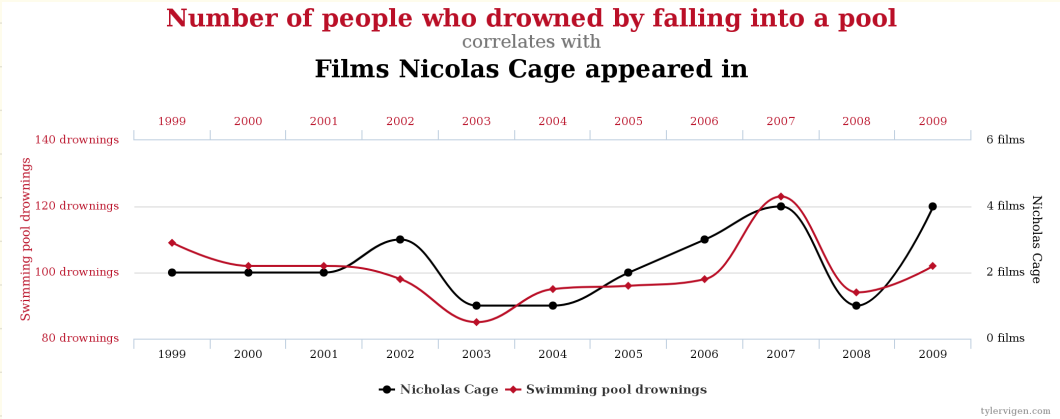
Cours de Détection de Variables et tests multiples

Patricia Reynaud-Bouret

Année 2016-2017

Mise en jambe



Qu'en pensez-vous ? (cf site http://tylervigen.com/spurious-correlations)

I Corrélation

$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$ où $Cov(X,Y) = E(XY) - E(X)E(Y)$

$\sigma_X = \sqrt{Var X}$, $\sigma_Y = \sqrt{Var Y}$

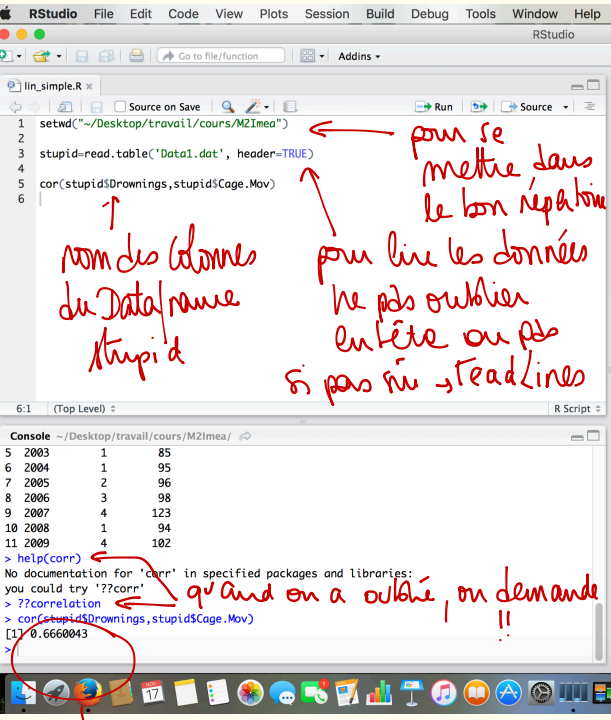
Version empirique

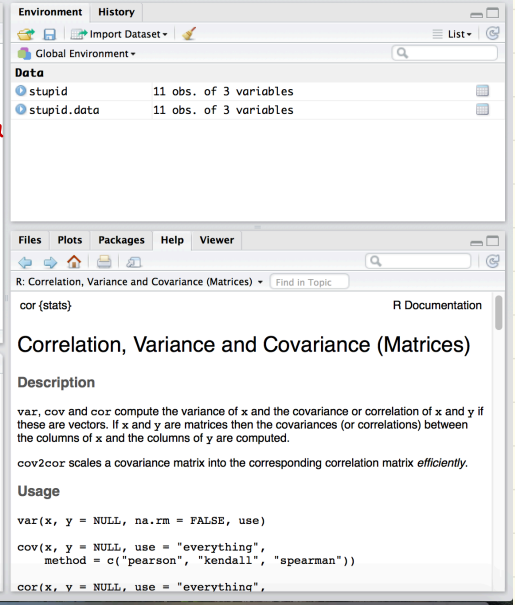
Si on a $(X_1,Y_1), \dots, (X_n,Y_n)$ iid on peut estimer r par

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

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

La "sagesse populaire" va dire qqch comme si $|\hat{r}| > 0.5$ dépendance forte entre les variables.





Une corrélation de 0.66 → fort ça doit bien dépendre l'un de l'autre!

A ce stade, vous devez vous dire "On est en train de faire n'importe quoi, il ne peut pas y avoir de liens entre les noyades et les films de Cage."

NB spurious = fallacieux

Quelles explications voyez vous?? Réponses possibles:

- Les bases de la science populaire ne veulent rien dire
- C'est qui une bonne base, que veut-on en faire?
- Il (Tyler Vigen) l'a cherché!! / a triché ??

↳ On va décortiquer les 3 points!!

II Corrélation, indépendance et lien de cause à effet

1/ Des trucs manifestement dépendants et qui ne se corrèlent pas

2/ Des trucs manifestement complètement "indépendants" et qui se corrèlent

Chercher ..... Comment simuler qqch comme ça ?

Pour 1/ Si on est indépendant (au vrai sens math)

Alors $E(XY) = E(X)E(Y)$ et donc $r = 0$, si bcp de données $\hat{r} \approx 0$

Mais on peut très bien imaginé des trucs dépendants qui peuvent vérifier $r = 0$ (juste premier moment)

Ex 

+ bruit eventuel → on vérifie par simu.

Pour 2/ Il suffit d'être lié à une 3e variable (ex année)

Cf toutes les grandes corrélation ↗ en fct de l'année

Ex dépense en sciences / suicides par pendaison

NB: A ce moment là, $\hat{r}$ pour estimer est en fait assez bof car pas iid. Conditionnellement à l'année.

Cas 1/

lin_simple.R — modlin_realdata.R

```r
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)
```

← aucune – vous à faire varier sigma la corrélation va augmenter avec le bruit!!

Environment — Global Environment

Data	
stupid	11 obs. of 3 variables
stupid.data	11 obs. of 3 variables

Values	
n	20
sigma	0.01
x	num [1:20] -1 -0.895 -0.789 -0.684 -0.579 ...
y	num [1:20] 1.014 0.828 0.62 0.451 0.337 ...

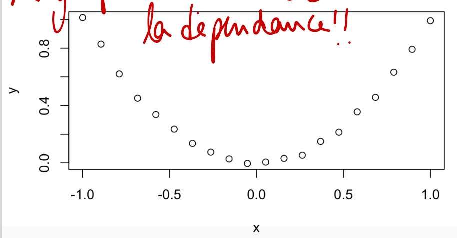
Console ~/Desktop/travail/cours/M2Imea/

```
> 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
>
```

← qd $\sigma = 0$!!

Corrélation ridicule!!

Le graphe montre clairement la dépendance!!



Cas 2

lin_simple.R* — modlin_realdata.R

```r
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)
```

← tirages aléatoires!!
← a n'est pas les mêmes!

Environment — Global Environment

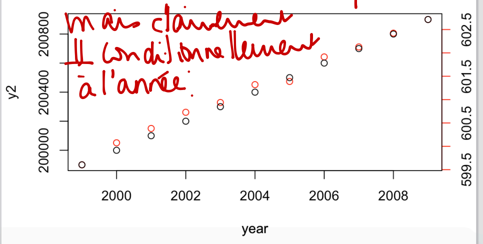
bruit1	num [1:11] -0.096 0.0756 0.0825 0.1285 0.0379 ...
bruit2	num [1:11] 0.0258 -0.1175 0.014 0.0153 0.025 ...
n	11L
sigma	0.1
x	num [1:20] -1 -0.895 -0.789 -0.684 -0.579 ...
y	num [1:20] 1.014 0.828 0.62 0.451 0.337 ...
y1	num [1:11] 600 600 600 601 601 ...
y2	num [1:11] 2e+05 2e+05 2e+05 2e+05 2e+05 ...
year	int [1:11] 1999 2000 2001 2002 2003 2004 2005 2006 …

Console ~/Desktop/travail/cours/M2Imea/

```
> 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
>
```

← Corrélation énorme!!

On croit à une dépendance mais clairement ⊥ conditionnellement à l'année.



3/ L'indépendance

Sans rentrer dans tous les tests possibles vous devez en connaître au moins un ??

