Introduction à la statistique non paramétrique

Catherine MATIAS

CNRS, Laboratoire Statistique & Génome, Évry

http://stat.genopole.cnrs.fr/∼cmatias

Atelier SFDS
27/28 septembre 2012

Partie 2 : Tests non paramétriques

Plan partie 2

Introduction I

Contexte

- Dans la suite, on observe soit un échantillon $X_1, \ldots, X_n$ de v.a. réelles i.i.d de même loi que X ou bien deux échantillons $X_1, \ldots, X_n$ de même loi que X et $Y_1, \ldots, Y_m$ de même loi que Y.
- Les tests sont non paramétriques lorsque la distribution des variables aléatoires n'est pas spécifiée sous au moins une des deux hypothèses (nulle ou alternative).

Introduction II

Exemples

- Tests d'adéquation à une loi : H_0 : "X suit la loi F_0" contre H_1 : "X ne suit pas la loi F_0".
- Tests d'adéquation à une famille de lois : H_0 : "X est gaussienne" (paramètres non spécifiés) contre H_1 : "X n'est pas gaussienne".
- Tests de comparaison (ou homogénéité) : H_0 :"X et Y ont la même loi" contre H_1 : "Xet Y n'ont pas la même loi".
- Tests d'indépendance : H_0 : $\{X_i\} \amalg \{Y_i\}$ contre H_1 : "$\{X_i\}$ ne sont pas indépendants des $\{Y_i\}$".

Principe

Principe général des tests

Trouver une statistique (de test) $T(X_1, \ldots, X_n)$ (ou bien $T(X_1, \ldots X_n, Y_1, \ldots, Y_m)$) dont la distribution sous H_0 ne dépend pas de la distribution des v.a. observées. On parle de statistique libre en loi.

Deux types de tests

- bilatères : lorsque sous l'alternative H_1, la statistique T n'est ni systématiquement "plus grande" ni "plus petite" que sous H_0.
- unilatères : lorsque sous l'alternative H_1, la statistique T est soit systématiquement "plus grande", soit "plus petite" que sous H_0.

Donner un sens à "plus grande" ou "plus petite" pour des variables aléatoires : notion d'ordre stochastique.

Ordre stochastique I

Définition

Si X v.a. réelle de fdr F et Y v.a. réelle de fdr G et si $\forall x \in \mathbb{R}$, on a $G(x) \leq F(x)$ avec inégalité stricte pour au moins un $x \in \mathbb{R}$, alors on dit que Y est stochastiquement plus grande que X et on note $Y \succ X$.

En particulier, si T est une v.a.r. de fdr F_0 sous l'hypothèse H_0 et de fdr F_1 sous l'hypothèse H_1 et si $\forall x \in \mathbb{R}$, $F_0(x) \leq F_1(x)$ avec inégalité stricte en au moins un point, alors T est stochastiquement plus grande sous H_0 que sous H_1.

Ordre stochastique II

Exemple

$T \sim \mathcal{N}(\theta, 1)$, $H_0 : \theta = 0$ et $H_1 : \theta = 1$.
T est stochastiquement plus petite sous H_0 que sous H_1.
En effet,
$F_1(x) = \mathbb{P}_{H_1}(T \leq x) = \mathbb{P}_{H_1}(T - 1 \leq x - 1) = \mathbb{P}(Z \leq x - 1)$ où $Z \sim \mathcal{N}(0, 1)$. De même $F_0(x) = \mathbb{P}_{H_0}(T \leq x) = \mathbb{P}(Z \leq x)$. Donc on obtient $F_1(x) = F_0(x - 1) < F_0(x)$, car F_0 strictement croissante.

Propriété

Si $X_1 \prec Y_1, X_2 \prec Y_2$ et $X_1 \amalg X_2$, $Y_1 \amalg Y_2$ alors
$X_1 + X_2 \prec Y_1 + Y_2$.

Choix de la région de rejet I

- C'est l'hypothèse alternative H_1 qui détermine si le test est bilatère ou unilatère.
- C'est aussi l'alternative H_1 qui détermine la région de rejet de l'hypothèse nulle H_0 : $\mathcal{R}_{H_0}$ choisie t.q. la densité de la stat. de test T sur $\mathcal{R}_{H_0}$ est plus grande sous H_1 que sous H_0.

Deux approches sont possibles :

1) On fixe un niveau α (erreur maximale de première espèce) et on cherche le seuil (donc la zone de rejet) tel que $\mathbb{P}_{H_0}(\text{Rejeter } H_0) \leq \alpha$. Ce test pourra être appliqué à tout jeu de données observées ultérieurement, et l'hypothèse testée au niveau α.

Choix de la région de rejet II

2) On observe une réalisation $x_1, \ldots, x_n$ et on calcule le degré de significativité (ou p-value) correspondant à cette réalisation, *i.e.* le plus petit niveau γ tel qu'on rejette le test à ce niveau avec les valeurs observées. Ce test est spécifique à l'observation mais permet de répondre au test pour toutes les valeurs de α, sur ce jeu de données.

