

# Statistiques avec



M2 Sciences du Langage

[Remi.lafitte@univ-grenoble-alpes.fr](mailto:Remi.lafitte@univ-grenoble-alpes.fr)

2023-2024

# Statistiques descriptives et inférentielles

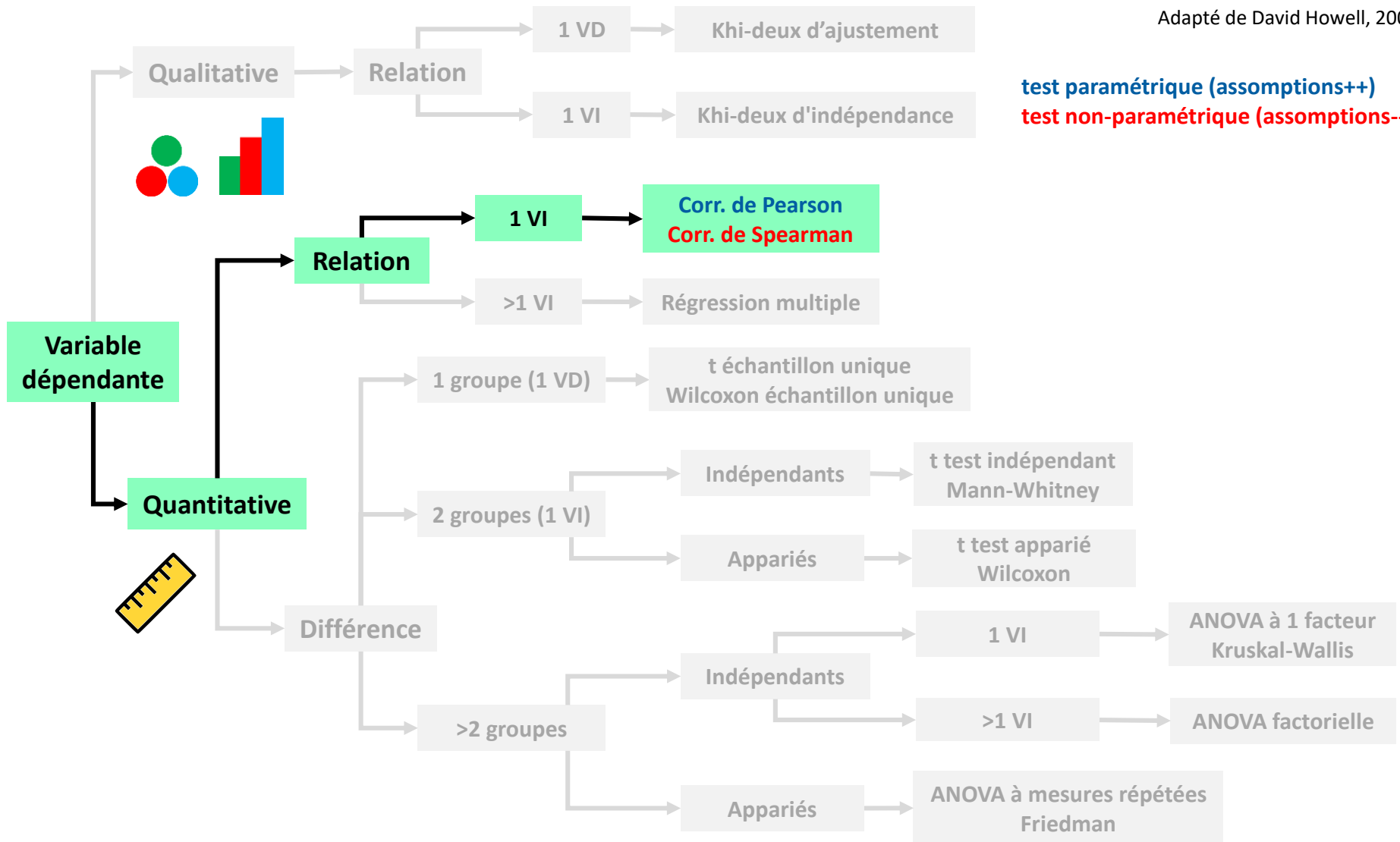


# Corrélations

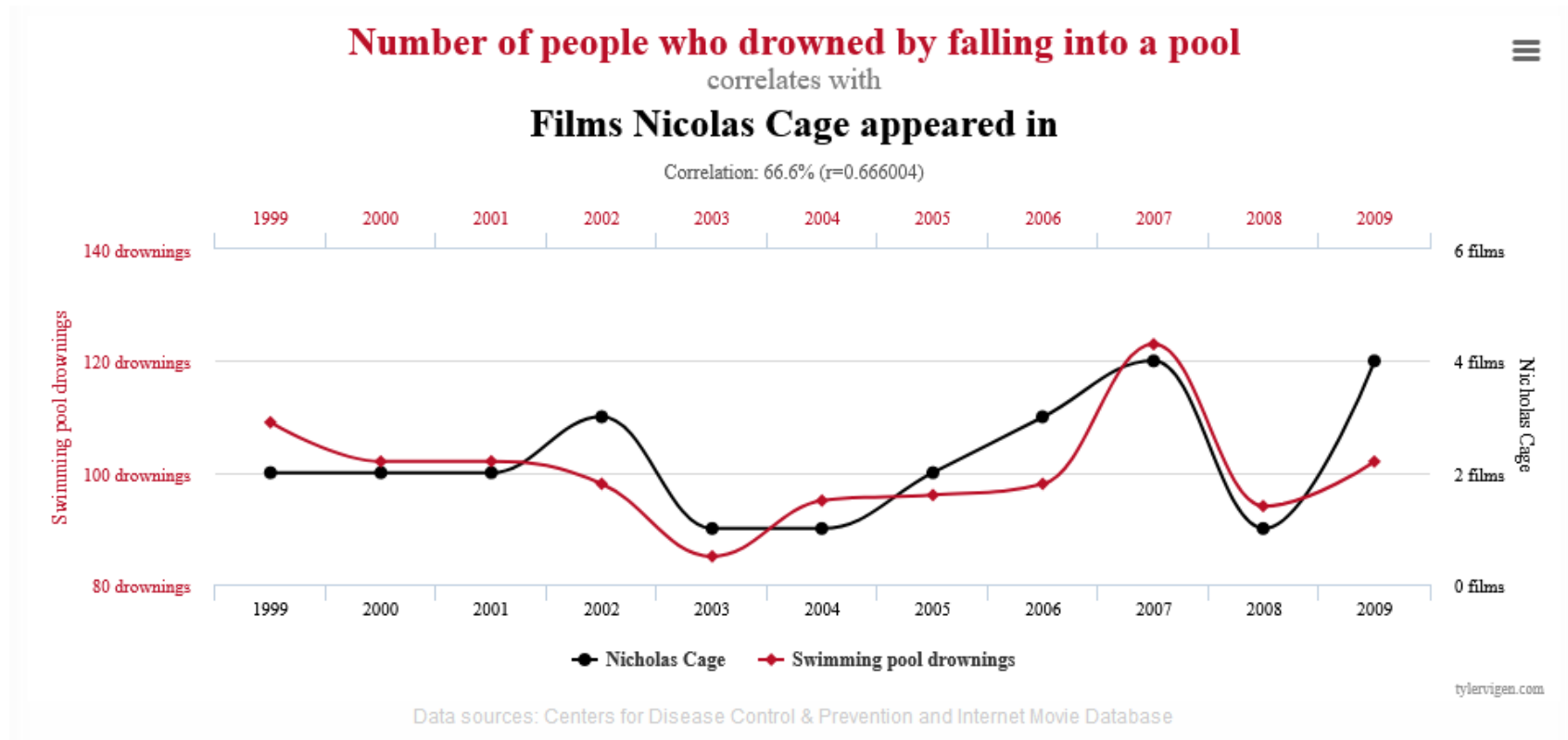


Adapté de David Howell, 2008

test paramétrique (assumptions++)  
test non-paramétrique (assumptions--)



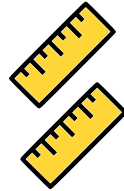
**/!\ corrélation n'est pas causalité /!\**



# Corrélation de Pearson

- Contexte

- 1 VD quantitative
- 1 VI quantitative



permutables

$H_0$  = hyp nulle

$H_1$  = hyp alternative

- $H_0/H_1$  : Il n'y a pas de / il y a une association LINEAIRE entre les deux variables dans la POPULATION

les deux variables augmentent ou diminuent simultanément et à une vitesse constante

- Exemple :

