: "greater", paired=true)

↑ X bigger than Y under H_1

(X_i, Y_i)

2- Rank sum test (Mann Whitney)

We observe $X_1 \dots X_n$ and $Y_1 \dots Y_n$ (not paired)

We assume that $X \perp\!\!\!\perp Y$ $\quad H_0$: X is not greater than Y vs H_1: X is greater than Y

Under H_0: $P(X>Y)+0.5P(X=Y) = P(Y<X)+0.5P(Y=X)$

H_1: $P(X>Y)+0.5P(X=Y) > 0.5$

$\Rightarrow$ Test Statistic $U = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\mathbb{1}_{Y_j<X_i} + 0.5\,\mathbb{1}_{Y_j=X_i}\right]$

reject when U's too large

↳ wilcox.test(X, Y, alternative=greater)

III Gaussianity tests (cf nortest package)

Observe $X_1 \dots X_n$ iid with cdf F. We denote Φ_{m,σ^2} cdf of $\mathcal{N}(m, \sigma^2)$

$\Rightarrow$ We want to test $H_0 : \left(\exists m, \sigma^2,\ F = \Phi_{m,\sigma^2}\right)$ vs H_1 : this is not the case

1. Lilliefors test

We compute

$$T_n = \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t} - \Phi_{\hat{m},\hat{\sigma}^2}(t) \right|$$

$$= \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq \hat{m} + u\hat{\sigma}} - \Phi_{0,1}(u) \right|$$

$$= \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \leq u} - \Phi_{0,1}(u) \right| \quad \text{with } Y_i = \frac{X_i - \hat{m}}{\hat{\sigma}}$$

The Y_i's are not iid but the dist° of $(Y_1 \dots Y_n)$ does not depend on m and σ

↳ one can compute quantiles for T_n without knowing m and σ

$\Rightarrow$ reject when T_n is too large / quantiles of this distribution under H_0

lillie.test (in the nortest package)

2. Shapiro and Wilk

↳ shapiro.test (the most powerful in general)