↳ Chi-deux d'indépendance

(X_i, Y_i) iid $i = 1..n$ et catégoriel valeur de $1...r$ pour X $1...s$ pour Y.

Alors si les $X_i \perp\!\!\!\perp Y_i$,

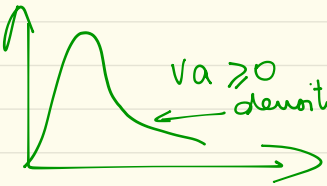
N_{jk} nb de couples qui valent (j,k) (ie $X_i = j, Y_i = k$)
$N_{j\cdot}$ nb de couples tq $X_i = j$
$N_{\cdot k}$ ——— tq $Y_i = k$

les f° de rép cv

Alors $$\sum_{j,k} \frac{\left(N_{jk} - \frac{N_{j\cdot} N_{\cdot k}}{n}\right)^2}{\frac{N_{j\cdot} N_{\cdot k}}{n}} \xrightarrow{\mathcal{L}} \chi^2\left((r-1)(s-1)\right)$$

nb observé — nb attendu si indep...

loi du χ^2 à (ici) $(r-1)\times(s-1)$ degré de liberté



Pour que approx valable il faut que tous les nb attendus ≥ 5.

Comme c'est pour des va catégorielles comment fait on ici?

→ Cage mov OK (a priori)
→ Drownings (on pourrait mais on va avoir du mal pour $N_{att} \geq 5$ → regrouper)

```
data=stupid[,-1]
```
← j'enlève la colonne année

```
tab1=table(data)   #matrice de contingence

chi1=chisq.test(tab1) # pval : 15 % (donc Indep acceptee) mais regardez le warning

chi1$expected # pas bon, pas >=5
```
(ça explique le warning)

```
data2=as.data.frame(cbind(data[,1],(data[,2]>100))) # j'ai fait deux categories sur les drownings
tab2=table(data2)

chi2=chisq.test(tab2)

chi2$expected # toujours pas bon

data3 =as.data.frame(cbind(data[,1]>2,(data[,2]>100))) # deux categories sur les deux valeurs
```
Donc il faut

```
tab3=table(data3)

chi3=chisq.test(tab3)

chi3$expected # toujours pas bon (allez voir ce que c'est que la Yates correction dont R parle)

# en fait quand on arrive a des tables 2 *2 on peut faire le test exact de fisher

fisher.test(tab3) # et la on ne rejette rien du tout,
```
(Allez voir ce que c'est si vous ne savez pas....)

pval 1 ...

⚠ A force de mettre des catégories de plus en plus grosses on perd de l'information

!! Un test est fait pour accepter son hypothèse nulle (ici X ⫫ et Y), il ne va donner des petites pvaleurs qu'à contre cœur... Donc pas d'info si pval grande !!

Pour les données parabol → pas catégorielles → on fait les catégories nous même

```
n= 100 # j'augmente un peu pour pas etre genee par les >=5
sigma=0.01
x=seq(-1,1,length.out=n)
y=x^2+rnorm(n,sd=sigma)

plot(x,y)

cor(x,y) # correlation ridicule, les varaibles sont pas correleees.

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) # pas de message d'erreur
chi.para # pvaleur clairement ridicule , on rejette l'independance, les varaibles sont dependantes
chi.para$expected # tout va bien pour l'approximation
```

Essayez tout seul sur les var qui croissent linéairement en fct de l'année → ça dira dépendance

Car y1 et y2 sont $\perp\!\!\!\perp$ conditionnellement à l'année mais pas inconditionnellement.

⇒ Le problème de la détection de variables est que c'est possible si on observe bien toutes les variables
Mais si des variables sont cachées on va trouver des liens absurdes → toujours faire attention avec les analyses.

4/ Cause à effet

Rien n'est plus trompeur que de confondre corrélation et lien de cause à effet.

Cause à effet : - au moins lien temporel : cause avant effet (et pas juste simultané)
- au mieux : une fois une dépendance remarquée voir si on peut faire l'expérience de ne manipuler qu'une variable et réaliser effet

III et la modélisation dans tout ça ?

Sur les stupid data, c'est pas parce que le test d'H accepte que ça dit quoique ce soit (un test qui accepte ne dit rien en fait)

↳ Y aurait-il un modèle statistique dans lequel les bonnes douceurs sur $\hat{r}$ prendrait un sens ?

1/ Le modèle de régression linéaire

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

avec ε_i bruit (ie $E(\varepsilon_i | X_i) = 0$)

a_0 et b_0 inconnu.

Pour deviner a_0 et b_0, on peut minimiser la distance ℓ^2

↳ on cherche a et b qui minimise $\sum_i (Y_i - (a + bX_i))^2$

↳ Comment fait-on ??

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

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

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

$$\frac{\partial}{\partial a} = -2 \sum Y_i + 2an + 2b \sum X_i \Rightarrow a = \overline{Y} - b\,\overline{X}$$

min en

$$\gamma(\underset{a_1}{\bar{y} - b\bar{x}}, b)$$

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

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

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

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

$\Rightarrow$ min (vérifier monotonie / chgt de signe dans le bon sens)

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

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

Lien avec la corrélation

$$\frac{\text{Var expliqué par le modèle}}{\text{Var intrinsèque}} = \frac{\sum(\bar{y} - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

où $\hat{y}_i$ = les préd du modèle $= \hat{a} + \hat{b}x_i$

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

Donc $\hat{r}$ est proportionnel à $\hat{b}$.

⇒ faire un test de la nullité de $\hat{b}$ ⟺ tester nullité de $\hat{r}$ et donc la corrélation éventuelle entre X et Y

2/ Le modèle linéaire gaussien (cas de la régression simple)

Comme on l'a vu avant, tester demande de connaître la loi (asymptotique éventuellement) de la stat. de test (ici $\hat{b}$)

↪ dépend forcément du bruit sous-jacent ε_i.

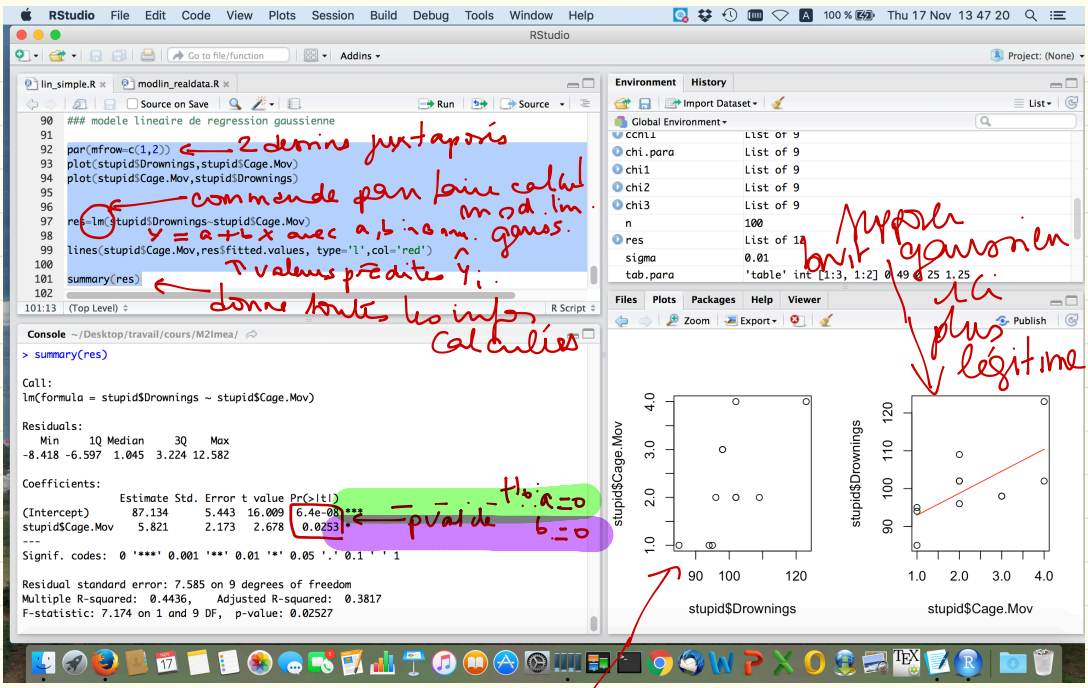
Si on suppose ε_i gaussien iid $\mathcal{N}(0, \sigma^2)$ σ^2 éventuellement inconnu

On peut alors connaître la loi de $\hat{b}$ / b → test de $b=0$

NB : La plus grosse partie théorique des modèles linéaires gaussiens consiste à savoir estimer les param. (ici b) et à donner leur loi !!