- Association linéaire entre le bonheur d'un pays et son nombre de groupes de métal ?

```
DF <- readxl::read_xlsx("metal_bands.xlsx")
head(DF); summary(DF) ; str(DF)
```

```
## # A tibble: 6 × 4
##   Territory    Bands Population Happiness
##   <chr>      <dbl>      <dbl>      <dbl>
## 1 Afghanistan     2    37466414      2.40
```



# Corrélation de Pearson

- NB : dans cet exemple nous allons supprimer les pays avec des données manquantes
- De cette façon c'est comme si notre DF contenait un **échantillon** de certains pays

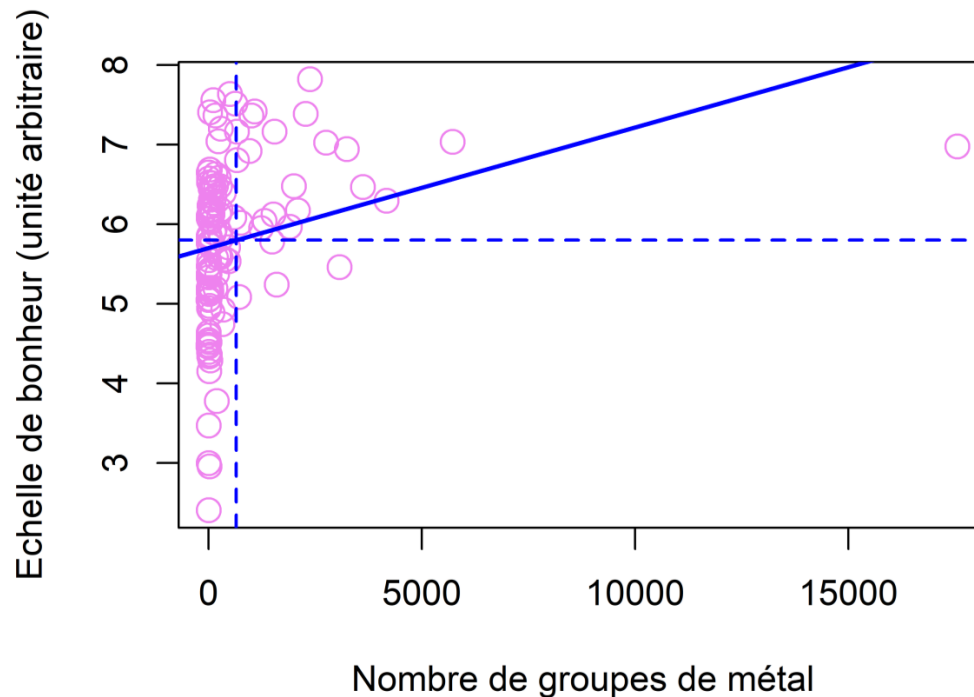
```
DF <- na.omit(DF[c("Territory", "Bands", "Happiness")])
```

- **Nous avons un total de 117 pays** (sur 197 selon wikipédia)

# Corrélation de Pearson

- Graphique
- NUAGE DE POINTS
- Possible de rajouter une **droite de régression** passant par les points

```
CORR_MOD <- lm(Happiness ~ Bands)      # modèle de régression linéaire  
abline(CORR_MOD, col = "blue", lwd=2) # droite estimée par le modèle
```



# Corrélation de Pearson

- Possible de rajouter une **droite de régression** passant par les points

```
coefficients(CORR_MOD)
```

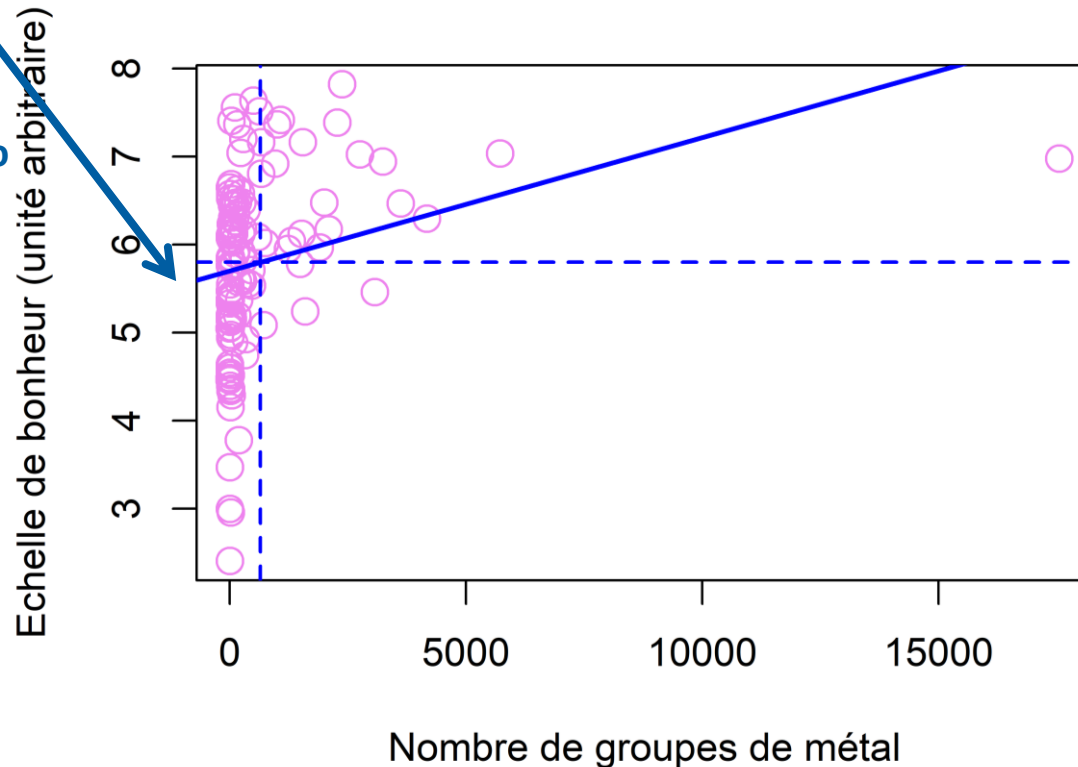
```
## (Intercept)
```

```
Bands
```

```
## 5.7023814491 0.0001514175
```

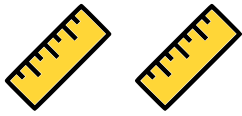
Augmentation prédite du  
bonheur pour chaque groupe de  
métal en plus

Bonheur  
prédit si zéro  
groupe de  
métal





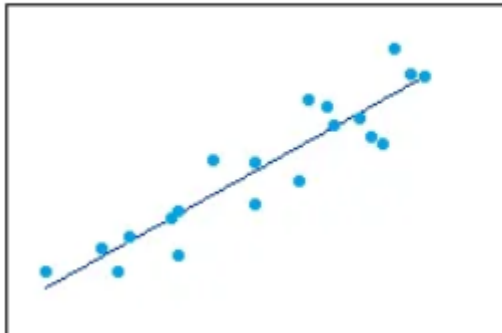
# Corrélation de Pearson

- **/!\** Conditions d'application
- Deux variables **continues** 
- Les paires de variables sont **indépendantes**  
ex : elles ne viennent pas du **même sujet**
- Relation **linéaire** entre les deux variables

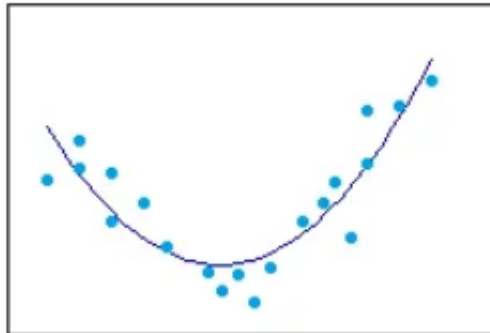
ces 2 paires d'observations ne sont pas indépendantes !