Exemple (degré de significativité)

- $\mathcal{R}_{H_0} = \{S_n \geq s\}$
- On observe la valeur de la statistique s^{obs}, alors $\gamma = \mathbb{P}_{H_0}(S_n \geq s^{\text{obs}})$ est le degré de significativité du test pour la valeur observée.
- Tout test de niveau $\alpha < \gamma$ accepte H_0 et tout test de niveau $\alpha \geq \gamma$ rejette H_0.

Choix de la région de rejet III

Cas des tests bilatères

- $\mathcal{R}_{H_0} = \{T_n \geq b\} \cup \{T_n \leq a\}$, avec $a \leq b$, seuils à déterminer.
- En pratique, si on se fixe un niveau α positif, alors on choisira a, b tels que $\mathbb{P}_{H_0}(T_n \geq b) = \mathbb{P}_{H_0}(T_n \leq a) \leq \alpha/2$.
- Si T_n a une distribution symétrique par rapport à m_0 sous l'hypothèse nulle H_0, on peut écrire de façon équivalente $\mathcal{R}_{H_0} = \{|T_n - m_0| \geq s\}$.
- Le degré de significativité n'a pas de sens pour un test bilatère. Une fois que les données sont observées, le test rejette pour une des deux alternatives, jamais les deux en même temps !

Puissance de test

- La fonction puissance est difficile à évaluer pour un test non paramétrique car l'ensemble des alternatives est très grand et contient des distributions très différentes.
- En particulier, il est difficile de comparer des tests de même niveau. On pourra plutôt en considérer plusieurs. Certains tests sont mieux adaptés à certaines alternatives que d'autres.
- Par construction, ces tests ne dépendent pas de la distribution des v.a. et ont les mêmes qualités quelle que soit cette distribution. En ce sens, ils sont dits robustes.

Efficacité asymptotique I

Définitions

- Pour tout test basé sur une stat T_n, on peut définir $N_T(\alpha, \beta, F)$: taille minimale de l'échantillon nécessaire pour que le test fondé sur T_n, de niveau α ait la puissance au moins β en l'alternative F.
- L'efficacité relative entre deux tests resp. fondés sur les stats T_n et U_n est la fonction

$$e_{T,U}(\alpha, \beta, F) = \frac{N_T(\alpha, \beta, F)}{N_U(\alpha, \beta, F)}.$$

- Pbm : en général, $e_{T,U}(\alpha, \beta, F)$ n'est pas calculable pour tout α, β, F. Par contre, on peut avoir les limites lorsque $\alpha \to 0, \beta \to 1$ ou $F \to F_0$.

Efficacité asymptotique II

Efficacité asymptotique au sens de Bahadur

Si pour $\beta \in (0,1)$ et $F \in H_1$, la limite

$$\lim_{\alpha \to 0} e_{T,U}(\alpha, \beta, F) := \mathcal{E}_{T,U}(\beta, F)$$

existe, alors c'est l'efficacité asymptotique relative (ARE), au sens de Bahadur, de $\{T_n\}$ relativement à $\{U_n\}$, aux points (β, F).

ARE des tests non paramétriques

- Il existe des résultats sur l'ARE des tests d'adéquation, des tests de symétrie et des tests de comparaison.
- Voir [Nikitin 95] pour plus de détails.

Correction du continu I

Contexte

- Stat. de test T_n discrète dont la loi approchée est continue.
- Ex : cadre asymptotique avec

$$\frac{T_n - \mathbb{E}_{H_0}(T_n)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(T_n)}} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0;1) \text{ sous } H_0.$$

- Si $\mathcal{R}_{H_0} = \{T_n \geq t\}$ et $\forall \alpha > 0$ on cherche le seuil t tel que $\mathbb{P}_{H_0}(T_n \geq t) \leq \alpha$.
- Or, $\forall u \in [0, 1[$, comme T_n est discrète,

$$\mathbb{P}_{H_0}(T_n \geq t) = \mathbb{P}_{H_0}(T_n \geq t - u).$$

- De même, si $\mathcal{R}_{H_0} = \{T_n \leq t\}$, alors $\forall u \in [0; 1[$ on a $\mathbb{P}_{H_0}(T_n \leq t) = \mathbb{P}_{H_0}(T_n \leq t + u)$.

Correction du continu II

Mise en œuvre

- La correction du continu consiste à remplacer la valeur par défaut $u = 0$ par $u = 1/2$.
- Ex : si $\mathcal{R}_{H_0} = \{T_n \geq t\}$, on cherche le seuil t tel que

$$\mathbb{P}_{H_0}(T_n \geq t - 0.5) \leq \alpha$$

$$\iff \mathbb{P}_{H_0}\left(\frac{T_n - \mathbb{E}_{H_0}(T_n)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(T_n)}} \geq \frac{t - 0.5 - \mathbb{E}_{H_0}(T_n)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(T_n)}}\right) \leq \alpha.$$

Test d'une hypothèse induite

Remarques

- Il arrive que pour tester H_0, on teste en fait H_0' telle que $H_0 \Rightarrow H_0'$.
- Exemple : H_0 : "Les variables sont gaussiennes" et H_0' : "le moment recentré d'ordre 3 est nul".
- Si on rejette H_0' alors on rejette nécessairement H_0. Par contre, si on accepte H_0', on n'accepte pas nécessairement H_0 !
- N.B. Lorsque H_0' est une hypothèse paramétrique, on sort du cadre des tests non paramétriques.
- Exemple : voir plus loin le test du signe.