(c'est beaucoup de proba. et d'algèbre linéaire mais dans un seul but : arriver à interpréter ce qui se passe sur les vraies données.

Pour le moment on va faire comme si vous saviez tout pour voir comment ça s'applique



commande pour faire calcul mod. lin.

2 dessins juxtaposés

$y = a + bx$ avec a, b inconnus.

$\hat{y}_i$ valeurs prédites

donne toutes les infos calculées

pval de $b=0$

$H_0 : b=0$

supposer bruit gaussien ici plus légitime

supposer bruit gaussien

dans Cage.mov = a + b Drownings + bruit

$\Rightarrow$ idiot Cage.mov prend 4 valeurs!

mais Drownings = a + b Cage.mov + bruit gaussien

est plus raisonnable

Retenir Code des pval (***.. etc)

ici le test de $b=0$ fait automatiquement pval : 0.02

le test accepte au niveau 1%

rejette au niveau 5% $\rightarrow$ le lien semble avéré ??

Oui mais

1/ et si pas Gaussien ??

test de shapiro et Wilk (on reparlera...)
(test le + puissant pour vérifier normalité !)
shapiro.test(res$residuals) ⟶ pval 26% semble Gaussien

⚠ semble... (bien sûr qu'un vrai de noyades pas gaussien c'est dis avec mais approx semble ok
cf approx Binomiale / gaussien)

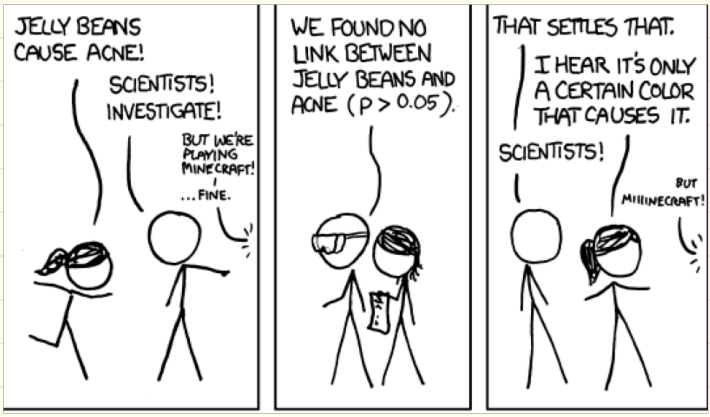
2/ Donc y a un lien !!?!!

⚠ A l'interpretation d'une pvaleur en particulier en cas de tests multiples non dits!

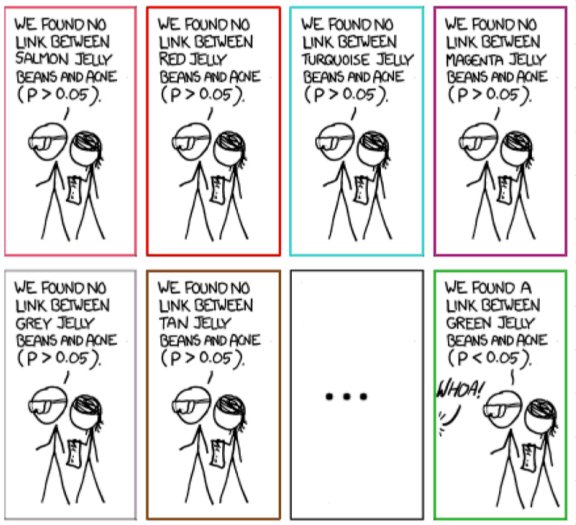
Prop pval sous H_0 $P(\text{pval} \leq \alpha) \simeq \alpha$ (on y reviendra)

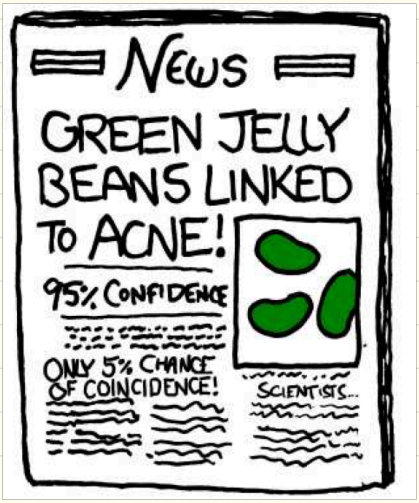
si $\alpha = 5\%$ (test au niveau 5%)

pval $\leq 5\%$ a 5% de chance de se produire !!



source : xkcd

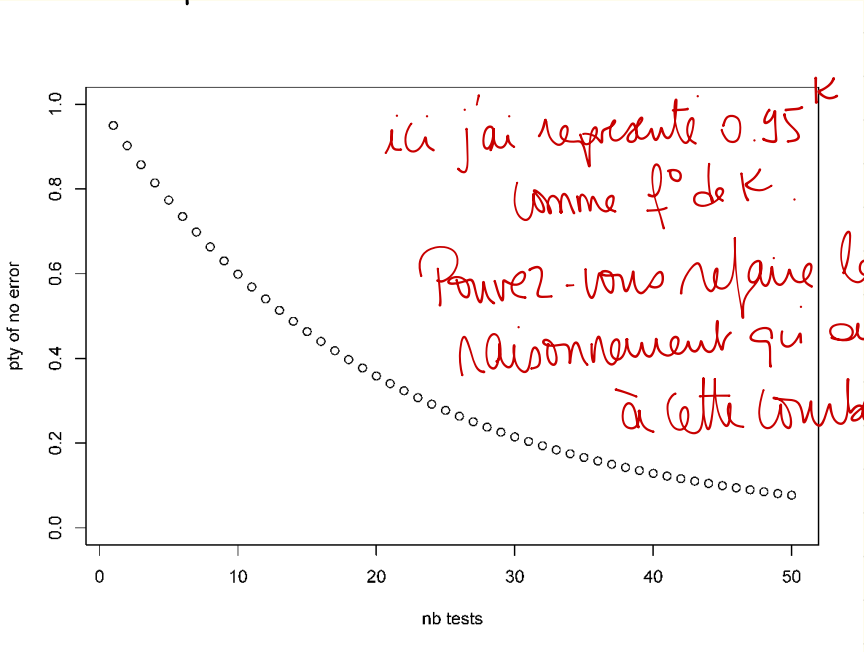




Or qu'a fait Tyler Vigen ? Combien a-t-il testé de courbes bizarres avant de trouver la paire qui va avec les films de Cage ???

Si chaque donnée est vraiment indépendante des films de Cage et qu'on teste K donnée différente, on trouve que la probabilité de trouver un lien alors qu'il n'y en a pas est

ne pas



ici j'ai représenté 0.95^K comme f° de K.

Pouvez-vous refaire le raisonnement qui amène à cette courbe ??

Que peut-on faire ?

Bonferroni (et on verra pourquoi) dit qu'il ne faut pas comparer à α mais α/K

avec pval = 2% il suffit de 3 tests et $\alpha = 5\%$ pour que ça soit plus significatif !!

→ Bref, "Vigen a triché..." en ne disant pas combien de données il a testé !!