##	Sujet	Age	Happiness	Height
##	1	20	2.4	165
##	1	30	4.7	165
##	2	45	3.6	178
##	3	19	2.6	185

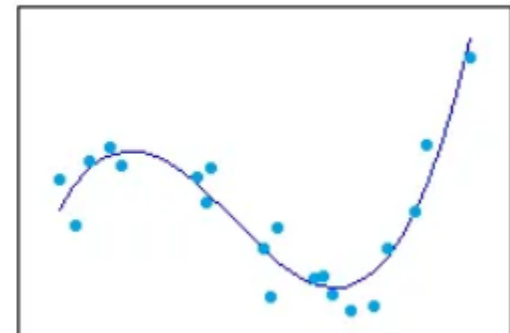
Linear



Quadratic



Cubic

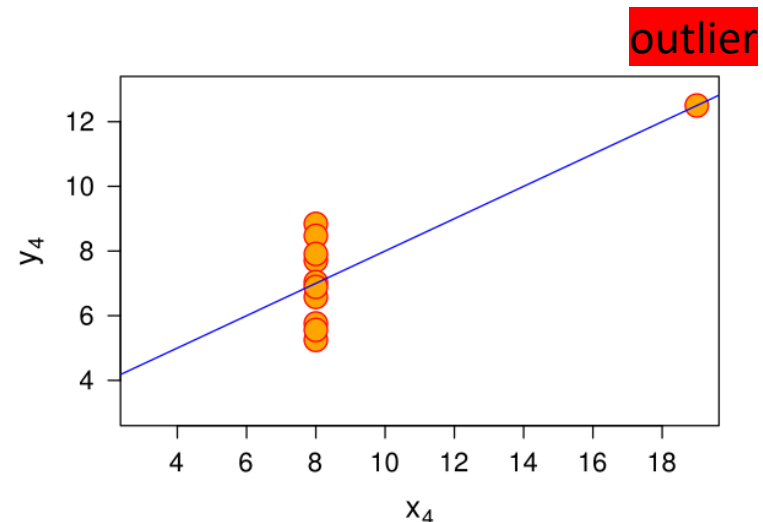
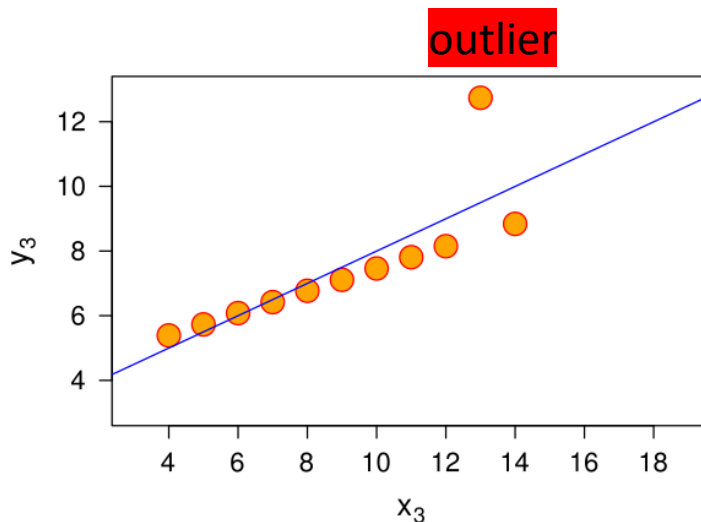


# Corrélation de Pearson

- **/!\** Conditions d'application
- Absence d'**outliers** ("cas influents")



*Outliers, by virtue of being **different from other cases** ... usually exert **disproportionate influence** on **substantive conclusions** regarding relationships among variables (Aguinis et al., 2013)*



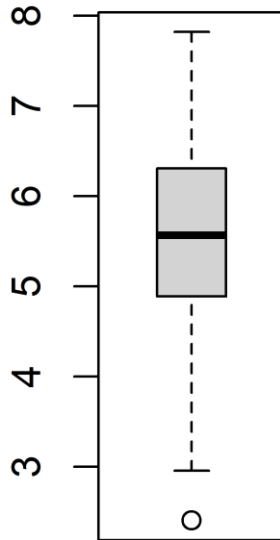
# Corrélation de Pearson

- (1) Détection des outliers

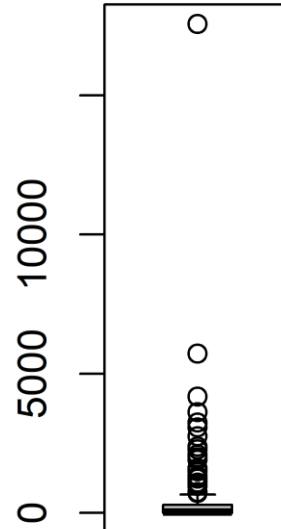
- Boxplots univariés

```
boxplot(Happiness)
boxplot(Bands)
```

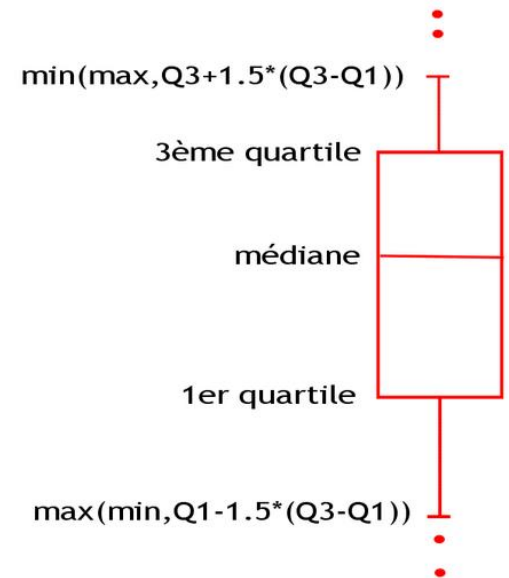
Echelle de bonheur (unité arbitraire)



Nombre de groupes de métal



## Anatomie du boxplot



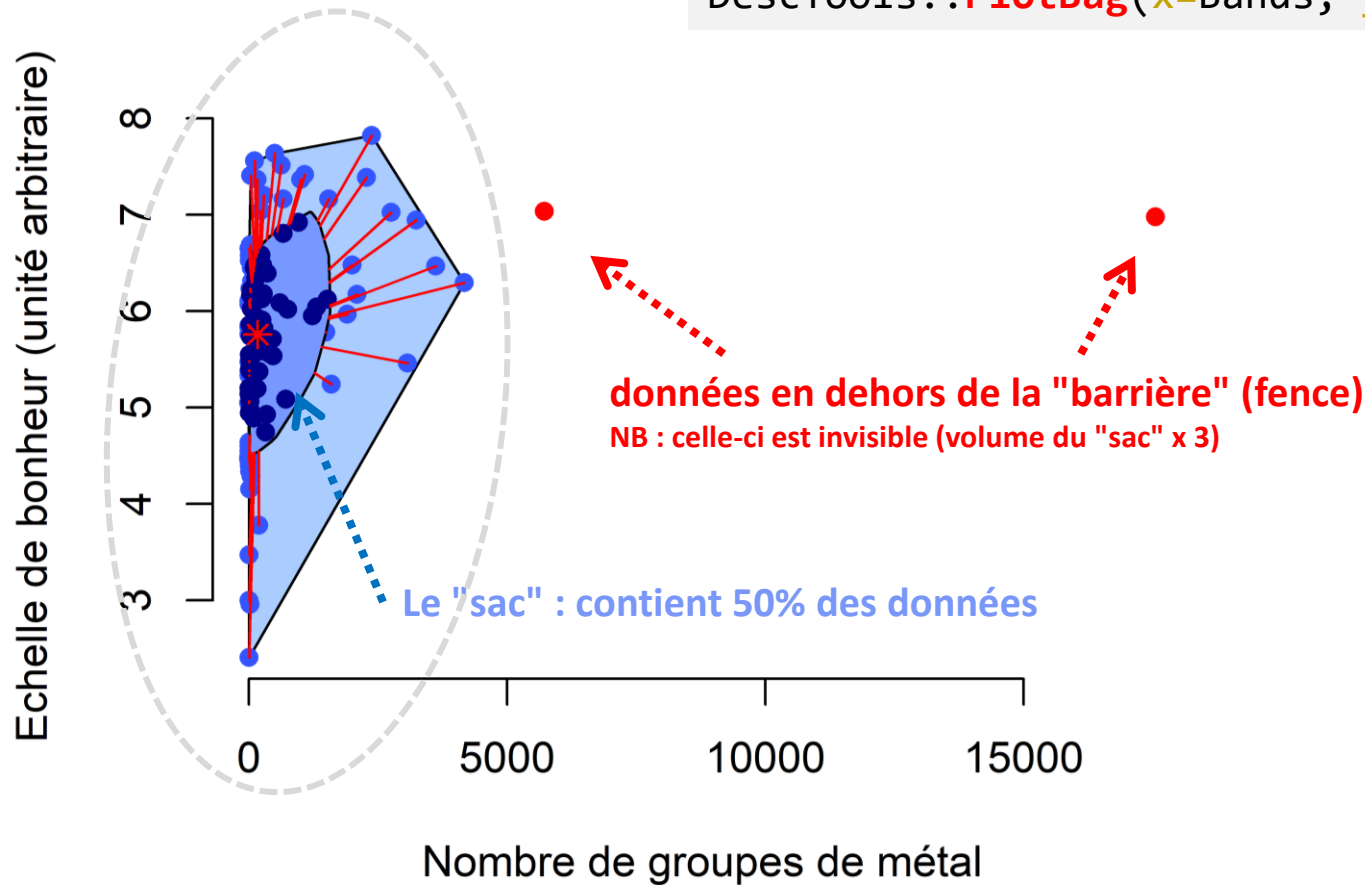
**!!** ne montre pas les **paires** d'observations déviantes ...