Tests par permutation

Pincipe

- Il s'agit d'une technique générale pour échantillonner la loi de la statistique de test sous H_0.
- Requiert que les individus soient i.i.d (ou plus généralement échangeables).
- Permet d'obtenir un test exact (par opposition à asymptotique) ;
- Peut être difficile à mettre en œuvre d'un point de vue puissance de calcul (si trop de catégories par exemple).

Exemple

- Pour tester que 2 échantillons ont la même loi, on peut permuter les individus entre les échantillons.

Plan partie 2

Rappel cas discret : Test d'adéquation du χ^2 de Pearson I

Description

Pour un échantillon de v.a. discrètes avec r modalités (qualitatives ou quantitatives) $X_1, \ldots, X_n$ et une distribution p_0 fixée, on teste $H_0 : "p = p_0"$ contre $H_1 : "p \neq p_0"$.

Statistique de Pearson

$$\chi^2 = n \sum_{k=1}^{r} \frac{(\hat{p}_k - p_0(k))^2}{p_0(k)}, \tag{1}$$

où $\hat{p}_k = \sum_{i=1}^{n} 1\{X_i = k\}/n$. On rejette H_0 pour les grandes valeurs de χ^2.

Rappel cas discret : Test d'adéquation du χ^2 de Pearson II

Propriétés

- Test asymptotique, fondé sur la loi limite de la statistique de test qui suit un $\chi^2(r-1)$.
- Sous certaines hyps, utilisable lorsque p_0 est définie à un paramètre près $\theta_0 \in \mathbb{R}^s$ qui est estimé (loi limite $\chi^2(r-s-1)$).
- C'est un test consistant : pour toute alternative $p \neq p_0$, la puissance $\beta_n(p) \to 1$.
- Remarque : C'est en fait un test paramétrique ! puisque la loi discrète des X_i dépend d'un nombre fini de paramètres.

Cas continu : test d'adéquation du χ^2 de Pearson avec catégories I

Description

- Pour un échantillon de v.a. continues $X_1, \ldots, X_n$ et une fdr F_0 fixée, on veut tester H_0 : "$F = F_0$" contre H_1 : "$F \neq F_0$".
- On découpe $\mathbb{R} = (-\infty, a_1) \cup (a_1, a_2) \cup \ldots (a_{r-2}, a_{r-1}) \cup (a_{r-1}, +\infty)$ en intervalles fixés pour obtenir r modalités :
  $\hat{p}_k = \sum_{i=1}^{n} 1\{X_i \in (a_{k-1}, a_k)\}/n$ et
  $p_0(k) = F_0(a_k) - F_0(a_{k-1})$.
- On teste alors l'hypothèse induite H_0' : "Les versions discrétisées de F et F_0 sont identiques".
- Même statistique qu'en (1) et même type de zone de rejet.

Cas continu : test d'adéquation du χ^2 de Pearson avec catégories II

Propriétés

- Test asymptotique, fondé sur la loi limite de la statistique de test qui suit un $\chi^2(r-1)$.
- Sous certaines hyps, utilisable lorsque F_0 est définie à un paramètre près $\theta_0 \in \mathbb{R}^s$ qui est estimé (loi limite $\chi^2(r-s-1)$).
- Ce test n'est pas consistant pour toute alternative $F \neq F_0$ (en fait H_1' est plus "petite" que H_1).
- Peut-être généralisé au cas où les a_k sont aléatoires et tels que le nombre de points dans chaque intervalle est fixé.
- La version du test à r fixé est paramétrique. Par contre, il existe des généralisations avec $r = r(n) \to +\infty$, qui sont alors non paramétriques [Morris 75].

Cas continu : test de Kolmogorov Smirnov (KS)

Description

Pour un échantillon de v.a. continues $X_1, \ldots, X_n$ et une fdr F_0 fixée, on teste $H_0 : "F = F_0"$ contre $H_1 : "F \neq F_0"$.

Statistique de KS

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$
$$= \max_{1 \leq i \leq n} \{|F_0(X_{(i)}) - i/n|, |F_0(X_{(i)}) - (i-1)/n|\}.$$

On rejette H_0 pour les grandes valeurs de D_n.

Propriétés

- Statistique libre en loi sous H_0. Cette loi est tabulée.
- Approximation pour n grand

$$\mathbb{P}_{H_0}(\sqrt{n} D_n > z) \underset{n \to \infty}{\to} 2 \sum_{j \geq 1} (-1)^{j-1} \exp(-2j^2 z^2).$$