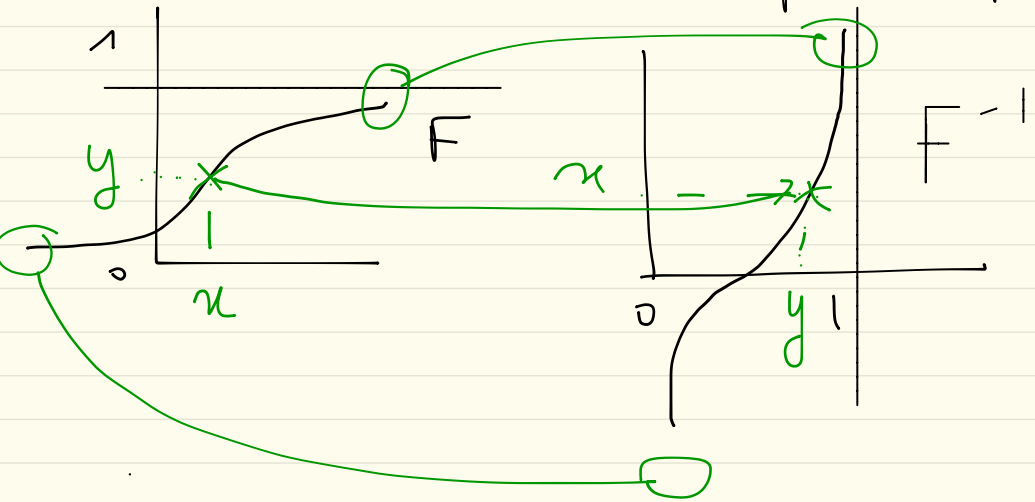
Tests non paramétriques

I Représentations graphiques et jugement qualitatif

1. Fonction de répartition

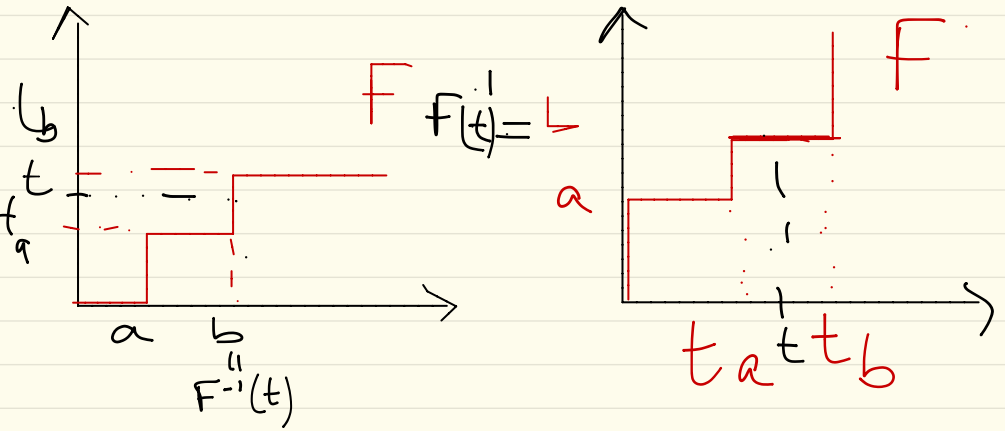
- la Fonction de répartition $t \mapsto P(X \leq t) = F(t)$ caractérise la loi de la variable réelle X.
- F est une fonction $\nearrow$ $\lim_{-\infty} F = 0$; $\lim_{+\infty} F = 1$
- La fonction quantile $F^{-1}(t) = \inf\{x / F(x) \geq t\}$
  - Si X a densité $f > 0$, F est bijective et continue

F^{-1} est la fonction réciproque



NB: Si $U \sim \mathcal{U}[0,1]$, $F^{-1}(U) \sim X$ & F c°, $F(X) \sim \mathcal{U}[0,1]$

Sinon F^{-1} est quand même bien défini



- Si $X_1, \dots, X_n$ iid de c.d.f F

alors la fonction de répartition empirique est donnée par

$$\forall t \in \mathbb{R}, \hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \le t}$$

C'est un estimateur sans biais et consistant de F

- Sans biais : $\forall t \quad \mathbb{E}\left[\hat{F}_n(t)\right] = F(t)$
- Consistant (dans l'espace des fonctions $\{F \uparrow \lim_{-\infty} = 0, \lim_{+\infty} = 1\} = \mathcal{F}$ muni de la $\|\|_{\infty}$)

$$\sup_{t \in \mathbb{R}} \left| \hat{F}_n(t) - F(t) \right| \xrightarrow[n \to +\infty]{P.S.} 0$$ (théorème de Glivenko Cantelli)

2. Représentation graphique et QQ plot

a) Si on pense que les X_i suivent une certaine loi de c.d.f. F_0

On peut toujours tracer $\hat{F}_n$ et F_0 et voir si ils sont proches

⇒ utiliser la commande ecdf de R
Simuler n va gaussiennes iid $\mathcal{N}(0,1)$
Vérifier le fit avec F_0 cdf de la $\mathcal{N}(0,1)$
regarder comment ça se déforme si on ne met pas la bonne moyenne ou la bonne variance

⇒ Si on veut vérifier que gaussien mais on ne connait pas moyenne et variance

↳ Que suggérez-vous ?

Estimer moyenne, variance, tracer F_0 la cdf d'une $\mathcal{N}(\hat{m}, \hat{\sigma}^2)$

b) Version estimation de densité

Ça vous dit qui ?
↳ histogramme
↳ estimateur à noyau

Que savez-vous sur les histogrammes ?

- en pratique : la f° hist, la f° density
- choix du pas ??

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

← avec 100 vous verrez qqch de convaincant mais si nombre d'observation petit moins clair

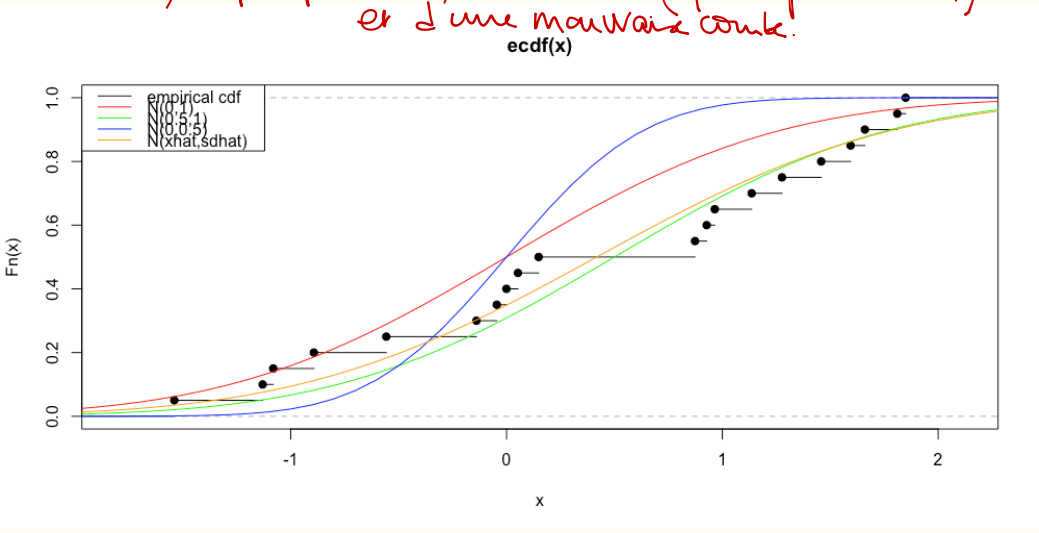
```
## fonction de repartition

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))
```

← regarder l'aide pour voir ce qu'est ecdf

regardez bien l'aide pour mettre au bon endroit.

← sur une seule ligne de commande

La courbe ci dessous a été tracée avec n = 20
On voit bien que la proximité avec la bonne courbe (rouge) dépendra, bcp + proche de l'estimé (qu'en pensez vous?) et d'une mauvaise courbe.



```
#### estimation de densite

hist(X,freq=FALSE)
rug(X)
lines(density(X),col='cyan')

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))
```

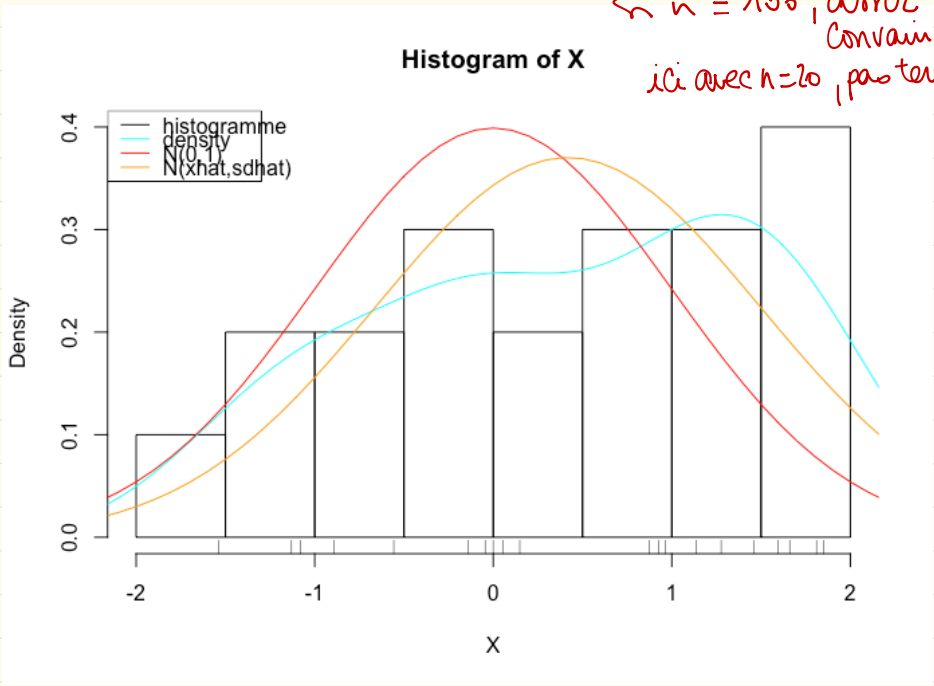
regarder bien l'aide, renvoyez règle de Sturges

pour voir où sont les points

attention ce n'est pas la vraie densité mais un estimateur, disséque l'aide pour voir ce que c'est!!!