# Corrélation

- (1) Détection des outliers
- Boxplot bivarié !

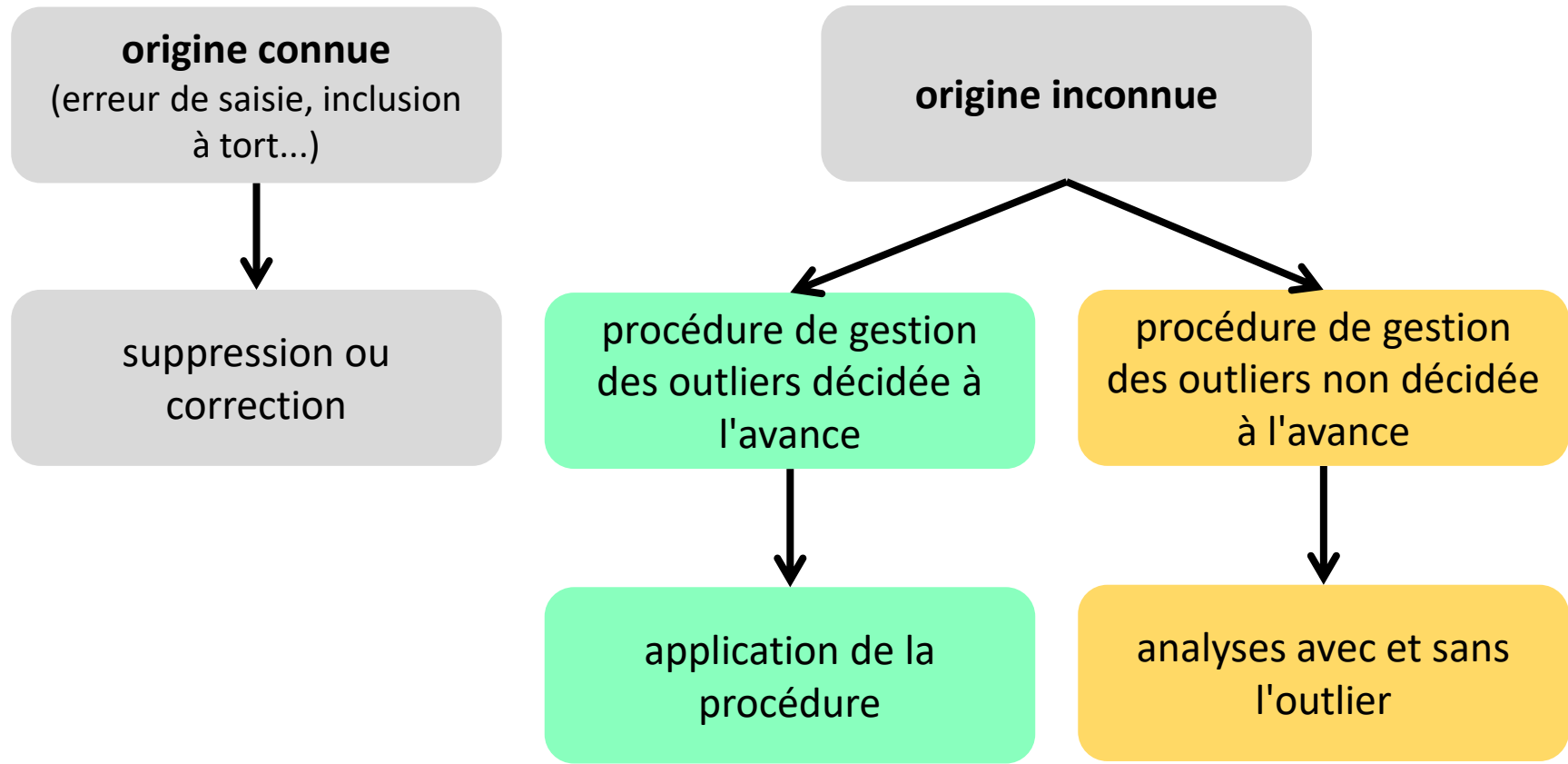
```
library(DescTools) # à installer
```

```
DescTools::PlotBag(x=Bands, y = Happiness)
```



# Corrélation de Pearson

- (2) Traitement des outliers



# Corrélation de Pearson

## Statistical models

The PSEs of the line midpoint and visual vertical will be separately analyzed by one-way repeated measures ANOVA with visual cue condition as factor. Pairwise comparisons will be performed with one-sided (line midpoint) or two-sided (visual vertical) paired t-tests.

In addition, for each participant the PSEs of the line midpoint and visual vertical will be regressed on the visual cue conditions, coded as [left-cue = 1, no-cue = 2, right-cue = 3]. This way each participant will have two slope values estimating the sensitivity of her/his subjective line midpoint and visual vertical to the visual cue conditions. The slope for the visual vertical PSE will be regressed on the slope for the line midpoint PSE.

*No files selected*

## Transformations

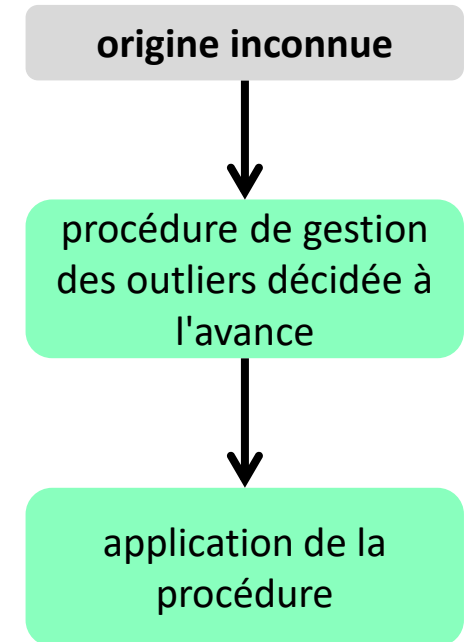
If the residuals of the linear tests are not normally distributed, according to visual inspection and Shapiro test, the dependent variable will be rendered more Gaussian with the Yeo-Johnson transformation. Sensitivity analyses will be performed to analyse how the results differ with this transformation. Non-parametric tests will be used as a last option if the transformations are ineffective.

## Inference criteria

We will use the standard alpha rate of 5% for determining whether the ANOVAs and the linear regression are significant or not. The alpha rate of the post-hoc t-tests will be corrected with the Holm-Bonferroni method.

## Data exclusion

For the ANOVAs we will examine the studentized residuals of each t-test (cutoff = 4) to check the presence of highly influential observation.

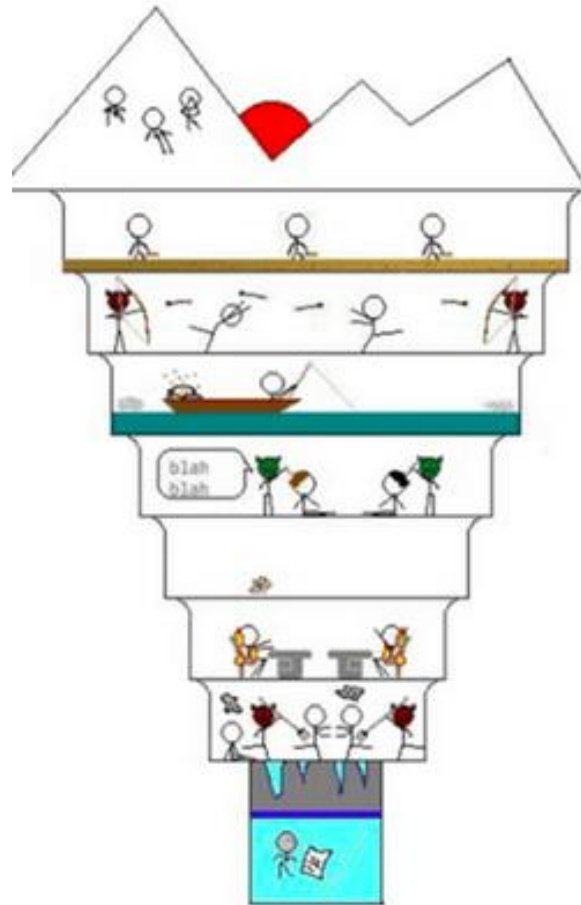


ex. de **pré-enregistrement**  
d'une étude expérimentale en  
psychologie

<https://osf.io/jtsuf>

# Corrélation de Pearson (aparté science ouverte)

## The Nine Circles of Scientific Hell



**First Circle: Limbo**

**Second Circle: Overselling**

**Third Circle: Post-Hoc Storytelling**

**Fourth Circle: P-Value Fishing**

**Fifth Circle: Creative Use of Outliers**

Those who “cleaned up” their results by excluding inconvenient data points are condemned here. Demons pluck out their hairs one by one, each time explaining that the sinner is better off without that hair, because there was something wrong with it.