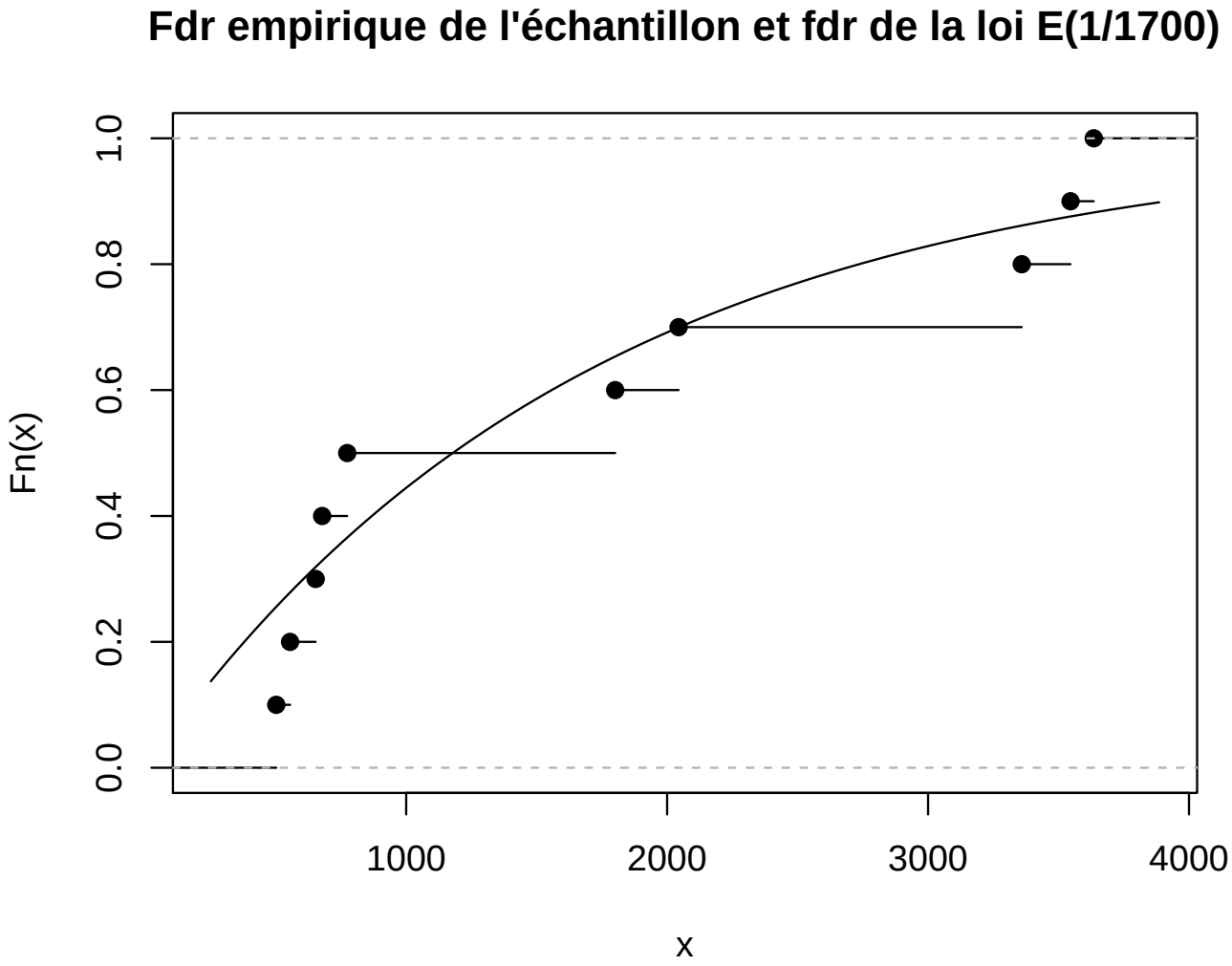
Test KS : exemple I

Contexte

Un fabricant garantit que la fiabilité des appareils qu'il vend est telle que leur durée de vie suit une loi exponentielle, de moyenne 1700 heures. Afin de tester cette affirmation, on mesure la durée de vie en heures de 10 de ces appareils pris au hasard. On obtient les valeurs suivantes $\{555, 653, 1801, 678, 3635, 502, 2044, 3359, 3546, 774\}$.

```
> x<-c(555,653,1801,678,3635,502,2044,3359,3546,774)
> plot.ecdf(x,main="Fdr empirique de l'échantillon
+ et fdr de la loi E(1/1700)")
> curve(pexp(q=x,1/1700),add=T)
```

Test KS : exemple II



Test KS : exemple III

Test "à la main"

```
> x <- sort(x) # on ordonne l'échantillon
> u <- (1:10)/10 # vecteur des valeurs i/n
> v <- (0:9)/10 # vecteur des valeurs (i-1)/n
> Fx <- pexp(x,1/1700) # vecteur des valeurs F_0(X_(i))
> diff <-pmax(abs(Fx-u), abs(Fx-v)) # vecteur des valeurs
> # max entre les deux arguments
> Dn <- max(diff) # KS stat
> Dn
```

```
[1] 0.2556874
```

La différence maximale est obtenue pour l'indice 1. Elle correspond donc soit à $|F_0(X_{(1)}) - 1/n|$ soit à $|F_0(X_{(1)}) - 0|$. En regardant le graphique, on constate qu'il s'agit de $|F_0(X_{(1)}) - 0|$. En consultant une table, on obtient $\mathbb{P}_{H_0}(D_n \geq 0,409) = 0,05$ donc la zone de rejet de H_0 au niveau $\alpha = 5\%$ est $\{D_n \geq 0,409\}$. La valeur observée D_n^{obs} ci-dessus (0.255) n'est pas dans la zone de rejet.