si n = 100, assez convaincant

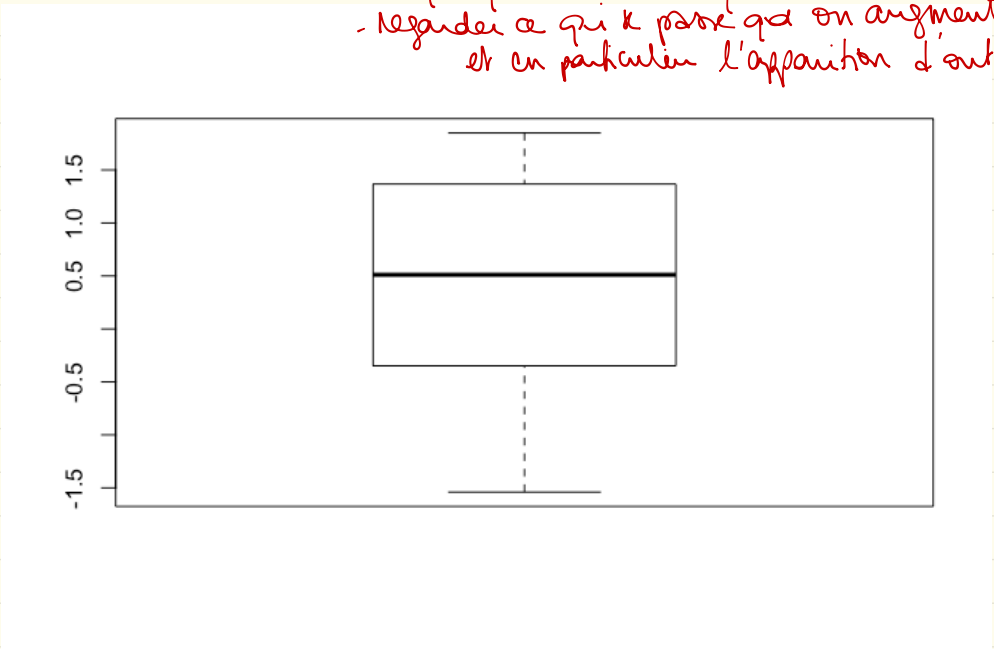
ici avec n=20, pas terrible



c) ### boxplot (boite a moustache)

boxplot(X)

Qd n petit, finalement boxplot assez suffisant
- Vérifier que vous savez ce qui est tracé
- regarder ce qui se passe qd on augmente n et en particulier l'apparition d'outliers



d) Qqplot

Idée du qqplot : on est visuellement plus sensible à "voir une droite"

→ Qqplot entre un échantillon $X_1 \dots X_n$ et une loi (cdf F_0)

$(x,y) \in$ Qqplot si $x = F_0^{-1}(t)$ et $y = \hat{F}_n^{-1}(t)$

F_0^{-1} : vrai quantile ; $\hat{F}_n^{-1}$: quantile empirique

Si $X_1, \dots, X_n \sim F_0$ le qqplot devrait s'aligner sur la diagonale $x=y$

Qqplot entre deux échantillons

$x_1 \dots x_n$ cdf empirique $\hat{F}_n$, $y_1 \dots y_n$ cdf empirique $\hat{G}_n$

$(x,y) \in$ Qqplot si $\exists t$, $x = \hat{F}_n^{-1}(t)$ et $y = \hat{G}_n^{-1}(t)$

Si les deux échantillons ont même loi ça devrait s'aligner sur la diagonale $x=y$

Qqplot pour loi normale

Si $X \sim \mathcal{N}(m,\sigma^2)$ alors $\frac{X-m}{\sigma} \sim \mathcal{N}(0,1)$

donc si Φ cdf de la $N(0,1)$

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

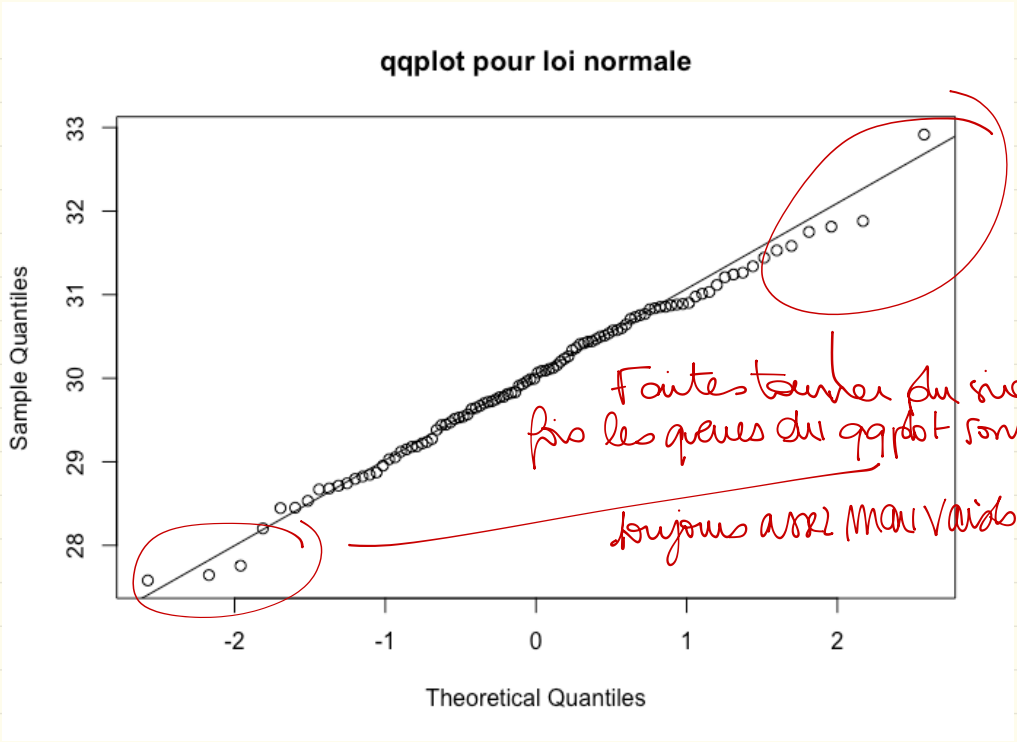
donc si $F^{-1}(u)$ quantile de X

$$F\left(F^{-1}(u)\right) = u = \Phi\left(\frac{F^{-1}(u)-m}{\sigma}\right) \quad \text{et} \quad F^{-1}(u) = m + \sigma \Phi^{-1}(u)$$

Donc si $X \sim \mathcal{N}(m,\sigma^2)$, les quantiles empiriques sont approximativement une fonction linéaire de $\Phi^{-1}(u)$ (quantile gaussien centré réduit)

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

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

← regarder l'aide ça rajoute la ligne qui passe par les 2 pts 1er et 3e quartiles (ie $t=0.25$ et $t=0.75$)

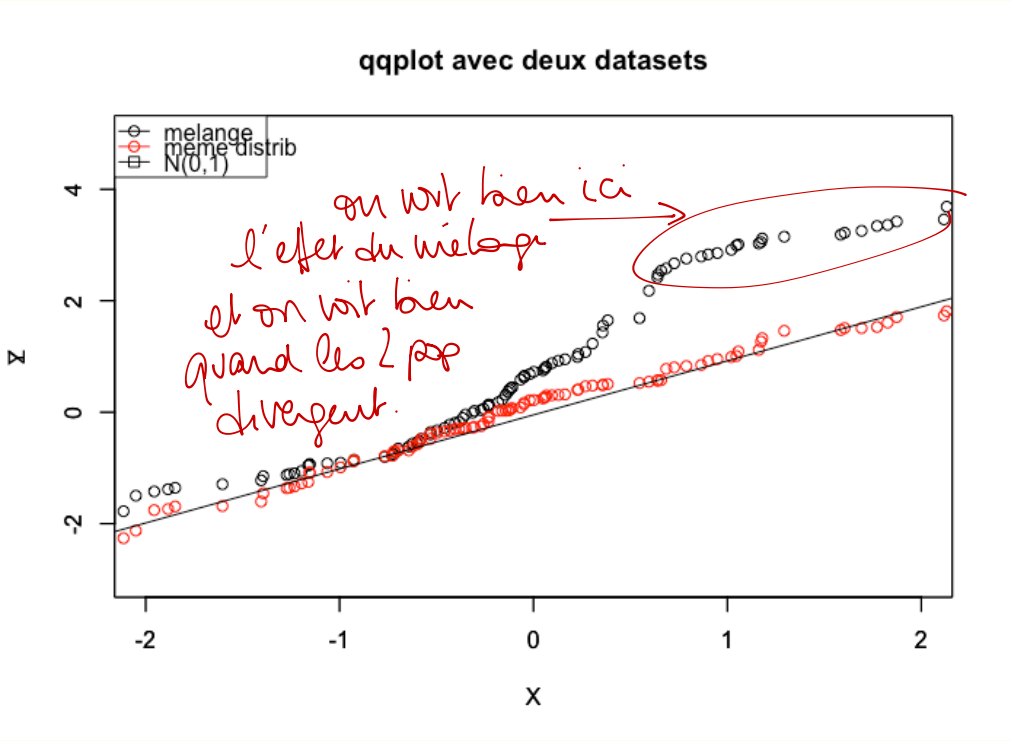



```
### qq plot de deux data sets

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'))
```

quelle loi ai-je simulé ??