origine inconnue



procédure de gestion  
des outliers **non**  
décidée à l'avance

**!! DANGER !!**

Neurosceptic (2012)

# Corrélation de Pearson (aparté science ouverte)



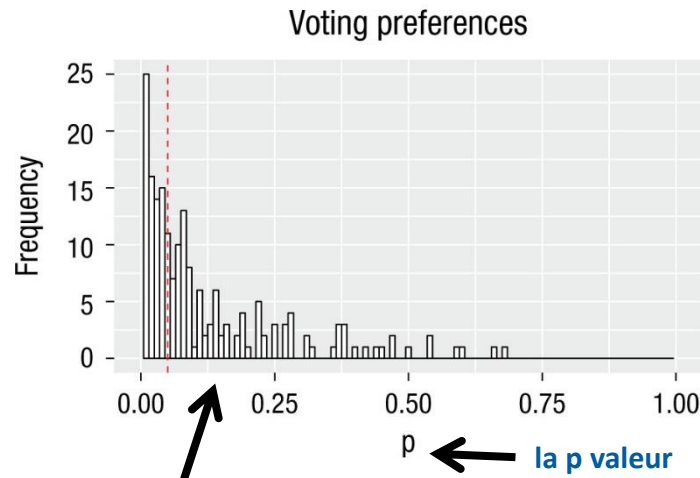


# Corrélation de Pearson (aparté science ouverte)

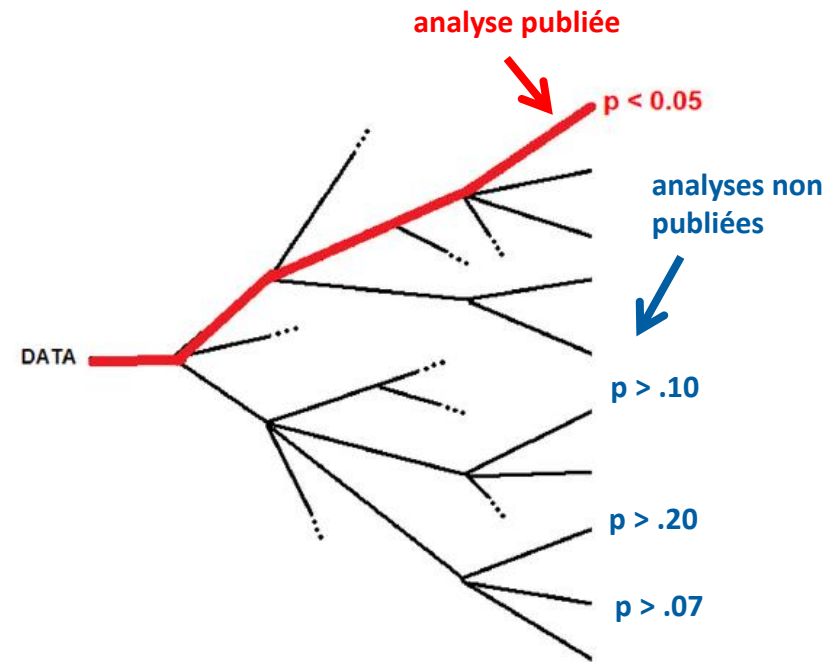
- Les dangers qui vous guêtent ...

## Une analyse implique de nombreux choix :

- choix de(s) la VD
- choix et codage des VI
- détection et traitement des outliers
- inclusion de variables contrôles



120 barres = 120 manières de traiter les données !

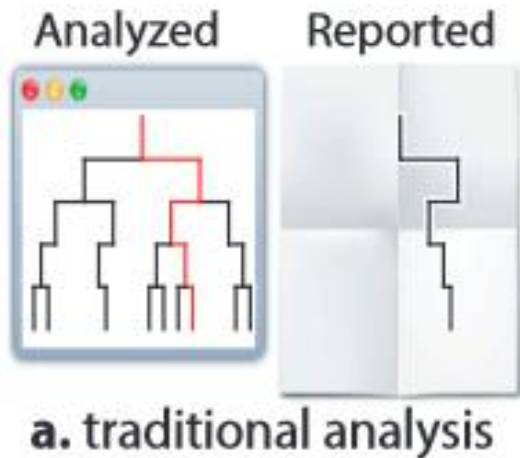


# Corrélation de Pearson (aparté science ouverte)

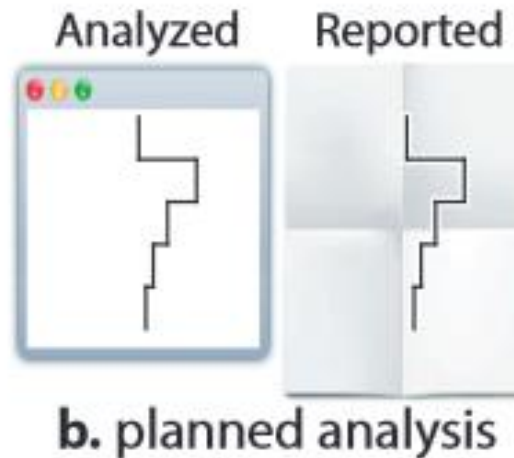


- Les dangers qui vous guêtent...

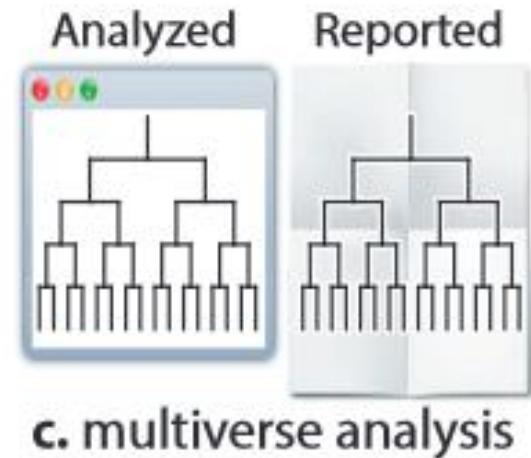
**Deux solutions : le pré-enregistrement et l'analyse "multivers"**