Test KS : exemple IV

Test via la fonction *ks.test* sous R

```
> ks.test(x/1700,"pexp",alternative="two.sided")

        One-sample Kolmogorov-Smirnov test

data:  x/1700
D = 0.2557, p-value = 0.4559
alternative hypothesis: two-sided

> # ou bien ks.test(x,"pexp",1/1700,alternative="two.sided"
```

Autres tests : Cramér von Mises et Anderson-Darling

Dans le même contexte que KS, on peut utiliser

Statistiques de test

- $CVM_n = n \int (\hat{F}_n(x) - F_0(x))^2 dF_0(x)$
- $A_n = n \int \frac{(\hat{F}_n(x) - F_0(x))^2}{F_0(x)(1-F_0(x))} dF_0(x)$

On rejette H_0 pour les grandes valeurs de CVM_n ou A_n.

Propriétés

- Statistiques libres en loi sous H_0.
- Tests souvent plus puissants que KS. En particulier, A_n sensible aux écarts à F_0 dans la queue de distribution.

Implémentation sous R

Fonction *ad.test* dans le paquet *ADGofTest*, *cvm.test* dans le paquet *dgof*.

Conclusions sur tests d'adéquation à une loi fixée

Puissance

- KS, CVM et AD sont souvent plus puissants que χ^2,
- CVM et AD souvent plus puissants que KS.

Adéquation à une famille de lois

- χ^2 se généralise au cas de $H_0' : F = F_0(\theta_1, \ldots, \theta_s)$ où $\theta_1, \ldots, \theta_s$ inconnus.
- Les tests KS, CVM et AD ne s'appliquent pas directement à ce cadre. En effet, leurs statistiques ne sont plus libres en loi (même asymptotiquement) sous H_0'. Il faut adapter ces tests pour chaque famille de loi considérée.

Plan partie 2

Tests d'adéquation à une famille de lois

- Le test du χ^2 se généralise au cas $H_0' : F = F_0(\theta_1, \ldots, \theta_s)$ où $\theta_1, \ldots, \theta_s$ inconnus.
- Les tests KS, CVM et AD doivent être modifiés dans ce cadre. Voir [De Wet and Randles 87] pour plus de détails.

Cas particulier : famille gaussienne

- Les tests de normalité permettent de tester H_0 : "F suit une loi gaussienne (de paramètres non spécifiés)" contre H_1 : "F n'est pas gaussienne".
- En pratique, on teste H_0' : "$F = \mathcal{N}(\hat{\theta}_n, \hat{\sigma}_n^2)$" (paramètres estimés).
- Il existe des généralisations de KS (test de Lilliefors) et CVM à ce cadre, implémentées sous R (Paquet *nortest*, fonctions *lillie.test* et *cvm.test*. Contient également *Pearson.test*).

