SCI03 - Analyse de données expérimentales

Tests d'hypothèses

Thierry Denœux

Automne 2014

Plan

1. Tests d'hypothèses : Principes généraux
2. Tests de conformité
3. Tests d'homogénéité
4. Tests d'indépendance

Première partie I

Tests d'hypothèses : Principes généraux

Plan

Exemple introductif
- Exposé du problème
- Première approche

Résolution d'un problème de test
- Cadre général
- Principaux problèmes de test

Influence des métaux lourds sur la santé

- L'exposition au métaux lourds augmente-t-elle le taux de cholestérol ?
- Taux de cholestrol (en mgm%) observés chez 137 ouvriers travaillant dans une usine de zinc en Tasmanie :

```
[1] 145 174 180 196 204 208 214 221 227 238 247 256 268 294 145 175 181 198
[19] 205 209 216 221 227 239 248 257 268 296 146 175 181 198 205 211 217 221
[37] 228 241 252 257 273 300 158 175 187 198 205 211 217 221 230 242 253 261
[55] 274 301 158 175 187 198 205 211 217 223 235 242 253 262 278 302 163 175
[73] 192 201 205 212 217 224 235 243 253 266 283 314 168 175 194 201 206 212
[91] 218 225 235 243 254 267 284 331 168 178 194 201 206 212 218 225 236 345
[109] 254 267 285 168 180 195 204 206 214 218 227 237 245 255 268 289 172 180
[127] 196 204 207 214 218 227 238 245 256 268 292
```

- Le taux de cholestérol moyen dans la population masculine de Tasmanie est de 218 mgm%.
- Que peut-on en conclure ?

Plan

Exemple introductif
- Exposé du problème
- Première approche

Résolution d'un problème de test
- Cadre général
- Principaux problèmes de test

Etude élémentaire de l'échantillon

- Résumé de l'échantillon :

```
> summary(chol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  145.0   198.0   218.0   224.4   253.0   345.0
```

- La moyenne dans l'échantillon est supérieure à la moyenne dans la population totale, mais il y a une grande variabilité des observations dans l'échantillon...
- L'écart à la valeur de référence (218) est-il accidentel, ou peut-il être attribué à l'effet de l'exposition au zinc ? L'écart est-il significatif ?

Méthode de résolution

- Soit X le taux de cholestérol d'un ouvrier pris au hasard.
- On suppose que $X \sim \mathcal{N}(\mu, \sigma^2)$.
- Si l'exposition au zinc n'a pas d'effet sur le taux de cholestérol, alors $\mu = 218$. C'est l'hypothèse nulle, notée H_0.
- Si l'exposition au zinc a un effet, alors $\mu > 218$: c'est l'hypothèse alternative notée H_1.
- On cherche une statistique de loi de probabilité connue sous l'hypothèse H_0, et dont la loi sous H_1 s'écarte de la loi sous H_0 dans un sens prévisible.
- Si cette statistique prend une valeur « étonnamment» grande ou petite sous l'hypothèse H_0, on décidera de rejeter cette hypothèse.

La statistique t

- On choisit comme statistique la différence normalisée entre la moyenne théorique $\mu_0 = 218$ et la moyenne empirique de l'échantillon :

$$T = \frac{\overline{X} - \mu_0}{S^*/\sqrt{n}}$$

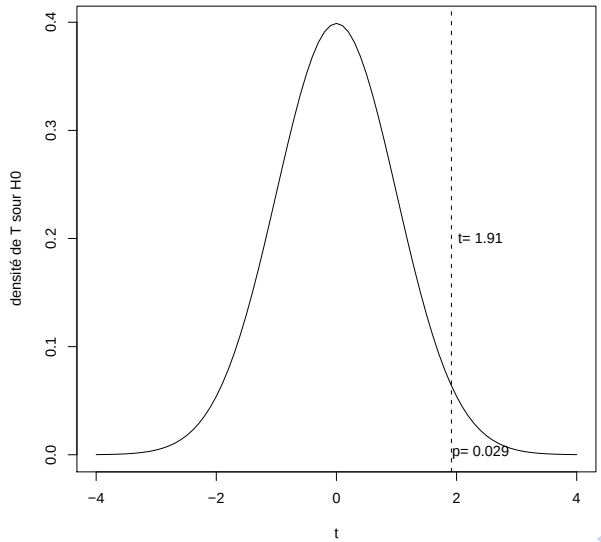
avec $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ et

$$S^* = \left(\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \right)^{1/2}.$$

- Sous H_0, T suit une loi de Student à $n-1$ degrés de liberté. Comme $n \geq 30$, cette loi est très bien approchée par une loi normale $\mathcal{N}(0,1)$.
- Sous H_1, T a tendance à prendre des valeurs plus grandes.

Principe du test t

- La valeur prise par la statistique T est $t = 1.91$. Si H_0 est vraie, alors la probabilité d'observer une valeur au moins aussi élevée est $p = 0.029$.
- Cette valeur étant assez petite, il est permis de mettre en doute l'hypothèse H_0. Si on s'était fixé comme seuil une probabilité de 0.05, on dit qu'on rejette H_0 au niveau de signification de 5 %.

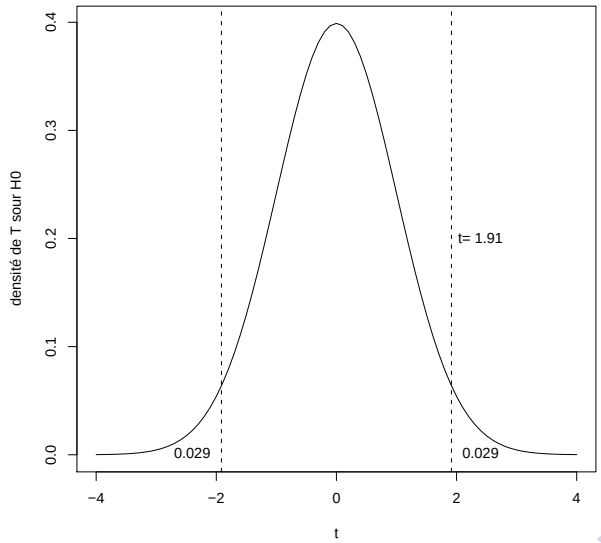


Importance du choix de l'hypothèse alternative

- En prenant comme hypothèse alternative $H_1 : \mu > \mu_0$, on a implicitement supposé que l'effet d'une exposition au zinc ne peut être qu'une augmentation du taux de cholestérol. On dit qu'on a fait un test unilatéral.
- Si on admet qu'un effet contraire est également possible, il faut prendre comme hypothèse alternative $H_1 : \mu \neq \mu_0$.
- Dans ce cas, le motif de rejet de H_0 est une valeur de T « étonnamment » grande en valeur absolue $\rightarrow$ test bilatéral.

Test t bilatéral

- La valeur prise par la statistique T est $t = 1.91$. Si H_0 est vraie, alors la probabilité d'observer une valeur au moins aussi élevée en valeur absolue est $p = 2 \times 0.029 = 0.058$.
- Cette valeur est encore assez petite, mais deux fois plus grande que précédemment. Elle est maintenant tout juste supérieure à 0.05 : on ne rejette pas H_0 au niveau de signification de 5 %.