(-) = Risques de p hacking  
(-) = Risques de conclusions fragiles



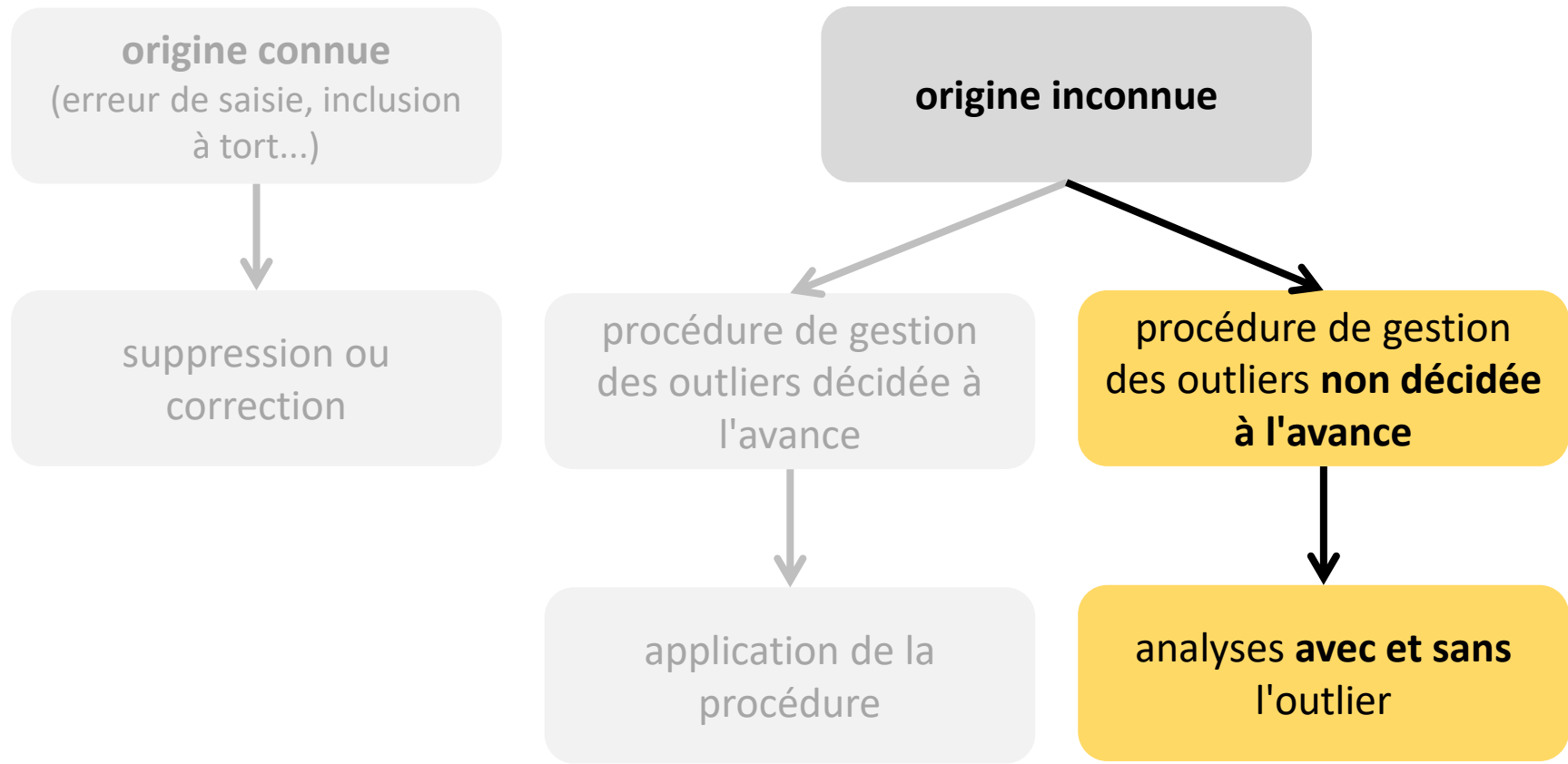
(-) = Contraignant  
(+) = ... mais possibilité de revenir sur ses choix si arguments  
(+) = une meilleure science



(-) = Beaucoup de travail !  
(+) = Epreuve la robustesse des résultats vis-à-vis des choix du chercheur

# Corrélation de Pearson

- Traitement des outliers



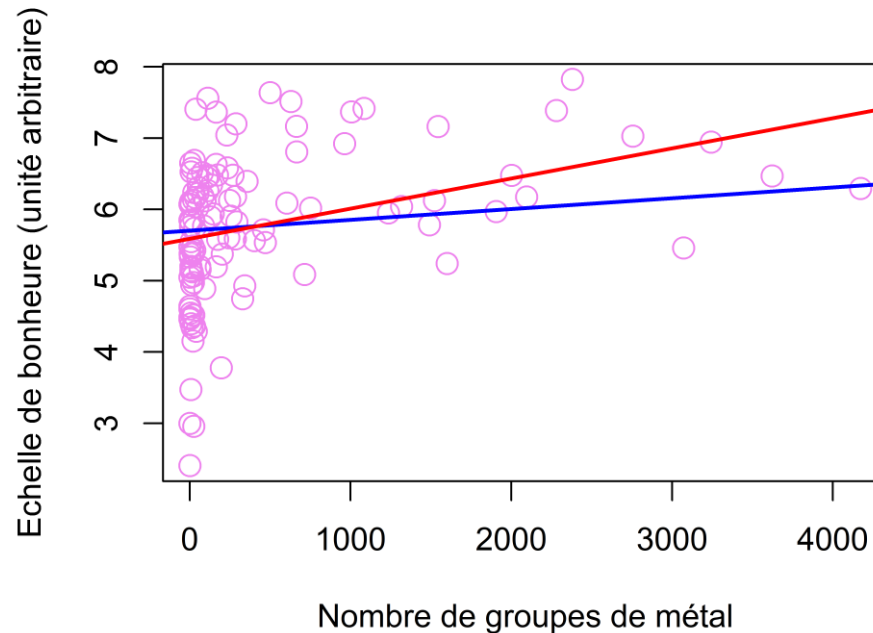
# Corrélation de Pearson

- Absence d'outliers : oui (puisque'on le supprime)

```
DF_stat2 <- DF_stat[!DF_stat$Territory %in% c("United States", "Germany"),]  
# exclusion
```


```
plot(x = DF_stat2$Bands, y = DF_stat2$Happiness)
```

```
CORR_MOD2 <- lm(Happiness ~ Bands, DF_stat2) # modèle linéaire sans l'outlier  
abline(CORR_MOD, col = "blue", lwd=2)  
abline(CORR_MOD2, col = "red", lwd=2)
```



**corrélation plus  
forte sans USA  
et Allemagne**

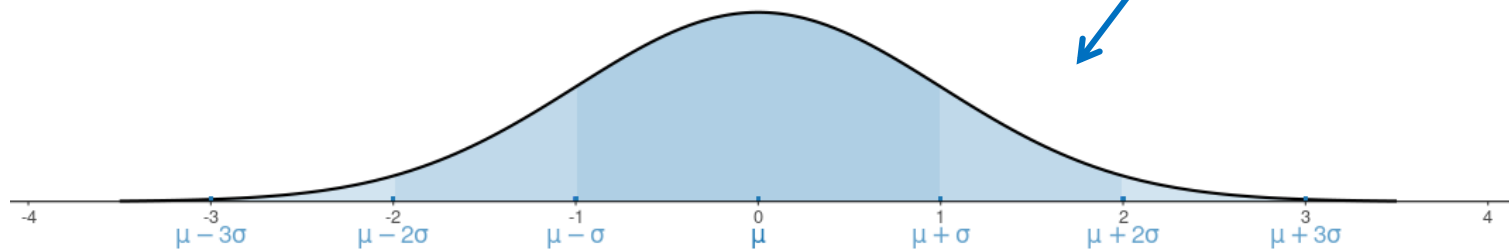
# Corrélation de Pearson

-  Conditions d'application
- Les distributions des 2 variables sont **approximativement** normales

<https://istats.shinyapps.io/NormalDist/>

## Normal Distribution

Mean :  $\mu = 0$ , Standard Deviation :  $\sigma = 1$



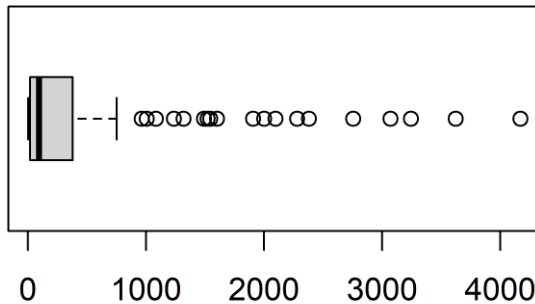
dans l'idéal on aimerait que nos variables soient distribuées de cette façon



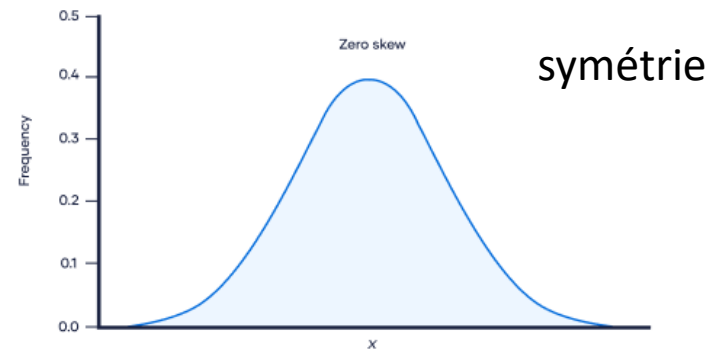
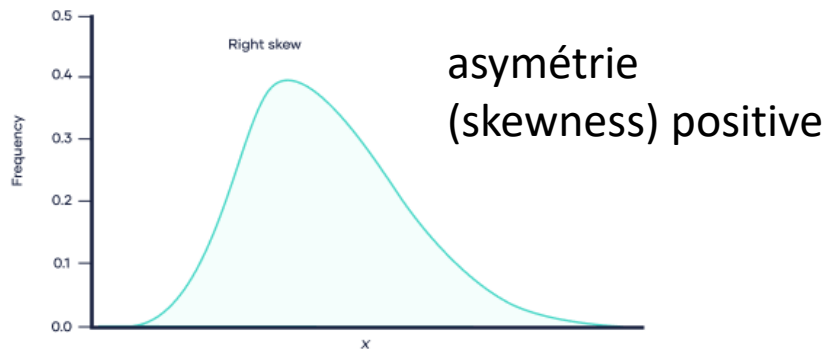
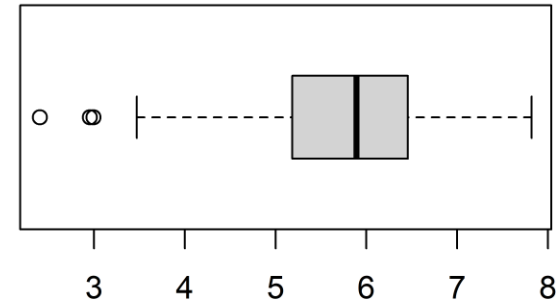
# Corrélation de Pearson