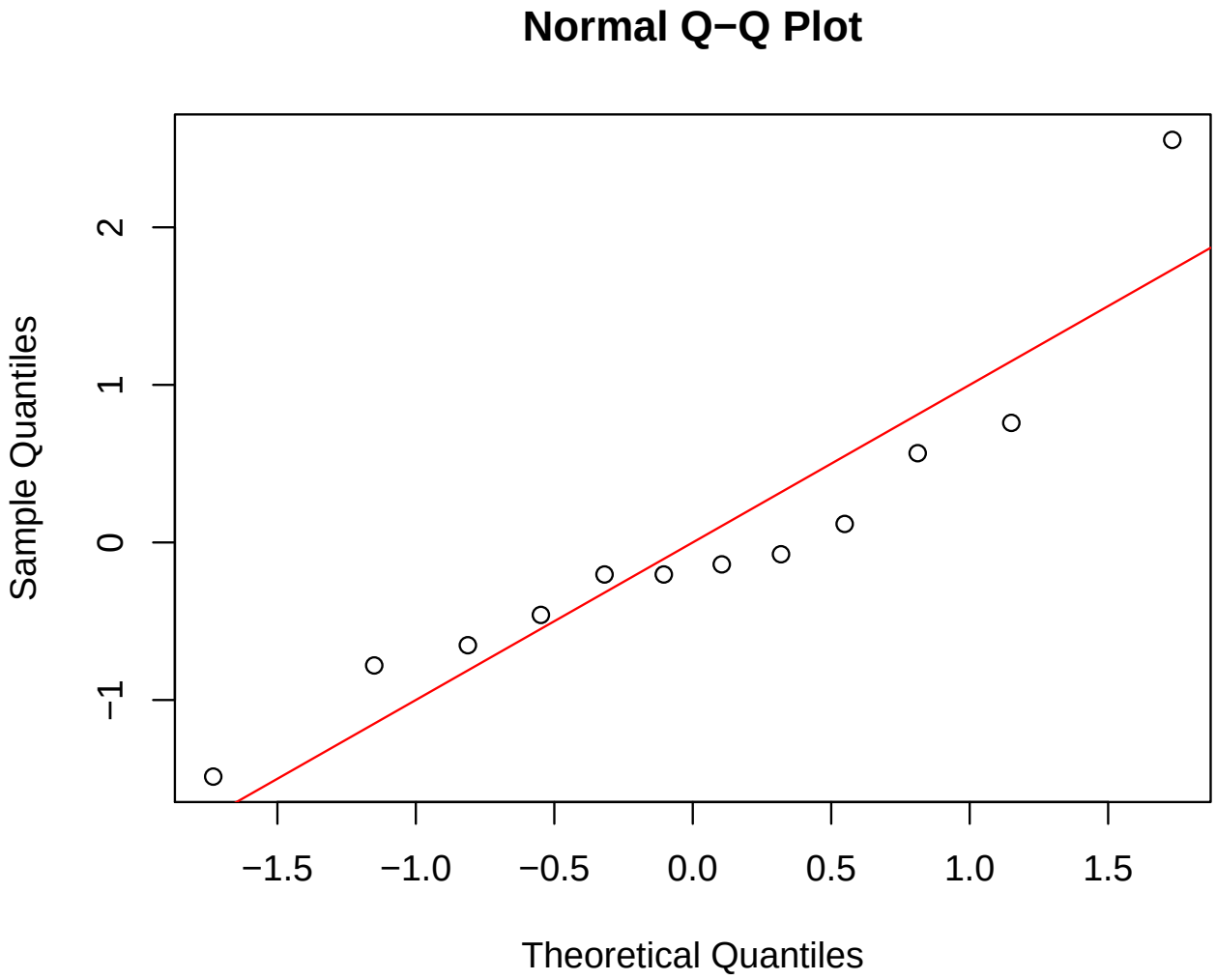
Tests de normalité : exemple I

Contexte

On cherche à déterminer si la taille d'une population de chats d'une race donnée peut-être considérée comme une variable aléatoire de distribution normale. On mesure un échantillon de 12 chats pris au hasard dans cette espèce. Les observations, en centimètres, sont les suivantes
$\{20.7, 18.7, 20.8, 20.7, 22.2, 20.9, 19.8, 20.3, 25, 20, 21.9, 21.2\}$.

```
> x <- c(20.7,18.7,20.8,20.7,22.2,20.9,19.8,20.3,25,20,21.9
> qqnorm(scale(x),col=1)
> abline(0,1,col=2)
```

Tests de normalité : exemple II



Tests de normalité : exemple III

```
> library(nortest)
> pearson.test(x)

        Pearson chi-square normality test

data:  x
P = 4, p-value = 0.2615

> lillie.test(x)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  x
D = 0.2032, p-value = 0.1863

> cvm.test(x)

        Cramer-von Mises normality test

data:  x
W = 0.0996, p-value = 0.09988
```

Tests de normalité : exemple IV

Conclusions

Sur cet exemple, on constate que

- Aucun des tests ne rejette l'hypothèse de normalité au niveau $\alpha = 5\%$.
- Le test de Pearson est le plus mauvais (degré de significativité le plus élevé). La construction des classes induit une perte d'information.
- CVM semble meilleur que KS (degré de significativité plus faible).

Plan partie 2

Rappel : test de signe I

Contexte

On observe un échantillon $X_1, \ldots, X_n$ de v.a. réelles i.i.d. On teste $H_0 : \mathbb{P}(X \leq 0) = 1/2$ *i.e.* "la médiane de la distribution est nulle" contre $H_1 : \mathbb{P}(X \leq 0) > 1/2$ *i.e.* "la médiane de la distribution est négative" ou $H_1' : \mathbb{P}(X \leq 0) < 1/2$ *i.e.* "la médiane de la distribution est positive".

Statistique de signe

$$S_n = \sum_{i=1}^{n} 1\{X_i \leq 0\} \sim \mathcal{B}(n, m),$$

où $m = \mathbb{P}(X \leq 0)$. Sous $H_0 : m = 1/2$, on a $S_n \sim \mathcal{B}(n, 1/2)$ et sous $H_1 : \mathbb{P}(X \leq 0) > 1/2$, la statistique S_n est stochastiquement plus grande que sous H_0. On rejette donc H_0 pour les grandes valeurs de S_n.

Rappel : test de signe II

Propriétés

- Pour les petites valeurs de n, la distribution $\mathcal{B}(n; 1/2)$ est tabulée. Pour les grandes valeurs de n, on a recours à une approximation Gaussienne.
- Ce test est très général, mais il utilise très peu d'information sur les variables (uniquement leur signe, pas leurs valeurs relatives). C'est donc un test peu puissant.
- Le test de signe et rang utilise plus d'information sur les variables.
- Remarque : c'est en fait un test paramétrique ! puisque la loi de S_n sous H_0 et sous l'alternative est paramétrique ($\mathcal{B}(n, m)$).

Digression : Statistiques d'ordre et de rang I

Définitions

Soient $X_1, \dots X_n$ v.a. réelles.

i) La statistique d'ordre $X^\star = (X_{(1)}, \dots, X_{(n)})$ est obtenue par réarrangement croissant des X_i.
Ainsi : $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ et
$\forall a \in \mathbb{R}, |\{i; X_i = a\}| = |\{i; X_{(i)} = a\}|$.

ii) Le vecteur des rangs R_X est une permutation de $\{1, \dots, n\}$ telle que $\forall i \in \{1, \dots, n\}$, on a $X_i = X^\star_{R_X(i)} = X_{(R_X(i))}$.