Plan

Exemple introductif
- Exposé du problème
- Première approche

Résolution d'un problème de test
- Cadre général
- Principaux problèmes de test

Définition d'un problème de test

- Un problème de test est une question que l'on se pose sur une ou plusieurs populations, par exemple :
  - La moyenne d'une variable X dans la population est-elle égale à une valeur donnée ?
  - Les proportions d'un certain caractère qualitatif dans deux populations sont-elles égales ou différentes ?
- On répond à cette question sur la base d'un ou plusieurs échantillons aléatoires obtenus par échantillonnage dans la ou les populations concernées.
- Formellement, on définit une hypothèse nulle H_0 et une hypothèse alternative H_1.
- Ces hypothèses sont supposées exclusives et exhaustives (une et une seule est vraie).

Statistique de test et degré de signification

- Pour résoudre un problème de test, on définit une statistique de test T dont on connaît la loi sous H_0.
- On définit le degré de signification p (*p-value*) comme la probabilité que T prenne une valeur au moins aussi extrême que la valeur observée t.
- La notion de « valeur extrême » dans la définition précédente dépend de l'hypothèse alternative H_1. Par exemple, si on s'attend à observer sous H_1 des valeurs plus grandes de T, on définit le degré de signification comme : $p = \mathbb{P}_{H_0}(T \geq t)$.
- Plus le degré de signification est petit, plus les observations contredisent l'hypothèse H_0.
- Au contraire, si le degré de signification est grand, cela n'accrédite pas forcément H_0 (on ne peut rien dire).

Niveau de signification

- Pour prendre une décision, on se fixe un niveau de signification α^* (le plus souvent : 1%, 5 % ou 10 %) et on compare le degré de signification p à α^*.
- Si $p \leq \alpha^*$, on rejette H_0 au niveau de signification α^*. Sinon, on ne rejette pas H_0.
- Terminologie :
  - $p \leq 0.01$: test très significatif ;
  - $0.01 < p \leq 0.05$: test significatif ;
  - $0.05 < p \leq 0.1$: test faiblement significatif.

Plan

Exemple introductif
- Exposé du problème
- Première approche

Résolution d'un problème de test
- Cadre général
- Principaux problèmes de test

Typologie des problèmes de test

- Tests de conformité : comparaison d'un paramètre à une valeur théorique, sur la base d'un seul échantillon :
  - Moyenne, médiane, variance, proportion, etc ;
  - Test d'adéquation : la variable considérée suit-elle une certaine loi (par exemple : loi normale) ?
- Tests d'homogénéité : comparaison des moyennes, variances, etc. dans deux ou plusieurs échantillons :
  - Echantillons appariés (deux grandeurs observées sur les mêmes individus, par exemple avant et après traitement) ;
  - Echantillons indépendants.
- Tests d'une liaison entre deux variables :
  - Tests sur la valeur d'un coefficient de corrélation ;
  - Tests d'indépendance.

Choix d'une méthode de test

- Type de données disponibles :
  - Un échantillon ;
  - Deux échantillons appariés ;
  - K échantillons indépendants, ...
- Traduire une question sous forme d'une hypothèse nulle H_0 et une hypothèse alternative H_1.
- Vérifier les conditions d'applicabilité du test :
  - Indépendance de l'échantillon ;
  - Normalité ;
  - Taille de l'échantillon, ...

Deuxième partie II

Tests de conformité

Plan

Test sur une moyenne

Test sur une médiane

Test sur une proportion

Tests d'adéquation

Test de Student

- Modèle : $X_1, \ldots, X_n$ échantillon i.i.d. de variable parente $X \sim \mathcal{N}(\mu, \sigma^2)$.
- Hypothèses :
  - $H_0 : \mu = \mu_0$
  - $H_1 : \mu \neq \mu_0$, $\mu < \mu_0$ ou $\mu > \mu_0$.
- Statistique de test :

$$T = \frac{\overline{X} - \mu_0}{S^*/\sqrt{n}} \overset{H_0}{\sim} \mathcal{T}_{n-1}.$$

- Degré de signification :

$$p = \begin{cases} \mathbb{P}_{H_0}(T \geq t) & \text{si } H_1 : \mu > \mu_0 \\ \mathbb{P}_{H_0}(T \leq t) & \text{si } H_1 : \mu < \mu_0 \\ \mathbb{P}_{H_0}(|T| \geq |t|) & \text{si } H_1 : \mu \neq \mu_0. \end{cases}$$

Test de Student : conditions d'application

- Le test s'applique en toute rigueur dans le cas où X suit une loi normale. Il est nécessaire de tester l'hypothèse de normalité si $n < 30$ (cf. test de normalité).
- Il peut également s'appliquer quelle que soit la loi de X si $n \geq 30$.
  - Le degré de signification calculé est alors approximatif (approximation généralement très bien vérifiée).
  - On peut dans ce cas remplacer la loi de Student par la loi normale $\mathcal{N}(0, 1)$.

Le test de Student en R

- Exemple du taux de cholestérol (test unilatéral) :

```
> t.test(chol,mu=218,alt="g")
One Sample t-test
data: chol
t = 1.9145, df = 136, p-value = 0.02883
alternative hypothesis: true mean is greater than 218
95 percent confidence interval:
218.8697 Inf
sample estimates:
mean of x
224.4453
```

- Paramètres : `alt` ("g" : greater ; "l" : lower ; "t" : two-sided, défaut) ; `conf.level` (défaut : 0.95).

Médiane et test non paramétrique

- Soit X une v.a. continue. La médiane est la valeur m telle que $\mathbb{P}(X \leq m) = 0.5$.
- Contrairement à la moyenne, la médiane conserve une signification claire quelle soit la forme de la distribution de X (en particulier, même si celle-ci est fortement dissymétrique).
- Le test de Wilcoxon signé que nous allons présenter est un test non paramétrique, c'est-à-dire qu'il ne repose pas sur une famille de lois paramétrées comme la loi normale.
- Les tests non paramétriques sont d'application plus générale que les tests paramétriques, mais ils sont généralement moins puissants que les tests paramétriques lorsque ceux-ci s'appliquent (ils détectent moins souvent le fait que H_0 n'est pas vérifiée).
- Les tests non paramétriques s'utilisent souvent dans le cas de petits échantillons.

Test de Wilcoxon signé

- Modèle : $X_1, \ldots, X_n$ échantillon i.i.d. de variable parente X de loi de probabilité continue et symétrique.
- Hypothèses :
  - $H_0 : m = m_0$
  - $H_1 : m \neq m_0$, $m < m_0$ ou $m > m_0$.
- Calcul de la statistique de test :
  - On ordonne les observations par ordre de distance croissante à m_0
  - $V =$ somme des rangs des observations supérieures à m_0 :

$$V = \sum_{X_i > m_0} \text{rang}\left(|X_i - m_0|\right).$$

- Une grande valeur de V accrédite l'hypothèse alternative $H_1 : m > m_0$.

Exemple : consommation énergétique

- Données : consommations énergétiques de 11 femmes (en kJ) :

```
> conso
[1] 5260 5470 5640 6180 6390 6515 6805 7515 7520 8230 8770

> stem(conso)
The decimal point is 3 digit(s) to the right of the |
5 | 356
6 | 2458
7 | 55
8 | 28
```

- La consommation médiane est-elle différente de la ration calorique recommandée de 7725 kJ ?