qq plot à 2 échantillon

je trace la droite qui passe par les 2 quartiles de la variable distribuée



II Tests de Kolmogorov Smirnov et variantes

1 Test de KS à un échantillon

On observe $X_1 \dots X_n$ iid et on se donne une c.d.f. F_0

On veut tester H_0 : "les X_i ont pour cdf F_0" contre ce n'est pas le cas

Cas particulier de tests plus généraux appelés goodness-of-fit

On va rejeter si la distance entre la cdf empirique et F_0 est trop grande

le test de KS consiste à prendre comme distance (et donc statistique de test)

empirical c.d.f.

$$D_n^{F_0} = \sup_{t \in \mathbb{R}} \left| \hat{F}_n(t) - F_0(t) \right|$$

cdf supposée qu'on veut tester

* Sous H_0, les $X_1 \dots X_n$ ont pour cdf F_0 et on peut calculer/simuler la loi de $D_n^{F_0} \Longrightarrow$ quantiles $q_n^{F_0}(t)$ et f° de rép° $Q_n^{F_0}(t)$

On rejette si $D_n > q_n(1-\alpha)$ et cela fait un test de niveau α

pvalue : $1 - Q_n^{F_0}(D_n)$ NB : pour simplifier, je fais comme si $Q_n^{F_0}$ c°.

* La statistique $D_n^{F_0}$ est calculable facilement

En effet, si on range l'échantillon → les statistiques d'ordre

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

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

↑ sur cet intervalle, c'est juste $\frac{i}{n}$

$$= \sup_{i=0,\dots,n} \sup_{t \in [X_{(i)}, X_{(i+1)}[} \left| \frac{i}{n} - F_0(t) \right|$$

mais F_0 ↗ sur l'intervalle donc le sup est atteint aux extrémités

$$D_n^{F_0} = \max_{i=0,\dots,n} \max \left[\left| \frac{i}{n} - F_0(X_{(i)}) \right| ; \left| \frac{i}{n} - \lim_{\substack{t \le X_{(i+1)} \\ t \to X_{(i+1)}}} F_0(t) \right| \right]$$

$= F_0(X_{(i+1)})$ si F_0 c°

x Si F_0 est C^0, la loi de $D_n^{F_0}$ ne dépend pas de F_0

(pratique pour la tabulation, c'est cette loi qui est utilisée dans R)
→ warnings si R voit qu'il y a des ex aequo car ça ne peut pas être continu)

Preuve (Sort vient du calcul précédent car les $F(X_i) \sim \mathcal{U}([0,1])$ ou sinon on utilise l'équivalence suivante:

(vraie TF même non continue)

$$F^{-1}(x) \leq t \iff x \leq F(t)$$ pour tt F cdf.

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

$$\sim \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{F_0^{-1}(U_i) \leq t} - F_0(t) \right| \quad (U_i \text{ iid} \sim \mathcal{U}([0,1])$$

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

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

Si F_0 C^0, $F_0(\mathbb{R}) = [0,1]$ et c'est $= D_n^{\mathcal{U}([0,1])}$

Sinon $F_0(\mathbb{R}) \subsetneq [0,1]$ et $\leq D_n^{\mathcal{U}([0,1])}$

Donc $P(D_n^{F_0} > t) \leq P(D_n^{\mathcal{U}([0,1])} > t)$

Les quantiles devraient donc s'appliquer sans pb.
En pratique dans R, ça ne marche pas car il ne calcule pas la bonne stat de test.

x asymptotiquement, TCL $\sqrt{n}\left(\hat{F}_n(t) - F_0(t)\right) \xrightarrow[\text{sous } H_0]{\mathcal{L}} \mathcal{N}(0, F_0(t)(1-F_0(t)))$

$\Rightarrow$ on peut mq $\sqrt{n} D_n \xrightarrow[\text{sous } H_0]{\mathcal{L}}$ une distribution connue et tabulée (F_0 est C^0)

NB parfois D_n = en fait $\sqrt{n} D_n$ (ie on renormalise par $\sqrt{n}$ la stat de test)

NB : Kolmogorov (33)

$$P\left(\sqrt{n} \sup_t \left|\hat{F}_n(t) - F_0(t)\right| \geq x\right) \xrightarrow[n \to +\infty]{} 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2x^2}$$

on reconnaît une "queue gaussienne"

Smirnov (42)

$$P\left(\sqrt{n} \sup_t \left[\hat{F}_n(t) - F_0(t)\right] \geq x\right) \to e^{-2x^2}$$

$D_n^+ = \sup_t \left(\hat{F}_n(t) - F_0(t)\right)$ utile pour

$H_0 : \forall t \; F(t) \leq F_0(t)$ contre $\exists t / F(t) > F_0(t)$

car F cdf des X, F_0 cdf de X_0

si $F(t) \leq F_0(t) \; \forall t$, alors pour $U \sim \mathcal{U}[0,1]$ $F^{-1}(U) \geq F_0^{-1}(U)$ ($F^{-1}(t) \geq F_0^{-1}(t)$)

ie on peut construire des versions de X et X_0 tq $X \geq X_0$

on dit alors que X est stochastiquement plus petit que X_0.

donc ici si X et X_0 sont des durées de vie

H_0 : "X_0 meurt avant X" H_1 : "une durée où X_0 atteint t plus souvent que X.

Massart (90) $\forall n \; P(\sqrt{n} D_n^+ \geq x) \leq e^{-2x^2}$ et

$$P(\sqrt{n} D_n \geq x)) \leq 2e^{-2x^2} \quad (*)$$

$\times$ on peut donc en déduire ce que fait le test basé sur D_n l'alternative $H_1 : \exists t, \; F(t) \neq F_0(t)$

Sous H_1,

$$< \|F - F_0\|_\infty \leq \underbrace{\|F - \hat{F}_n\|_\infty} + \|\hat{F}_n - F_0\|_\infty$$

$\xrightarrow{PS} 0$ qd $n \to +\infty$ (à la vitesse $\sqrt{n}$ vu l'étude précédente)
(Glivenko Cantelli)

Donc $P\left(\|\hat{F}_n - F_0\|_\infty \geq \|\frac{F - F_0}{2}\|_\infty \right) \xrightarrow[n \to +\infty]{} 1$

Or par $*$ par ex, en posant $\alpha = 2e^{-2x^2}$

ie $x_\alpha = \sqrt{\frac{\ln(\frac{2}{\alpha})}{2}}$, $P\left(D_n \geq \frac{x_\alpha}{\sqrt{n}} \right) \leq \alpha$

donc $q_n^{F_0}(1-\alpha) \leq \frac{x_\alpha}{\sqrt{n}}$

donc $P(\text{test rejette}) = P\left(\|\hat{F}_n - F_0\|_\infty \geq q_n^{F_0}(1-\alpha) \right)$

$$\geq P\left(\|\hat{F}_n - F_0\|_\infty \geq \frac{x_\alpha}{\sqrt{n}} \right)$$

$$\geq \text{(pour } n \text{ assez grand)} \; P\left(\|\hat{F}_n - F_0\|_\infty \geq \frac{\|F - F_0\|_\infty}{2} \right)$$