Digression : Statistiques d'ordre et de rang II

Exemple

$n = 7, x = (4, 2, 1, 1, 2, 0, 1)$. Alors $x^\star = (0, 1, 1, 1, 2, 2, 4)$ et par exemple

X	4	2	1	1	2	0	1
R_X	7	5	2	3	6	1	4

Ici on a $x_1 = 4 = x_{(7)}$ et $R_x(1) = 7$.

Remarques

- Cette notion est dépendante de n qui doit être fixé.
- S'il y a des ex-æquos, le vecteur des rangs n'est pas unique.

Digression : Statistiques d'ordre et de rang III

Théorème

Soient $X_1, \dots X_n$ v.a. réelles, i.i.d. de statistique d'ordre $X^\star$ et vecteur des rangs R_X. Alors on a $X^\star$ et R_X sont indépendants et de plus R_X est distribué uniformément sur l'ensemble $\mathcal{S}_n$ des permutations de $\{1, \dots, n\}$, i.e.

$$\mathbb{P}(R_1 = r_1, \dots, R_n = r_n) = 1/n!$$

Test de signe et rang (ou Wilcoxon signed rank test) I

Contexte

Soit $X_1, \dots X_n$ un échantillon de v.a. réelles de loi supposée diffuse et symétrique par rapport à (la médiane) m. On veut tester $H_0 : m = 0$ contre $H_1 : m \neq 0$.

Statistique de Wilcoxon

$$W_n^+ = \sum_{i=1}^{n} R_{|X|}(i) 1\{X_i > 0\},$$

où $R_{|X|}$ le vecteur des rangs associé à l'échantillon.

Test de signe et rang (ou Wilcoxon signed rank test) II

Exemple

$n = 5$ et on observe $\{-0.15; -0.42; 0.22; 0.6; -0.1\}$. Alors,

X_i	-0.15	-0.42	0.22	0.6	-0.1
$\lvert X_i \rvert$	0.15	0.42	0.22	0.6	0.1
$R_{\lvert X \rvert}(i)$	2	4	3	5	1
$X_i > 0$	$-$	$-$	$+$	$+$	$-$

et $W_5^+ = 3 + 5 = 8$ tandis que $W_5^- = 2 + 4 + 1 = 7$.

Test de signe et rang (ou Wilcoxon signed rank test) III

Remarques

- $W_n^+ + W_n^- = n(n+1)/2$ p.s.
  En effet, comme X est diffuse, on a $X_i \neq 0$ p.s. et donc $W_n^+ + W_n^- = \sum_{i=1}^n R_{|X|}(i) = \sum_{i=1}^n i = n(n+1)/2$.
- $0 \leq W_n^+ \leq n(n+1)/2$ p.s.
  Le cas $W_n^+ = 0$ correspond à tous les $X_i < 0$ et le cas $W_n^+ = n(n+1)/2$ à tous les $X_i > 0$.

Cas des ex-æquos

- Normalement le test s'applique à des variables diffuses, donc pas d'ex-æquos en théorie.
- En pratique, on affecte des rangs moyens en cas d'égalité et il existe des corrections des lois limites dans ce cas.

Test de signe et rang (ou Wilcoxon signed rank test) IV

Théorème

Sous l'hypothèse H_0 : "La loi de X est symétrique par rapport à 0", les statistiques W_n^+ et W_n^- ont la même distribution et sont des statistiques libres en loi. De plus, on a

$$\mathbb{E}_{H_0}(W_n^+) = \frac{n(n+1)}{4}, \quad \mathbb{V}\mathrm{ar}_{H_0}(W_n^+) = \frac{n(n+1)(2n+1)}{24}$$

$$\text{et} \quad \frac{W_n^+ - \mathbb{E}_{H_0}(W_n^+)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(W_n^+)}} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0,1) \text{ sous } H_0.$$

Conséquences

- Test de H_0 : "$m = 0$" contre $H_1 : m \neq 0$ en rejetant H_0 pour les grandes valeurs de $|W_n^+ - n(n+1)/4|$.
- Test exact via table pour $n \leq 20$, asymptotique (via l'approximation gaussienne) sinon.

Plan partie 2

Tests de comparaison de deux populations

Dans la suite, on distinguera :

- Le cas de deux populations appariées, qui se ramène au cas des tests sur une population.
- Le cas de deux populations non appariées : il s'agit là vraiment de tests sur deux populations.

Échantillons appariés I

On considère $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_n$ deux échantillons indépendants de v.a. diffuses de lois respectives F et G.

Appariement

- Soit il s'agit des mêmes individus, par exemple sur lesquels on applique des traitements à 2 temps différents,
- Soit les individus sont différents et alors pour que l'appariement soit valable, il faut avoir collecté puis regroupé les individus en fonction de covariables (sexe, âge, etc).