Test de Wilcoxon signé en R

- Test en R :

```
> wilcox.test(conso,mu=7725)

Wilcoxon signed rank test

data: conso
V = 8, p-value = 0.02441
alternative hypothesis: true mu is not equal to 7725
```

- On rejette H_0 au niveau de 5% (résultat significatif).
- Paramètre : `alt=` "g" (greater) ; "l" (lower) ; "t" (two-sided, défaut).

Exemple

- Dans les affaires de divorce, la garde est accordée à la mère dans 70% des cas.
- Un juge a accordé la garde à la mère dans 12 cas sur 30 affaires.
- Peut-on mettre en cause l'impartialité du juge ?

Formalisation du problème

- Pour une affaire de divorce prise au hasard, soit X la v.a. définie par :

$$X = \begin{cases} 1 & \text{si le juge accorde la garde à la mère} \\ 0 & \text{sinon.} \end{cases}$$

- Soit π la probabilité que le juge donne la garde à la mère. X suit une loi de Bernoulli de paramètre π : $X \sim \mathcal{B}(\pi)$.
- Hypothèse $H_0 : \pi = \pi_0 = 0.7$ (le juge est impartial).
- Hypothèse $H_1 : \pi < \pi_0$ (le juge est partial en faveur des pères).

Résolution en R

```
> binom.test(12,30,0.7,alt="l")

Exact binomial test

data: 12 and 30
number of successes = 12, number of trials = 30, p-value = 0.000626
alternative hypothesis: true probability of success is less than
0.7
95 percent confidence interval:
0.0000000 0.5660547
sample estimates:
probability of success
0.4
```

Problème posé

- Beaucoup de méthodes statistiques reposent sur l'hypothèse que la loi de la variable d'intérêt X appartient à une famille de lois, par exemple : la famille des lois normales $\mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$.
- Les tests d'adéquation permettent de tester ce type d'hypothèse.
- Nous nous contenterons d'étudier le cas de l'adéquation à une loi normale (test de normalité).
- L'hypothèse nulle est alors

$$H_0 : X \sim \mathcal{N}(\mu, \sigma^2), \quad (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+.$$

- L'hypothèse H_1 est simplement la négation de H_0.

Méthode graphique : diagramme de normalité

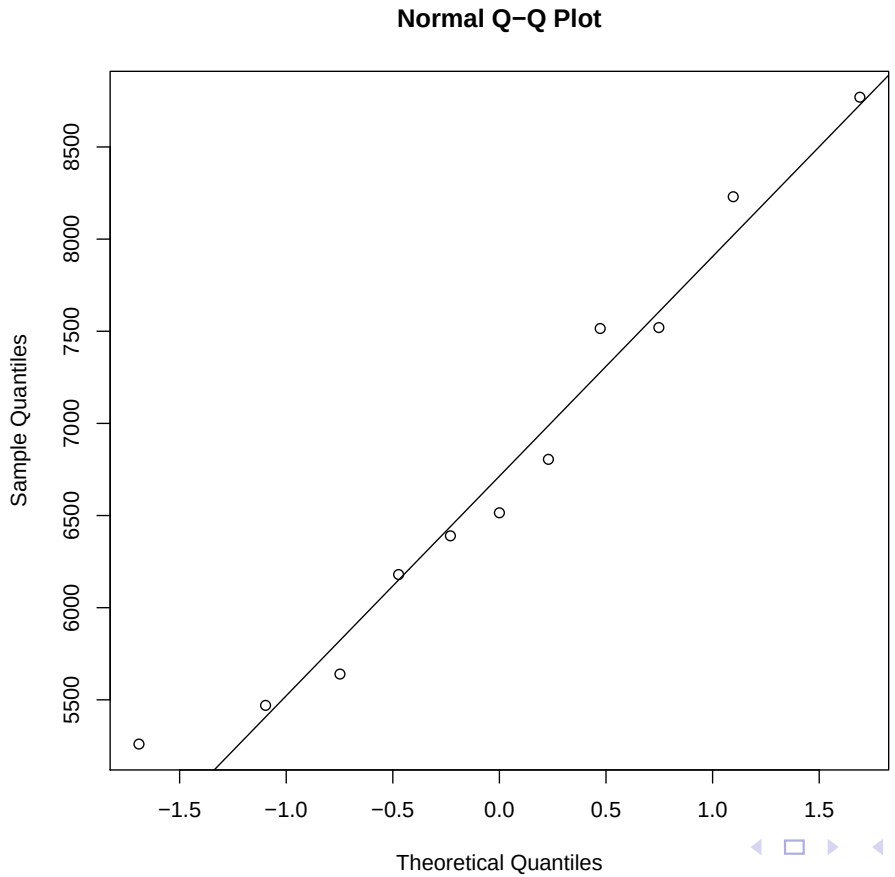
- Soient $x_{(1)}, \ldots, x_{(n)}$ les n observations de l'échantillon triées par ordre croissant.
- Soit $a_n(i)$ l'espérance de la i-ème observation dans un échantillon de taille n extrait de la loi $\mathcal{N}(0,1)$ (scores normaux).
- Si X suit une loi normale, on doit avoir

$$\frac{x_{(i)} - \overline{x}}{s^*} \approx a_n(i)$$

  c'est-à-dire $x_{(i)} \approx a_n(i)s^* + \overline{x}$.
- Il suffit donc de représenter les points $(a_n(i), x_{(i)})$ sur un graphe, et de vérifier si les points sont approximativement alignés.

Exemple : données de consommation calorique

```
> qqnorm(conso)
> qqline(conso)
```



Test de Shapiro-Wilk

- Basé sur une statistique W fonction des observations ordonnées et des scores normaux (même idée que dans le Q-Q plot).
- Des simulations ont montré que W prend en moyenne des valeurs plus petites dans le cas non normal que dans le cas normal.
- Degré de signification : $p = \mathbb{P}_{H_0}(W \leq w)$.
- Exemple :

```
> shapiro.test(conso)

Shapiro-Wilk normality test

data: conso
W = 0.9524, p-value = 0.675
```

Troisième partie III

Tests d'homogénéité

Plan

Echantillons appariés
- Exemple introductif
- Test de Student
- Test de Wilcoxon signé

Deux échantillons indépendants
- Comparaison de deux variances
- Comparaison de deux moyennes
- Test de Wilcoxon
- Comparaison de deux proportions

K échantillons indépendants
- Comparaison de K moyennes
- Comparaison de K variances
- Test de Kruskall-Wallis

Plan

Echantillons appariés
- Exemple introductif
- Test de Student
- Test de Wilcoxon signé

Deux échantillons indépendants
- Comparaison de deux variances
- Comparaison de deux moyennes
- Test de Wilcoxon
- Comparaison de deux proportions

K échantillons indépendants
- Comparaison de K moyennes
- Comparaison de K variances
- Test de Kruskall-Wallis

Influence du vendredi 13 sur les comportements

- Source : Scanlon, T.J., Luben, R.N., Scanlon, F.L., Singleton, N. (1993), "Is Friday the 13th Bad For Your Health ?," BMJ, 307, 1584-1586.
- Données du Ministère Britanique des Transports :

date	Vendredi 6	Vendredi 13	routes
1990, July	139246	138548	7 to 8
1990, July	134012	132908	9 to 10
1991, September	137055	136018	7 to 8
1991, September	133732	131843	9 to 10
1991, December	123552	121641	7 to 8
1991, December	121139	118723	9 to 10
1992, March	128293	125532	7 to 8
1992, March	124631	120249	9 to 10
1992, November	124609	122770	7 to 8
1992, November	117584	117263	9 to 10