- Les distributions des variables sont **approximativement** normales
- (1) boxplots

nombre de groupes de métal



bonheur ressenti

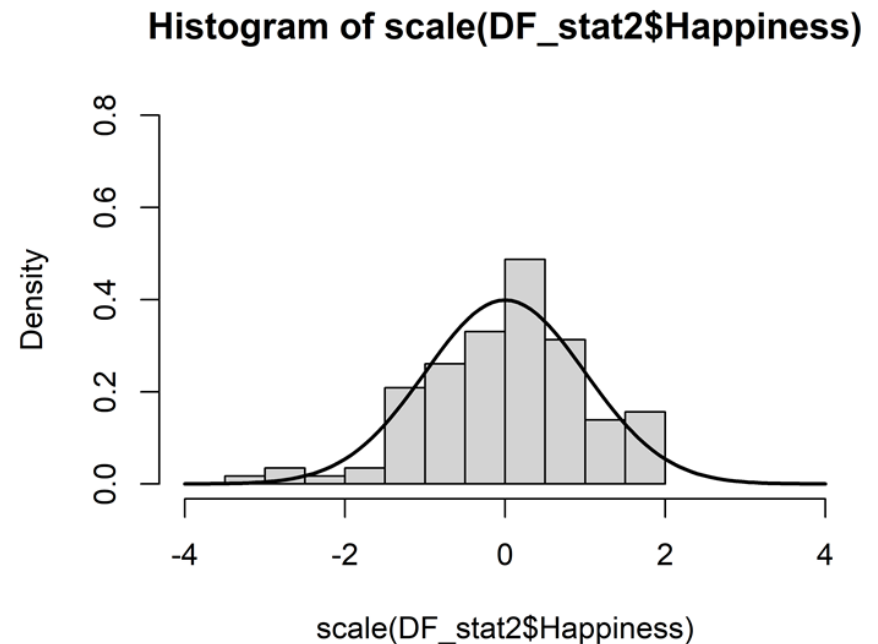
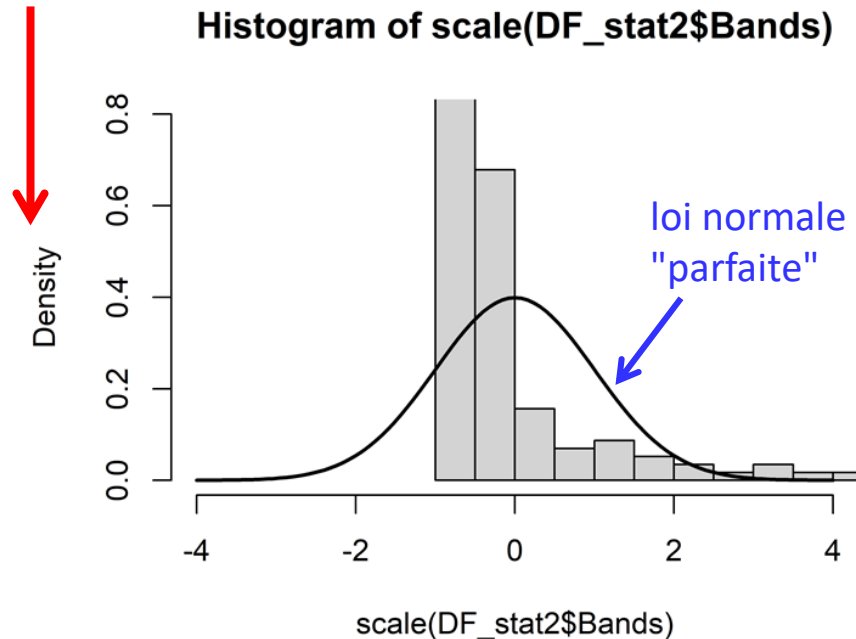


# Corrélation de Pearson


- Les distributions des variables sont **approximativement** normales
- (2) histogrammes** stratégie : centrer et réduire nos variables puis comparer leur histogramme avec une distribution normale

```
hist(scale(Bands),prob=T) # histogramme de densité au lieu de fréquence  
x <- scale(Bands)  
curve(dnorm(x),add=T, lwd=2)
```

aire totale sous la courbe = 1



# Corrélation de Pearson

- Possibilité d'utiliser des tests statistiques "formels" ...
- ...  mais toujours à croiser avec l'approche graphique

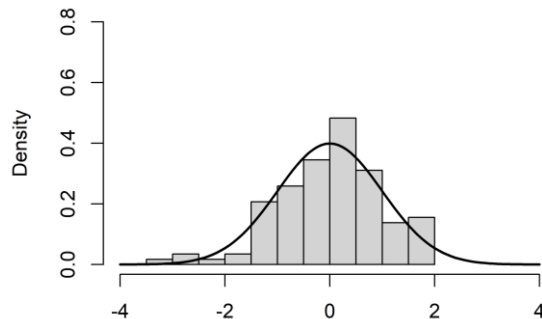
si  $n < 50$  : test de **Shapiro-Wilk**

**H0 : les données viennent d'une distribution normale**

```
shapiro.test(DF_stat2$Happiness)


##
##  Shapiro-Wilk normality test
##
## data:  DF_stat2$Happiness
## W = 0.9747, p-value = 0.02816
```

**Le Shapiro-Wilk est injustement sévère...**





# Corrélation de Pearson

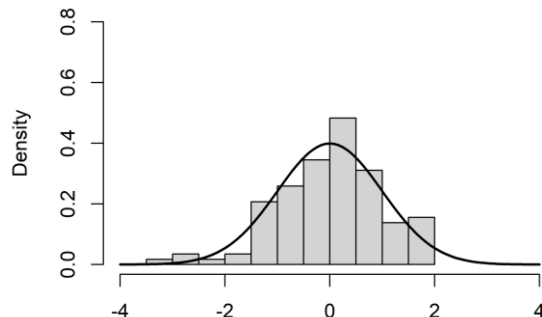
- Possibilité d'utiliser des tests statistiques "formels" ...
- ...  mais toujours à croiser avec l'approche graphique

si  $n < 50$  : test de **Shapiro-Wilk**

```
shapiro.test(DF_stat2$Happiness)

##
##  Shapiro-Wilk normality test
##
## data:  DF_stat2$Happiness
## W = 0.9747, p-value = 0.02705
```

**Le Shapiro-Wilk est injustement sévère...**



si  $n > 50$  : test de **Kolmogorov-Smirnov**

**H0 : les données viennent d'une distribution normale**

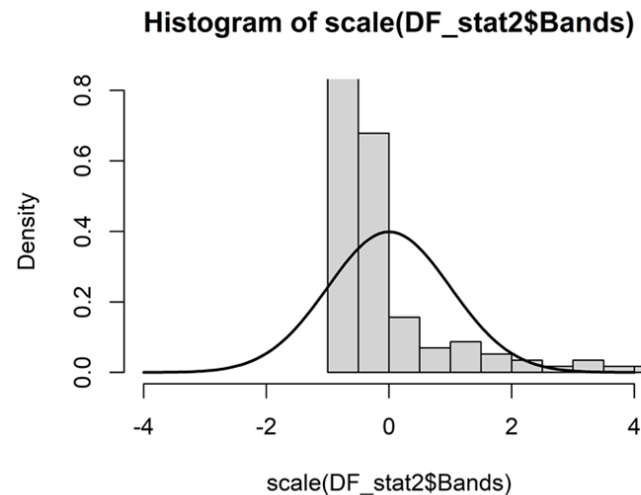
```
ks.test(scale(DF_stat2$Happiness),
        "pnorm", mean=0, sd =1)

## Asymptotic one-sample Kolmogorov-Smirnov
## test
##
## data:  scale(DF_stat2$Happiness)
## D = 0.058053, p-value = 0.8331
## alternative hypothesis: two-sided
```