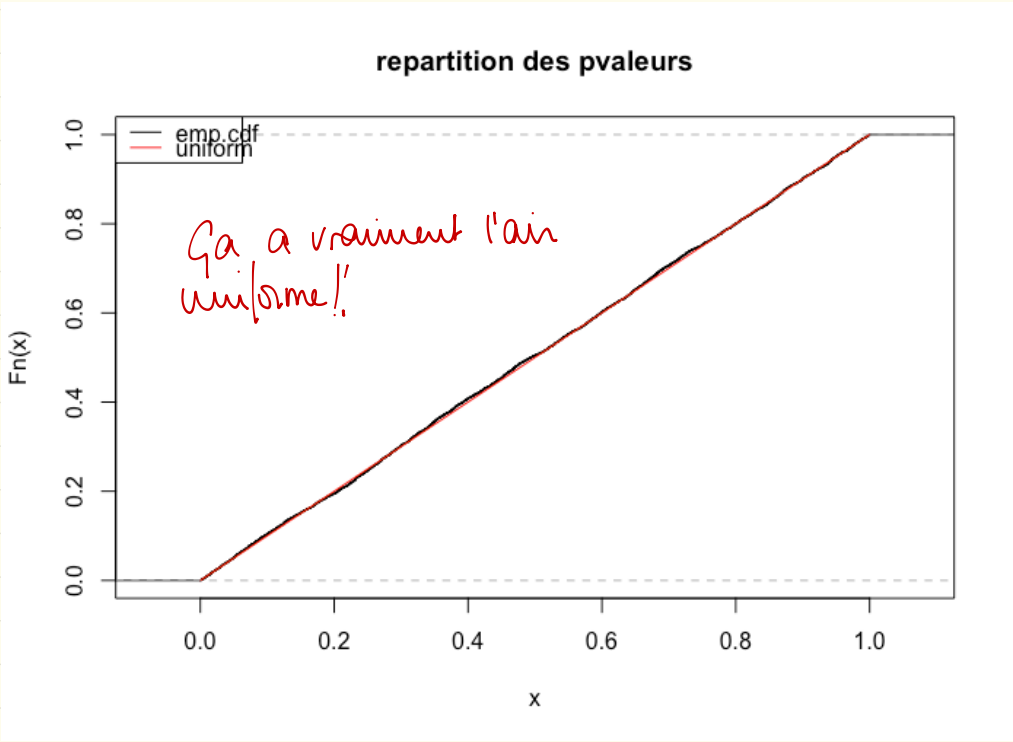
$$\xrightarrow[n \to +\infty]{} 1$$

Donc sous H_1, la puissance du test de KS tend vers 1

→ Comment le feriez vous ?
↳ méthode de Monte Carlo

```
# verification sous H0 du niveau du test de KS

Nsimu=5000 ← jouer avec Nb de simu
n=30 # taille de l'echantillon ← sur lequel on fait KS

pval_cont=rep(0,Nsimu) # on initialise


for(i in 1:Nsimu)
{
  X=rnorm(n) # l'echantillon gaussien F0 continu

  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))
```

NB : amusez vous à prendre avoir une distribution discontinue pour voir ce qui k passe



```
alpha=0.05 # je verifie que  le test est de niveau 5%

niveau_emp=mean(pval_cont<alpha) # comment savoir si la deviation est raisonnable ?

# test du "niveau = 5%"

# sous H0, niveau_emp est presqu'une gaussienne N(0.05, 0.05*0.95/Nsimu)

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

abs(alpha-niveau_emp)>erreur_adm # c'est le rejet de ce test qui est faux donc on accepte

#on peut aussi verifier plus globalement que les pvaleurs sont uniformes

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

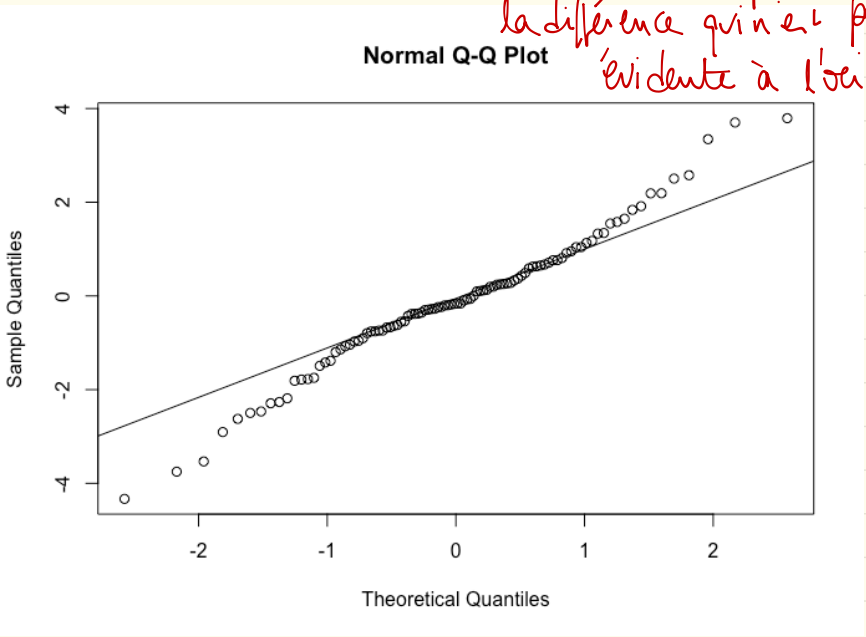
↑ c'est jamais 5% pile ! Regarder la variation en fct de Nsimu pourquoi 5000 est plutôt bon ??

On teste le test :)

```
## sous H1

n=100 # taille de l'echantillon

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

qqnorm(X)
qqline(X)
```

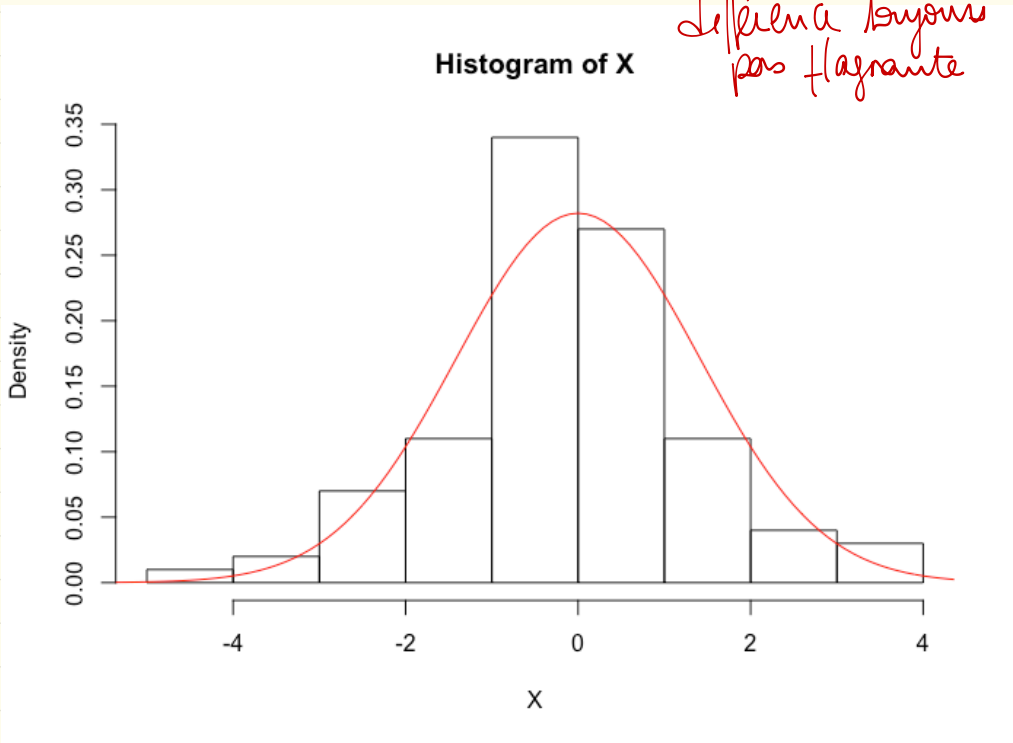
← c'est une loi de Laplace. Devinez sa densité.

↑ test nécessaire pour voir la différence qui n'est pas évidente à l'oeil nu



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

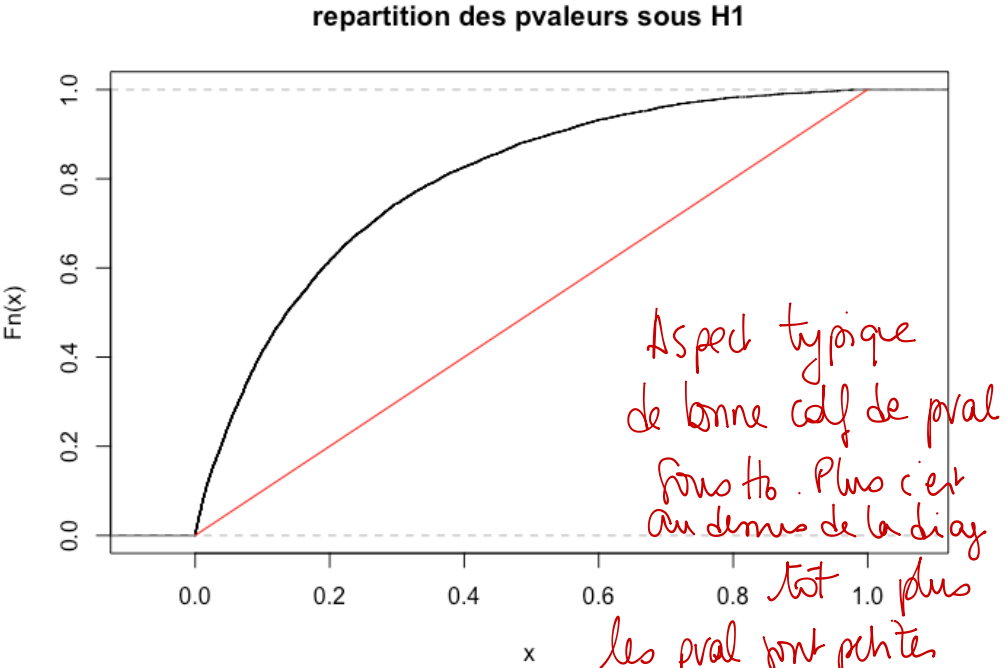
différence toujours pas flagrante




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

← KS voit la différence en principe.