Plan

Echantillons appariés
- Exemple introductif
- Test de Student
- Test de Wilcoxon signé

Deux échantillons indépendants
- Comparaison de deux variances
- Comparaison de deux moyennes
- Test de Wilcoxon
- Comparaison de deux proportions

K échantillons indépendants
- Comparaison de K moyennes
- Comparaison de K variances
- Test de Kruskall-Wallis

Test de Student (échantillons appariés)

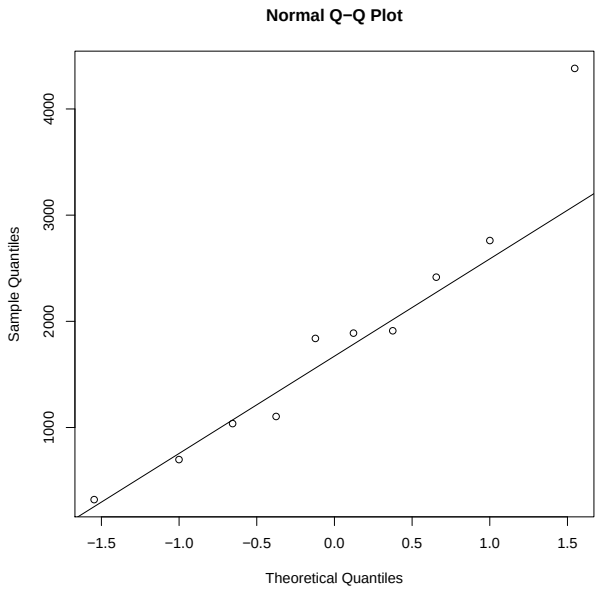
- Modèle : $(X_1, Y_1), \ldots, (X_n, Y_n)$ échantillon i.i.d. d'un couple de variables aléatoires (X, Y).
- Soit $D = X - Y$. On suppose $D \sim \mathcal{N}(\mu_D, \sigma^2)$, avec $\mu_D = \mu_X - \mu_Y$.
- Hypothèses :
  - $H_0 : \mu_D = 0$ ($\Leftrightarrow \mu_X = \mu_Y$)
  - $H_1 : \mu_D \neq 0$ ($\mu_X \neq \mu_Y$), $\mu_D < 0$ ($\mu_X < \mu_Y$) ou $\mu_D > 0$ ($\mu_X > \mu_Y$).
- On fait un test de Student sur l'échantillon des différences $D_1, \ldots, D_n$ avec $D_i = X_i - Y_i$.

Vendredi 13 : test de normalité de D

```
> shapiro.test(v6-v13)

Shapiro-Wilk normality test

data: v6 - v13
W = 0.9316, p-value = 0.4636
```



Vendredi 13 : test de Student

```
> t.test(v6,v13,paired=T)

Paired t-test

data: v6 and v13
t = 4.9364, df = 9, p-value = 0.0008062
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
994.5304 2677.0696
sample estimates:
mean of the differences
1835.8
```

Plan

Echantillons appariés
- Exemple introductif
- Test de Student
- Test de Wilcoxon signé

Deux échantillons indépendants
- Comparaison de deux variances
- Comparaison de deux moyennes
- Test de Wilcoxon
- Comparaison de deux proportions

K échantillons indépendants
- Comparaison de K moyennes
- Comparaison de K variances
- Test de Kruskall-Wallis

Test de Wilcoxon signé (échantillons appariés)

- Modèle : $(X_1, Y_1), \ldots, (X_n, Y_n)$ échantillon i.i.d. d'un couple de variables aléatoires (X, Y).
- Soit $D = X - Y$. On suppose que la loi de D est continue et symétrique, de médiane m.
- Hypothèses :
  - $H_0 : m = 0$ ($\Leftrightarrow \mathbb{P}(X \leq Y) = 1/2$)
  - $H_1 : m \neq 0$ ($\mathbb{P}(X \leq Y) \neq 1/2$), $m < 0$ ($\mathbb{P}(X \leq Y) > 1/2$) ou $m > 0$ ($\mathbb{P}(X \leq Y) < 1/2$).
- On fait un test de Wilcoxon signé sur l'échantillon des différences $D_1, \ldots, D_n$ avec $D_i = X_i - Y_i$.

Vendredi 13 : test de Wilcoxon signé

```
> wilcox.test(v6,v13,paired=T)

Wilcoxon signed rank test

data: v6 and v13
V = 55, p-value = 0.001953
alternative hypothesis: true location shift is not equal to 0
```

Plan

Echantillons appariés
- Exemple introductif
- Test de Student
- Test de Wilcoxon signé

Deux échantillons indépendants
- Comparaison de deux variances
- Comparaison de deux moyennes
- Test de Wilcoxon
- Comparaison de deux proportions

K échantillons indépendants
- Comparaison de K moyennes
- Comparaison de K variances
- Test de Kruskall-Wallis

Exemple : dosage de l'azote

- Deux méthodes de dosage de l'azote ont été répétées, à partir d'un même échantillon, 25 fois avec la méthode A, 30 fois avec la méthode B. On veut comparer la variabilité (précision) des deux méthodes.
- Données :

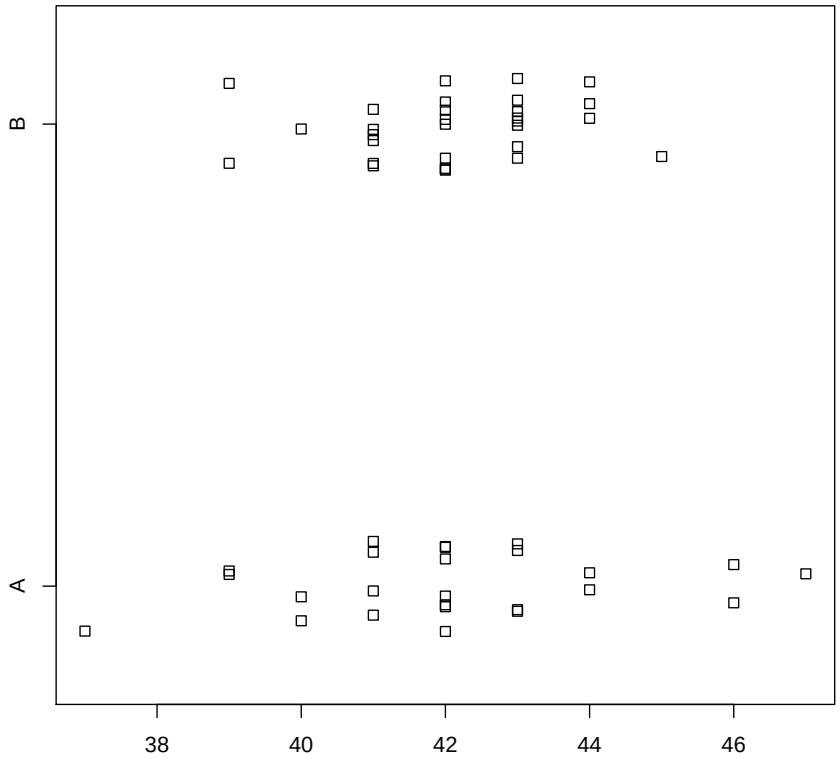
```
> A
[1] 37 39 39 40 40 41 41 41 41 42 42 42 42 42 42 42 43
[18] 43 43 43 44 44 46 46 47
> B
[1] 39 39 40 41 41 41 41 41 41 42 42 42 42 42 42 42 42
[18] 42 43 43 43 43 43 43 43 43 44 44 44 45
```