Le test de **Kolmogorov-Smirnov est peu puissant**,  
donc "moins sévère" avec les distributions  
**approximativement normales**

# Corrélation de Pearson

- Absence d'outliers : oui (après suppression)
- Les distributions des variables sont **approximativement** normales : **NON !**



## Solutions ?

- (1) Forcer la variable à devenir normale ?!
- (2) Utiliser le test de corrélation de **Spearman** qui ne requiert pas ces 2 assumptions

# Corrélation de Pearson

- La transformation de Box-Cox

```
library(car)

# étape 1 : trouver Le paramètre Lambda qui permet La transformation optimale
LAMBDA <- car::powerTransform(DF_stat2$Bands ~ 1)

# étape 2 : appliquer La transformation de BoxCox
DF_stat2$Bands_BOXCOX <- car::bcPower(DF_stat2$Bands, LAMBDA$lambda)
```

```
fxhistnorm(DF_stat2$Bands_BOXCOX)

ks.test(scale(DF_stat2$Bands_BOXCOX),
mean=0, sd=1, "pnorm")
```

- Bands\_BOXCOX est-elle devenue normale ? vérifiez avec boxplot() et/ou hist()

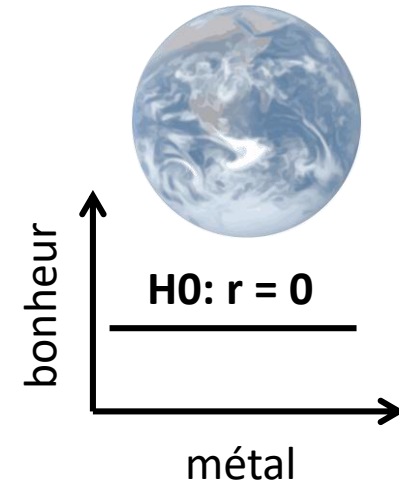
# Corrélation de Pearson

- Statistiques descriptives
- Le r de Pearson

```
cor(DF_stat2$Happiness,DF_stat2$Bands_BOXCOX)  
## [1] 0.5244629
```

**Est ce que cette corrélation de 0.52 permet de rejeter l'hypothèse nulle et être généralisée à TOUS les PAYS ?**

NB : les pays manquants dans notre échantillon sont ceux que nous avons exclus pour cause de données manquantes. Si nous avions eu accès aux données de TOUS les pays, nous n'aurions pas besoin d'un test d'inférence statistique !



représentation imagée de H0

# Corrélation de Pearson

- Statistiques inférentielles
- Test de corrélation de Pearson

variable 1

variable 2  
(transformée)

```
cor.test(DF_stat2$Happiness, DF_stat2$Bands_BOXCOX)

##
##  Pearson's product-moment correlation
##
## data:  DF_stat2$Happiness and DF_stat2$Bands_BOXCOX
## t = 6.7341, df = 114, p-value = 7.021e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3889778 0.6523339
## sample estimates:
##      cor
## 0.5334628
```

p valeur significative

r de Pearson

intervalle de confiance du r de Pearson

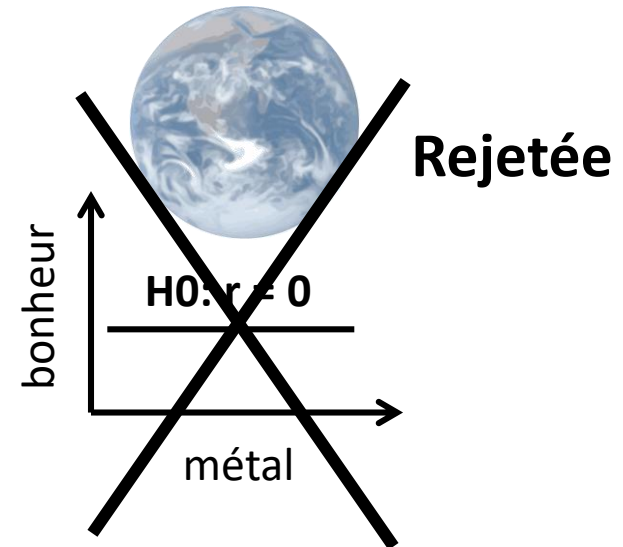
# Corrélation de Pearson

- Statistiques inférentielles
- Test de corrélation de Pearson

```
cor.test(DF_stat2$Happiness, DF_stat2$Bands_BOXCOX)

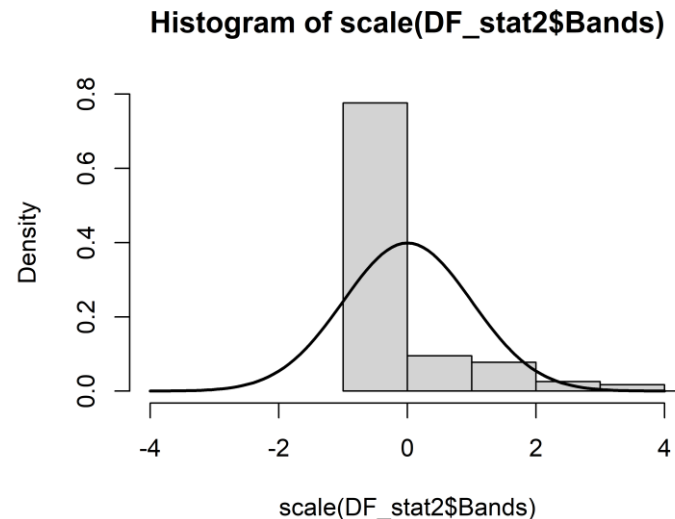
##
## Pearson's product-moment correlation
##
## data: DF_stat2$Happiness and DF_stat2$Bands_BOXCOX
## t = 6.7341, df = 114, p-value = 7.021e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3889778 0.6523339
## sample estimates:
## cor
## 0.5334628
```

Est ce que cette corrélation de 0.53 permet de rejeter l'hypothèse nulle et être généralisée à la population ?  
**OUI !!**



# Corrélation de Pearson

- Absence d'outliers : oui (après suppression)
- Les distributions des variables sont **approximativement** normales : **NON !**



## Solutions ?

- (1) Forcer la variable à devenir normale ?!
- (2) Utiliser le test de corrélation de Spearman qui ne requiert pas ces 2 assumptions

# Corrélation de Spearman

- équivaut à une corrélation de Pearson sur 2 variables **transformées en rangs**
- **teste une relation MONOTONE** (si x augmente, y augmente)
- test légèrement moins puissant (p valeur plus grosse)

par défaut, méthode = "pearson"

```
cor.test(DF_stat$Happiness, DF_stat$Bands, method="spearman")  
## Spearman's rank correlation rho  
##  
## data: DF_stat$Happiness and DF_stat$Bands  
## S = 125768, p-value = 8.858e-10  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
##      rho  
## 0.5288108
```

p valeur

rho de Spearman



# Corrélation

- Exemples de rédaction

Nous avons conduit une corrélation de **Pearson** pour tester l'existence d'une association (**linéaire**) entre le nombre de groupes de métal (GM) et le bonheur ressenti (BR) par pays. La relation entre GM et BR était significative ( **$r(113) = .52, p < .001$** ).

 **ddl = n - 2**

Nous avons conduit une corrélation de **Spearman** pour tester l'existence d'une association (**monotone**) entre le nombre de groupes de métal (GM) et le bonheur ressenti (BR) par pays. La relation entre GM et BR était significative ( **$r_s(114) = .53, p < .001$** ).

 **NB : pas de 0 avant virgule si chiffre borné entre -1 et +1**

# Corrélation

- Mémo

VI-VD	stat descriptive	stat inférentielle
2 VD/VI ordinaire intervalle rapport (quantitative)	plot() lm() abline() cor()	cor.test()

# Exercice 6

- Importez "reading\_skills2.csv"
- Tester l'hypothèse d'une corrélation entre les variables "iq" et "age"
- Tester l'hypothèse d'une corrélation entre les variables "iq" et "accuracy"
- **Suivez les étapes :**
  - graphique bivarié
  - outlier ?
  - normalité des distributions ? (si besoin, transformez la variable en question)
  - test inférentiel : Pearson ou Spearman ?
- Question de statistiques bonus : Timéha trouve une corrélation de **0.01** entre le **nombre de groupes de jazz par pays** et le **bonheur ressenti par pays**. Elle a les données de tous les pays. **Dans son article Timéha rejette H0 sans même calculer de p-valeur...** S'est-elle trompée !?