```
for(i in 1:Nsimu)
{
  X=(2*rbinom(n,1,0.5)-1)*rexp(n,1) # l'echantillon expo et on va se demander si gaussien

  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')
```



Aspect typique de bonne calf de pval sous H_0. Plus c'est au dessus de la diag tôt plus les pval sont petites et plus on détecte la différence

2. Les variantes (change la métrique et les quantités mais ne change rien au raisonnement)

Cramer-von Mises utilise la distance L_2

$$n \int \left(\hat{F}_n(t) - F_0(t)\right)^2 dF_0(t)$$

↗ si besoin de la renormalisation

NB: dans le package gof test on les trouve

NB2: gof = goodness of fit

Anderson-Darling

$$n \int \frac{\left(\hat{F}_n(t) - F_0(t)\right)^2}{F_0(t)\left(1 - F_0(t)\right)} dF_0(t)$$

3. Le test de KS à deux échantillons

$X_1 \dots X_n$ iid de cdf F et $Y_1 \dots Y_m$ iid de cdf G

On veut tester H_0 : "$F = G$" contre H_1 ce n'est pas le cas

C'est une hypothèse nulle composite

↳ il faut trouver une statistique dont la loi ne dépend pas de la valeur sous-jacente de $F = G$

$$\mathcal{T}_{n,m} = \sup_t |\hat{F}_n(t) - \hat{G}_m(t)|$$

Sa loi ne dépend pas de $F = G$ si $\underline{C^0}$

→ on rejette quand > au quantile d'ordre $1-\alpha$

```
n=1000
X=rnorm(n)
Y=(2*rbinom(n,1,0.5)-1)*rexp(n,1)

ks.test(X,Y)
```

← jouez avec n pour vous rendre compte qu'il va être beaucoup plus difficile de voir que "deux échantillons sont $\neq$" par rapport à "cet échantillon ne suit pas cette loi"

NB : En pratique si les F pas C^0, c'est que X est discret
⇒ tests du χ^2 (à revoir)

II Les tests de rang de Wilcoxon (1945)

1. Le signed-rank test

Si $X_1 \dots X_n$ iid on veut tester

H_0 : la distribution de X est symmétrique autour de 0
Contre H_1 : ce n'est pas le cas.

Idée : Si la distribution de X est symmétrique

on peut toujours la voir comme $\text{sign}(X)\,|X|$

avec $\text{sign}(X)$ Rademacher (ie vaut ± 1 avec proba $1/2$) sous H_0

Pour rendre libre de la loi de $|X|$ on transforme les $|X_i|$ en leur rang R_i. NB rang de $|X_i|$ ou rang de $F_{|1|}(|X_i|) \sim U_{i\,iid}[0,1]$ (cdf de $|X|$) c'est pareil.

Si $X_i \neq 0$ ps $\quad W = \sum_{i=1}^{n} \text{sign}(X_i)\, R_i$: la loi ne dépend pas de la distrib de $|X|$

$\Rightarrow$ accès au quantile et rejet quand trop grand. ou trop petit

Intérêt principal = Comparaison de deux échantillons appariés

à chaque exp i, on observe X_i et Y_i, on veut savoir si $X_i > Y_i$
(X_i, Y_i) iid

On pose $Z = X - Y$ et on teste

H_0 : distrib de Z symmétrique
X est autant au dessus qu'au dessous de Y

H_1 : elle est décentré vers $\mathbb{R}_+$
X est plus souvent au dessus de Y

revient à rejeter pour grande valeur de W unsigned

↳ WilCox.test(x, y, alternative = "greater", paired = TRUE)

2. Le rank-sum test (aussi appelé test de Mann-Whitney)

On observe $X_1, \dots, X_n$ iid indépendants de $Y_1, \dots, Y_m$ (et donc sont pas appariés !!)
$\sim X$ $\sim Y$

On suppose que $X \perp\!\!\!\perp Y$ et l'on veut savoir si X est significativement plus grand que Y

Sous H_0 : $X \sim Y$, $P(X > Y) = P(Y > X)$

donc $P(X > Y) + 0.5\,P(X = Y) = P(Y > X) + 0.5\,P(Y = X)$

et comme la somme = 1, $P(X > Y) + 0.5\,P(X = Y) = 0.5$

l'alternative est donc comprise comme H_1 : $P(X > Y) + 0.5\,P(X = Y) > 0.5$
pour dire X plus "grand" que Y

Stat de test $U = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}_{y_j < x_i} + 0.5\,\mathbb{1}_{y_j = x_i}$

NB si $X \sim Y$, la loi de U ne dépend pas de la loi sous jacente

↳ on utilise son quantile pour rejeter qd U dépasse ce quantile d'ordre $1-\alpha$

↳ WilCox.test $(X, Y,$ alternative = greater$)$

↑ par défaut pas apparié

NB : variante two sided / lower etc...
(default)

NB2 : Ce n'est pas la même alternative que KS deux échantillons
one-sided.

III Tests de normalité (cf package nortest)

On observe $X_1 \dots X_n$ iid de cdf F, on note Φ_{m,σ^2} la cdf de $N(m,\sigma^2)$

On veut tester

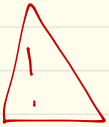
H_0 : "$\exists m, \sigma^2 / F = \Phi_{m,\sigma^2}$" contre ce n'est pas le cas

1. Lilliefors

Si on connaissait m et σ^2 on pourrait utiliser

$$D_n = \sup_t \left| \hat{F}_n(t) - \Phi_{m,\sigma^2}(t) \right|$$

↳ Comme on ne les connaît pas, on pourrait faire du plugin
= remplacer m et σ^2 par $\hat{m}$ et $\hat{\sigma}^2$



ça change la loi de D_n (ce qui n'est pas grave en soi)
↳ il faut vérifier que ça ne dépend pas des m et σ^2 inconnus sous H_0

$$T_n = \sup_t \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \le t} - \Phi\left(\frac{t-\hat{m}}{\hat{\sigma}}\right) \right| \quad \text{on pose } u = \frac{t-\hat{m}}{\hat{\sigma}}$$

Φ ← cdf de $N(0,1)$

$$= \sup_u \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \le \hat{m} + u\hat{\sigma}} - \Phi(u) \right|$$

$$= \sup_u \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \le u} - \Phi(u) \right| \quad \text{où } Y_i = \frac{X_i - \hat{m}}{\hat{\sigma}}$$

Si $\hat{m} = \bar{X}$ et $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

alors la loi de $\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ ne dépend pas de m et σ^2 ni $\hat{m}$ ni les cdf ne sont plus $\mathbb{1}$.

$\Rightarrow$ la loi de T_n sous H_0 est tabulée $\Rightarrow$ le test (écrivez le ...)

⇒ Lilliefors test du package nortest en R.

2. Les variantes

On peut faire le même genre de manip sur les tests de Cramer Von Mises et Anderson (par abus de langage, on les appelle tjs pareil)

↳ aussi dans le package nortest

3. Chi-deux

On peut tjs faire du test du chi deux en regroupant par classe. (même valable pour KS à 1 éch)

Si m et σ^2 pas connus → Test avec estimateurs de paramètres !
du chi deux (+/- vrai car distrib pas discrète)

(Vous avez dû les voir l'année prochaine)

↳ aussi dans le package nortest

4. Shapiro et Nilk

↳ shapiro.test Connu pour être le + puissant

aller voir la formule si vous êtes curieux.