Mise en forme des données

```
> dosage=c(A,B)
> methode=c(rep("A",25),rep("B",30))
> dosage
[1] 37 39 39 40 40 41 41 41 41 42 42 42 42 42 42 42 43
[18] 43 43 43 44 44 46 46 47 39 39 40 41 41 41 41 41 41
[35] 42 42 42 42 42 42 42 42 42 43 43 43 43 43 43 43 43
[52] 44 44 44 45
> methode
[1] "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A"
[14] "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "B"
[27] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"
[40] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"
[53] "B" "B" "B"
```

Représentation des données

```
> stripchart(dosage ~ methode,method="jitter")
```



Test de Fisher

- Modèle : $X_1, \ldots, X_{n_X}$ échantillon i.i.d. de X, $Y_1, \ldots, Y_{n_Y}$ échantillon i.i.d. de Y. Les deux échantillons sont indépendants.
- On suppose $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ et $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.
- Hypothèses :
  - $H_0 : \sigma_X^2 = \sigma_Y^2$
  - $H_1 : \sigma_X^2 \neq \sigma_Y^2$, $\sigma_X^2 > \sigma_Y^2$ ou $\sigma_X^2 < \sigma_Y^2$.
- Statistique de test :

$$F = \frac{S_X^{*2}}{S_Y^{*2}} \overset{H_0}{\sim} \mathcal{F}(n_X - 1, n_Y - 1).$$

Test de Fisher en R

```
> var.test(dosage ~ methode)

F test to compare two variances

data: dosage by methode
F = 2.6392, num df = 24, denom df = 29, p-value =
0.01367
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
1.225230 5.852172
sample estimates:
ratio of variances
2.639153
```

Plan

Echantillons appariés
- Exemple introductif
- Test de Student
- Test de Wilcoxon signé

Deux échantillons indépendants
- Comparaison de deux variances
- Comparaison de deux moyennes
- Test de Wilcoxon
- Comparaison de deux proportions

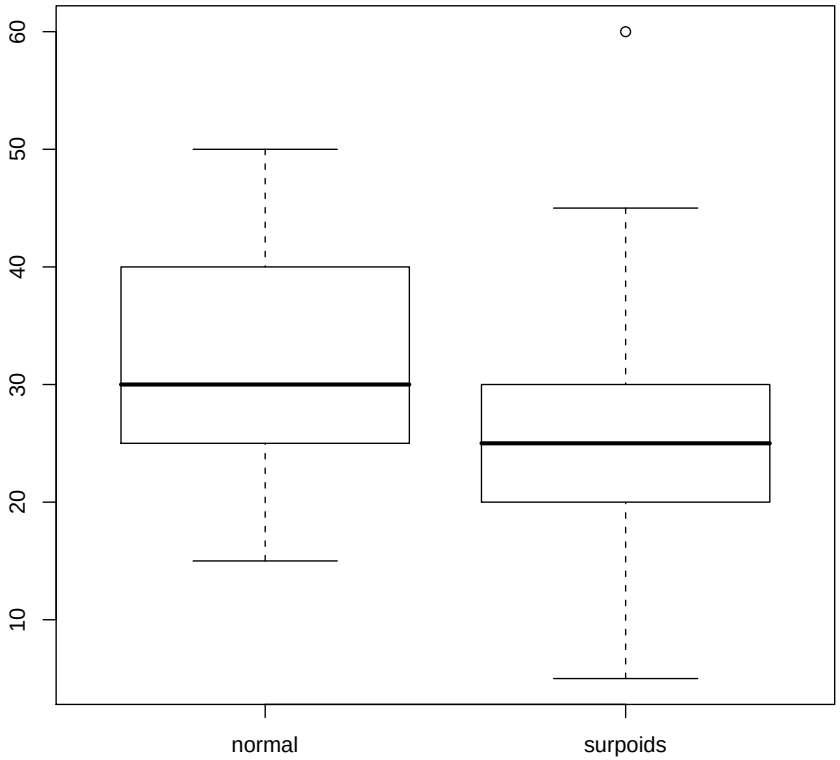
K échantillons indépendants
- Comparaison de K moyennes
- Comparaison de K variances
- Test de Kruskall-Wallis

Réaction des médecins au poids des patients

- Question : les patients obèses sont-ils victimes de discrimination lors des consultations médicales ?
- Cadre expérimental
  - $n_X = 33$ médecins reçoivent le dossier médical d'un patient de poids normal
  - $n_Y = 38$ médecins reçoivent le dossier médical d'un patient présentant un surpoids (mêmes symptômes : migraine)
  - questionnaire : estimation du temps devant être passé lors d'une consultation avec le patient.

Données

`boxplot(time ~ poids)`



Test de Student (2 échantillons indépendants)

- Modèle : $X_1, \ldots, X_{n_X}$ échantillon i.i.d. de X, $Y_1, \ldots, Y_{n_Y}$ échantillon i.i.d. de Y. Les deux échantillons sont indépendants.
- On suppose $X \sim \mathcal{N}(\mu_X, \sigma^2)$ et $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$ (les variances sont égales).
- Hypothèses :
  - $H_0 : \mu_X = \mu_Y$
  - $H_1 : \mu_X \neq \mu_Y$, $\mu_X < \mu_Y$ ou $\mu_X > \mu_Y$.
- Statistique de test :

$$T = \frac{|\overline{X} - \overline{Y}|}{S^* \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \overset{H_0}{\sim} \mathcal{T}_{n_X+n_Y-2}$$

avec $S^{*2} = \left((n_X - 1)S_X^{*2} + (n_Y - 1)S_Y^{*2}\right) / (n_X + n_Y - 2)$.

Test d'homogénéité des variances

```
> var.test(time ~ poids)

F test to compare two variances

data: time by poids
F = 1.0443, num df = 32, denom df = 37, p-value = 0.893
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.5333405 2.0797269
sample estimates:
ratio of variances
1.044316
```

Test de Student en R

```
> t.test(time ~ poids,var.equal=T)

Two Sample t-test

data: time by poids
t = 2.856, df = 69, p-value = 0.005663
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
1.997955 11.255633
sample estimates:
mean in group normal mean in group surpoids
31.36364 24.73684
```

Test de Student : conditions d'application

- L'hypothèse de normalité n'est plus nécessaire lorsque $n_X \geq 30$ et $n_Y \geq 30$.
- Il existe une variante du test de Student adaptée au cas où les variances sont différentes (le degré de signification est alors approché). C'est l'option par défaut en R lorsqu'on ne précise pas `var.equal=T`.

Plan

Echantillons appariés
- Exemple introductif
- Test de Student
- Test de Wilcoxon signé

Deux échantillons indépendants
- Comparaison de deux variances
- Comparaison de deux moyennes
- Test de Wilcoxon
- Comparaison de deux proportions

K échantillons indépendants
- Comparaison de K moyennes
- Comparaison de K variances
- Test de Kruskall-Wallis

Test de Wilcoxon (2 échantillons indépendants)

- Modèle : $X_1, \ldots, X_{n_X}$ échantillon i.i.d. de X, $Y_1, \ldots, Y_{n_Y}$ échantillon i.i.d. de Y (v.a. continues). Les deux échantillons sont indépendants.
- On suppose qu'il existe une fonction de densité f de moyenne nulle, telle que $X \sim f(x - \mu_X)$ et $Y \sim f(y - \mu_Y)$ (les deux distributions sont « décalées »).
- Hypothèses :
  - $H_0 : \mu_X = \mu_Y$
  - $H_1 : \mu_X \neq \mu_Y$, $\mu_X < \mu_Y$ ou $\mu_X > \mu_Y$.
- Statistique de test : somme des rangs des observations du premier échantillon dans la série des $n_X + n_Y$ observations triées par ordre croissant, moins le minimum théorique.

Test de Wilcoxon en R (réaction au poids des patients)

```
> wilcox.test(time ~ poids)

Wilcoxon rank sum test with continuity correction
data: time by poids
W = 866, p-value = 0.003985
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(15L, 15L, 45L, 40L, 45L, 20L, 40L
impossible de calculer la p-value exacte avec des ex-aequos
```

Plan

Echantillons appariés
- Exemple introductif
- Test de Student
- Test de Wilcoxon signé

Deux échantillons indépendants
- Comparaison de deux variances
- Comparaison de deux moyennes
- Test de Wilcoxon
- Comparaison de deux proportions

K échantillons indépendants
- Comparaison de K moyennes
- Comparaison de K variances
- Test de Kruskall-Wallis

Exemple

- Dans un groupe de 200 malades, on a constitué par tirage au sort une série traitée (soumise à un nouveau traitement) et une série témoin (soumise au traitement classique). On a observé :
  - chez les traités (102 sujets) : 20 décès ;
  - chez les témoins (98 sujets) : 29 décès.
- La différence est-elle significative ?

Test de comparaison de deux proportions

- Modèle : $X \sim \mathcal{B}(n_X, p_X)$, $Y \sim \mathcal{B}(n_Y, p_Y)$.
- Hypothèses :
  - $H_0 : p_X = p_Y$
  - $H_1 : p_X \neq p_Y$, $p_X < p_Y$ ou $p_X > p_Y$.
- Statistique de test :

$$\frac{\widehat{p}_X - \widehat{p}_Y}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_X} + \frac{1}{n_Y})}} \overset{H_0}{\sim} \mathcal{N}(0,1) \quad \text{(approx.)}$$

avec $\widehat{p}_X = X/n_X$, $\widehat{p}_Y = Y/n_Y$, $\widehat{p} = (X+Y)/(n_X+n_Y)$.

Test de comparaison de deux proportions en R

```
> prop.test(c(20,29),c(102,98))
2-sample test for equality of proportions with
continuity correction
data: c(20, 29) out of c(102, 98)
X-squared = 2.1806, df = 1, p-value = 0.1398
alternative hypothesis: two.sided
95 percent confidence interval:
-0.22860281 0.02892294
sample estimates:
prop 1 prop 2
0.1960784 0.2959184
```

Plan

Echantillons appariés
- Exemple introductif
- Test de Student
- Test de Wilcoxon signé

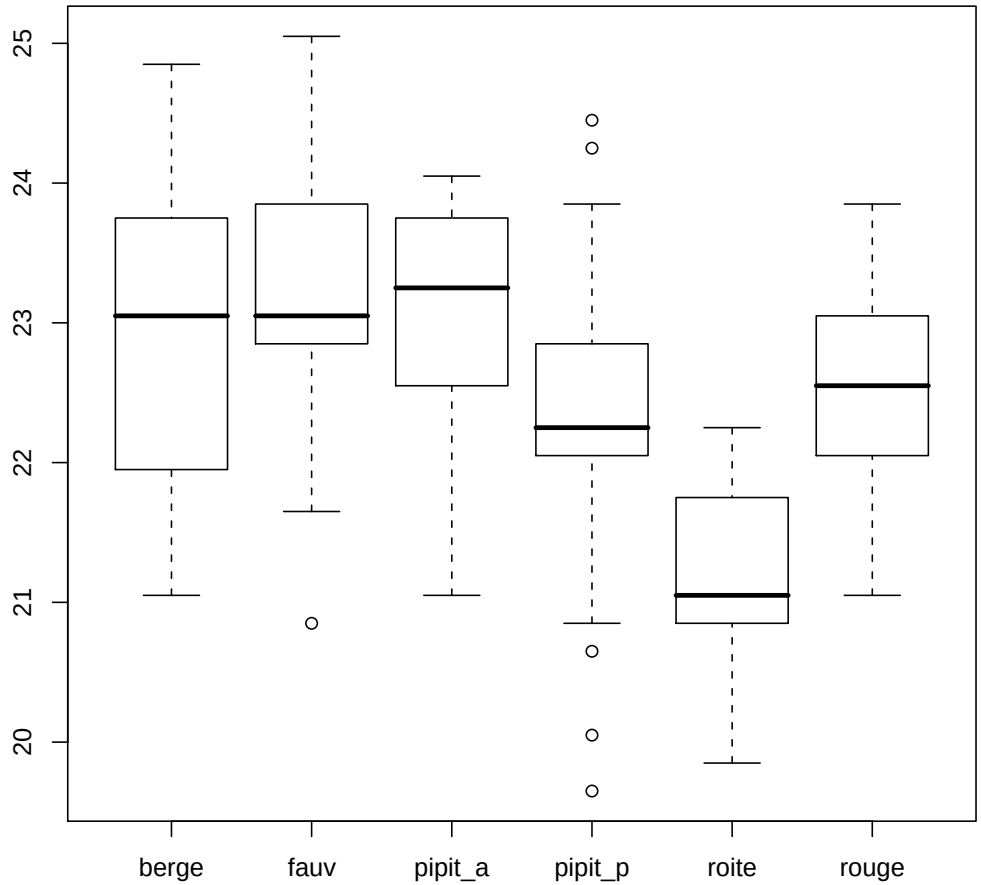
Deux échantillons indépendants
- Comparaison de deux variances
- Comparaison de deux moyennes
- Test de Wilcoxon
- Comparaison de deux proportions

K échantillons indépendants
- Comparaison de K moyennes
- Comparaison de K variances
- Test de Kruskall-Wallis

Exemple : nids de coucous

- Une étude menée par E.B. Chance en 1940 a montré que les coucous retournent chaque année dans le même territoire et pondent leurs oeufs dans les nids d'espèces hôtes particulières.
- Des sous-espèces géographiques se développeraient ainsi, pondant des oeufs susceptibles d'être adoptés par les espèces hôtes.
- Données : tailles de 115 oeufs de coucous relevées dans les nids de six espèces différentes (d'après O.M. Latter, 1902) :
  - pipit des prés ;
  - pipit des arbres ;
  - fauvette des haies ;
  - rouge-gorge ;
  - bergeronnette de Yarrell ;
  - roitelet.

Nids de coucous : données



Analyse de la variance à un facteur

- Modèle : K v.a. gaussiennes de même variance $X_k \sim \mathcal{N}(\mu_k, \sigma^2)$.
- Pour chaque X_k on dispose d'un échantillon i.i.d. $X_k^1, \ldots, X_k^{n_k}$ de taille n_k, et on note $N = n_1 + \cdots + n_K$.
- Hypothèses :
  - $H_0 : \mu_1 = \ldots = \mu_K$
  - $H_1 : \exists k, \ell$ t.q. $\mu_k \neq \mu_\ell$.
- Statistique de test :

$$F = \frac{MSB}{MSW} \overset{H_0}{\sim} \mathcal{F}_{K-1,N-K}$$

avec

$$MSB = \frac{1}{K-1} \sum_{k=1}^{K} n_k (\overline{X}_k - \overline{X})^2, \quad MSW = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_k^i - \overline{x}_k)^2$$

- Degré de signification : $p = \mathbb{P}_{H_0}(F > f)$.

Analyse de la variance en R

```
> anova(lm(length ~ bird))
Analysis of Variance Table

Response: length
           Df  Sum Sq Mean Sq F value    Pr(>F)
bird        5  42.940   8.588  10.388 3.152e-08 ***
Residuals 114  94.248   0.827
```

Comparaisons par paires

```
> pairwise.t.test(length,bird)

Pairwise comparisons using t tests with pooled SD

data: length and bird

          berge    fauv     pipit_a  pipit_p  roite
 fauv     1.00000  -        -        -        -
 pipit_a  1.00000  1.00000  -        -        -
 pipit_p  0.22181  0.03787  0.03817  -        -
 roite    6.2e-06  5.6e-07  5.6e-07  0.00038  -
 rouge    1.00000  0.72321  0.72321  1.00000  0.00027
P value adjustment method: holm
```

Plan

Echantillons appariés
- Exemple introductif
- Test de Student
- Test de Wilcoxon signé

Deux échantillons indépendants
- Comparaison de deux variances
- Comparaison de deux moyennes
- Test de Wilcoxon
- Comparaison de deux proportions

K échantillons indépendants
- Comparaison de K moyennes
- Comparaison de K variances
- Test de Kruskall-Wallis

Test de Bartlett

- Modèle : K v.a. gaussiennes $X_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$.
- Pour chaque X_k on dispose d'un échantillon i.i.d. $X_k^1, \ldots, X_k^{n_k}$ de taille n_k, et on note $N = n_1 + \cdots + n_K$.
- Hypothèses :
  - $H_0 : \sigma_1^2 = \ldots = \sigma_K^2$
  - $H_1 : \exists k, \ell$ t.q. $\sigma_k^2 \neq \sigma_\ell^2$.
- Statistique de test :

$$B = (N-K)\ln(MSW) - \sum_{k=1}^{K}(n_k - 1)\ln(S_k^{*2}) \overset{H_0}{\sim} \chi^2_{K-1} \quad \text{(approx.)}$$

- Degré de signification : $p = \mathbb{P}_{H_0}(B > b)$.

Test de Bartlett en R

```
> bartlett.test(length ~ bird)

Bartlett test of homogeneity of variances

data: length by bird
Bartlett's K-squared = 4.4794, df = 5, p-value = 0.4826
```

Plan

Echantillons appariés
- Exemple introductif
- Test de Student
- Test de Wilcoxon signé

Deux échantillons indépendants
- Comparaison de deux variances
- Comparaison de deux moyennes
- Test de Wilcoxon
- Comparaison de deux proportions

K échantillons indépendants
- Comparaison de K moyennes
- Comparaison de K variances
- Test de Kruskall-Wallis

Test de Kruskall-Wallis

- Modèle : K v.a. X_k de fonction de répartition F_k.
- Pour chaque X_k on dispose d'un échantillon i.i.d. $X_k^1, \ldots, X_k^{n_k}$ de taille n_k, et on note $N = n_1 + \cdots + n_K$.
- Hypothèses :
  - $H_0 : F_1 = \ldots = F_K$
  - $H_1 : \exists k, \ell$ t.q. $F_k \neq F_\ell$.
- Statistique de test :

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{K} \frac{1}{n_k} \left(\sum_{i=1}^{n_k} R_k^i \right)^2 - 3(N+1) \overset{H_0}{\approx} \chi^2_{K-1} \text{ (approx.,)}$$

avec R_k^i = rang de l'observation X_k^i dans la série des N observations triées par ordre croissant.

- Degré de signification : $p = \mathbb{P}_{H_0}(H > h)$.

Test de Kruskall-Wallis en R

```
> kruskal.test(length ~ bird)

Kruskal-Wallis rank sum test

data: length by bird
Kruskal-Wallis chi-squared = 35.0403, df = 5, p-value = 1.477e-06
```

Quatrième partie IV

Tests d'indépendance

Plan

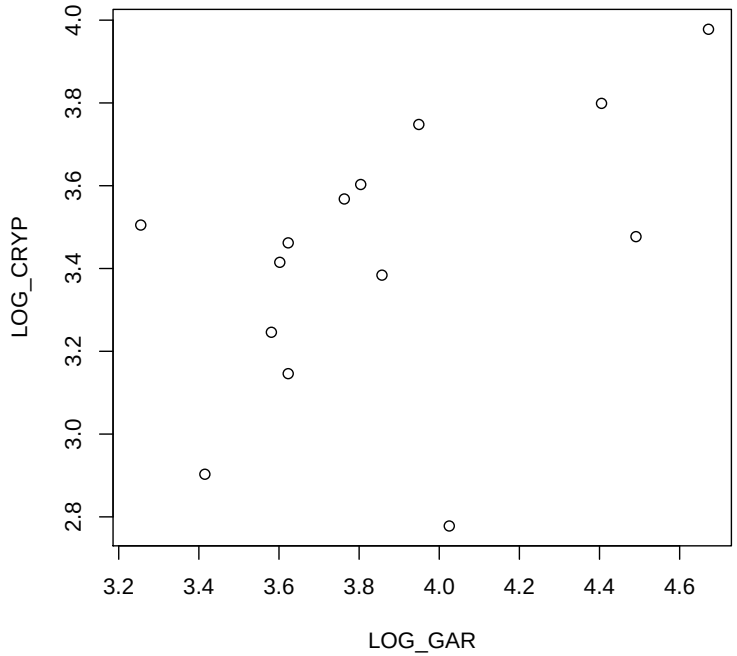
Variables quantitatives

Variables quantitatives ou ordinales

Variables qualitatives nominales

Relation entre cryptosporidium et giardia

- 14 mesures de concentration en cryptosporidium et giardia (microorganismes pathogènes) dans des échantillons d'eau de Seine.



- Existence d'une liaison entre ces deux variables ?

Coefficient de corrélation de Pearson

- Soient X et Y deux v.a. Leur coefficient de corrélation est

$$\rho = \frac{\mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]}{\sigma_X \sigma_Y}.$$

- $\rho \in [-1, 1]$, et $\rho = \pm 1$ ssi les v.a. X et Y dont liées par une relation linéaire.
- Estimateur de ρ : coefficient de corrélation de Pearson

$$R = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{S_X S_Y}.$$

- $R \in [-1, 1]$, et $R = \pm 1$ ssi les points (X_i, Y_i) sont alignés.

Test de significativité du coefficient de Pearson

- Modèle : $(X_1, Y_1), \ldots, (X_n, Y_n)$ échantillon i.i.d. du couple (X, Y) suivant une loi normale bidimensionnelle.
- Hypothèses :
  - $H_0 : \rho = 0$ ($\Leftrightarrow$ X et Y sont indépendantes)
  - $H_1 : \rho \neq 0$, $\rho < 0$ ou $\rho > 0$.
- Statistique de test : R. On a

$$\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \overset{H_0}{\sim} \mathcal{T}_{n-2}.$$

- Degré de signification :

$$p = \begin{cases} \mathbb{P}_{H_0}(R \geq r) & \text{si } H_1 : \rho > 0 \\ \mathbb{P}_{H_0}(R \leq r) & \text{si } H_1 : \rho < 0 \\ \mathbb{P}_{H_0}(|R| \geq |r|) & \text{si } H_1 : \rho \neq 0. \end{cases}$$

Test de significativité du coefficient de Pearson en R

```
> cor.test(LOG_GAR,LOG_CRYP,method='p')

Pearson's product-moment correlation

data: LOG_GAR and LOG_CRYP
t = 1.9725, df = 12, p-value = 0.07205
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.04850212 0.81216695
sample estimates:
cor
0.494811
```

Si (X, Y) est non gaussien, le test est utilisable si $n \geq 30$ (mais la non corrélation n'implique plus alors l'indépendance).

Plan

Variables quantitatives

Variables quantitatives ou ordinales
- Coefficient de corrélation de Spearman
- Coefficient de corrélation de Kendall

Variables qualitatives nominales

Coefficient de corrélation de Spearman

- Echantillon initial : $(X_1, Y_1), \ldots, (X_n, Y_n)$
- Nouvel échantillon : $(R_1, S_1), \ldots, (R_n, S_n)$
  - R_i : rang de X_i dans l'echantillon associé à X
  - S_i : rang de Y_i dans l'échantillon associé à Y
- Coeff. de corrélation de Spearman R_s : coefficient de corrélation linéaire calculé sur les rangs. On peut montrer que l'on obtient

$$R_s = 1 - \frac{6 \sum_i D_i^2}{n^2(n-1)}$$

avec $D_i = R_i - S_i$.

Test de significativité du coefficient de Spearman

- Modèle : $(X_1, Y_1), \ldots, (X_n, Y_n)$ échantillon i.i.d. du couple (X, Y).
- Hypothèses :
  - H_0 : X et Y sont indépendantes
  - H_1 : dépendance quelconque (test bilatérale), positive ou négative (tests unilatéraux).
- Statistique de test : R_s. On a

$$\frac{R_s\sqrt{n-2}}{\sqrt{1-R_s^2}} \overset{H_0}{\sim} \mathcal{T}_{n-2} \quad \text{approx.}$$

- Degré de signification :

$$p = \begin{cases} \mathbb{P}_{H_0}(R_s \geq r_s) & \text{si } H_1 : \text{dépendance positive} \\ \mathbb{P}_{H_0}(R_s \leq r) & \text{si } H_1 : \text{dépendance négative} \\ \mathbb{P}_{H_0}(|R_s| \geq |r|) & \text{si } H_1 : \text{dépendance quelconque.} \end{cases}$$

Test de significativité du coefficient de Spearman en R

```
> cor.test(LOG_GAR,LOG_CRYP,method='s')

Spearman's rank correlation rho

data: LOG_GAR and LOG_CRYP
S = 241.7655, p-value = 0.09097
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.4686471

Warning message:
In cor.test.default(LOG_GAR, LOG_CRYP, method = "s") :
Impossible de calculer les p-values exactes avec des ex-aequos
```

Plan

Variables quantitatives

Variables quantitatives ou ordinales
- Coefficient de corrélation de Spearman
- Coefficient de corrélation de Kendall

Variables qualitatives nominales

Coefficient de corrélation de Kendall

- Coefficient de corrélation de Kendall théorique :

$$\tau = 2\mathbb{P}\left((X_1 - X_2)(Y_1 - Y_2) > 0\right) - 1.$$

- τ compris entre -1 et 1, nul si X et Y sont indépendantes.
- Estimation : On considère tous les couples (X_i, Y_i).
- On note $+1$ si deux individus i et j sont dans le même ordre pour les deux variables : $X_i < X_j$ et $Y_i < Y_j$ et -1 si deux individus i et j sont dans des ordres différents : $X_i < X_j$ et $Y_i > Y_j$.
- Soit S la somme des $n(n-1)/2$ couples distincts.
- On pose :

$$\widehat{\tau} = \frac{2S}{n(n-1)}.$$

Test de significativité du coefficient de Kendall

- Modèle : $(X_1, Y_1), \ldots, (X_n, Y_n)$ échantillon i.i.d. du couple (X, Y).
- Hypothèses :
  - $H_0 : \tau = 0$
  - $H_1 : \tau \neq 0, \tau > 0, \tau < 0.$
- Statistique de test :

$$\widehat{\tau} \overset{H_0}{\sim} \mathcal{N}\left(0, \frac{2(2n+5)}{9n(n-1)}\right) \quad \text{approx.}$$

- Approximation gaussienne valable pour $n \geq 10$.

Test de significativité du coefficient de Kendall en R

```
> cor.test(LOG_GAR,LOG_CRYP,method='k')

Kendall's rank correlation tau

data: LOG_GAR and LOG_CRYP
z = 1.9738, p-value = 0.04841
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.3977961

Warning message:
In cor.test.default(LOG_GAR, LOG_CRYP, method = "k") :
Impossible de calculer la p-value exacte avec des ex-aequos
```

Régime méditerranéen et maladies cardiovasculaires

- 605 survivants de crise cardiaque partionnés aléatoirement en 2 groupes suivant :
  - le régime standard préconisé par l'American Heart Association (AHA) ;
  - un régime méditerranéen (pain, fruit, poisson, huile d'olive, vin, etc.)
- Au bout de quatre ans, évaluation de l'état de santé des patients.
- Source : De Lorgeril, M., Salen, P., Martin, J., Monjaud, I., Boucher, P., Mamelle, N. (1998). Mediterranean Dietary pattern in a Randomized Trial. Archives of Internal Medicine, 158, 1181-1187.

Données

```
> med.regime
```

	Cancers	Décès	Mal. bén.	Bonne santé
AHA	15	24	25	239
medit.	7	14	8	273

Test du χ^2

- Modèle :
  - X et Y variables aléatoires qualitatives prenant r et s valeurs
  - Échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$
- Hypothèses :
  - H_0 : X et Y sont indépendantes
  - H_1 : X et Y ne sont pas indépendantes

Statistique de test

- Tableau de contingence :

X^Y	1		j		s	
1			.			
i		...	N_{ij}			$N_{i.}$
r						
			$N_{.j}$			

- Statistique de test

$$D^2 = \sum_i \sum_j \frac{(N_{ij} - \frac{N_{i.}N_{.j}}{n})^2}{\frac{N_{i.}N_{.j}}{n}} \overset{H_0}{\sim} \chi^2_{(r-1)(s-1)} \quad \text{approx}$$

- Degré de signification : $p = \mathbb{P}_{H_0}(D^2 > d^2)$.

Test du χ^2 en R

```
> chisq.test(med.regime)

Pearson's Chi-squared test

data: med.regime
X-squared = 16.5545, df = 3, p-value = 0.0008726
```

Interprétation du résultat

```
> chisq.test(med.regime)$observed

         Cancers  Décès  Mal. bén.  Bonne santé
AHA           15     24         25          239
medit.         7     14          8          273
> chisq.test(med.regime)$expected

         Cancers    Décès  Mal. bén.  Bonne santé
AHA     11.01818 19.03140   16.52727     256.4231
medit.  10.98182 18.96860   16.47273     255.5769
```