Échantillons appariés II

Tests d'homogénéité

- On veut tester H_0 : "$F = G$" contre H_1 : "$F \neq G$".
- Après appariement des variables, on construit $Z_i = X_i - Y_i$.
- Sous l'hypothèse H_0, la loi de Z est symétrique par rapport à sa médiane m, qui vaut 0. D'où le test induit H_0' : "$m = 0$" contre H_1' : "$m \neq 0$".
- On applique le test de signe et rang de Wilcoxon.

Échantillons non appariés I

Contexte

- On considère $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ deux échantillons indépendants de v.a. diffuses de lois respectives F et G.
- On veut tester $H_0 : "F = G"$ contre $H_1 : "F \neq G"$.
- Remarque : Les échantillons n'ont a priori pas la même taille et il n'existe pas d'appariement naturel entre les variables.

Échantillons non appariés II

Exemple

- Population de N individus sur lesquels on veut tester un nouveau traitement.
- On forme un groupe de n individus qui reçoivent le nouveau traitement et $m = N - n$ forment le groupe "contrôle", recevant un placebo.
- On mesure une quantité relative au traitement.
- L'hypothèse nulle H_0 : "$F = G$" est privilégiée : si on la rejette, le nouveau traitement est déclaré efficace. On ne veut pas d'un nouveau médicament si on n'est pas sûrs qu'il a un effet.

Échantillons non appariés III

Approches possibles

- Tests de Kolmogorov-Smirnov, Anderson-darling ou Cramér-von-Mises de comparaison de 2 échantillons.
- Test de la somme des rangs de Wilcoxon (ou test de Mann-Whitney).

Test de Kolmogorov Smirnov (et variantes) de comparaison de 2 échantillons

Statistiques pour deux échantillons

- Kolmogorov-Smirnov : $D_{n,m} = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - \hat{G}_m(x)|$,
- Cramér von Mises et Anderson-Darling : se généralisent à 2 échantillons.

On rejette H_0 pour les grandes valeurs de ces statistiques.

Propriétés

- Statistiques libre en loi sous H_0. Cette loi est tabulée.

Implémentation R

Fonction *ks.test()*. Paquets *adk* et *CvM2SL2Test*.

KS test pour deux échantillons : exemple I

Souris infectées par des larves

On s'intéresse à l'effet d'une dose faible de Cambendazole sur les infections des souris par la Trichinella Spiralis. Seize souris ont été infectées par un même nombre de larves de Trichinella et ensuite réparties au hasard entre deux groupes. Le premier groupe de huit souris a reçu du Cambendazole, à raison de 10 mg par kilo, 60 heures après l'infection. Les autres souris n'ont pas reçu de traitement. Au bout d'une semaine, toutes les souris ont été sacrifiées et les taux d'infection suivants (exprimés en pourcentages) sont mesurés :
Souris non traitées $\{79.5, 80.6, 61.4, 70.3, 73.8, 94.8, 49.6, 82.5\}$
Souris traitées $\{64.7, 49.2, 63.3, 52.4, 40.5, 76.8, 64.8, 64.1\}$.
Que conclure ?

KS test pour deux échantillons : exemple II

```
> x <- c(79.5,80.6,61.4,70.3,73.8,94.8,49.6,82.5)
>                   # souris non traitees
> y <- c(64.7,49.2,63.3,52.4,40.5,76.8,64.8, 64.1)
>                   # souris traitees
> ks.test(x,y, alternative="l")

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D^- = 0.625, p-value = 0.04394
alternative hypothesis: the CDF of x lies below that of y
```

Le test unilatère conclut à l'efficacité du traitement.

Test de la somme des rangs de Wilcoxon (ou test de Mann-Whitney) I

Procédure

On classe les variables $\{X_i, Y_j\}$ par leur rang global (*i.e.* on considère le vecteur des rangs $R_{X,Y}$) et on note $R_1, R_2, \ldots, R_n$ les rangs associés au premier échantillon (i.e. les X_i) et $N = n + m$.

Exemple

$X_1 = 3.5; X_2 = 4.7; X_3 = 1.2; Y_1 = 0.7; Y_2 = 3.9$ alors $Y_1 \leq X_3 \leq X_1 \leq Y_2 \leq X_2$ et les rangs associés à l'échantillon des X_i sont $R_1 = 3, R_2 = 5, R_3 = 2$.

Remarque

Suivant le contexte, X et Y peuvent mesurer des choses très différentes. Par contre, le rang relatif de ces variables est une quantité qui ne dépend pas de la nature (de la loi) des variables de départ.

Test de la somme des rangs de Wilcoxon (ou test de Mann-Whitney) II

Statistique de Mann-Whitney W_{YX}

On note $\Sigma_1 = R_1 + \cdots + R_n$ la somme des rangs du premier échantillon. On a p.s.

$$\frac{n(n+1)}{2} \leq \Sigma_1 \leq \frac{(n+m)(n+m+1)}{2} - \frac{m(m+1)}{2} = nm + \frac{n(n+1)}{2},$$

On définit

$$W_{YX} = \Sigma_1 - \frac{n(n+1)}{2},$$

On a $0 \leq W_{YX} \leq nm$ p.s.. On définit de façon symétrique $W_{XY} = \Sigma_2 - m(m+1)/2$ où Σ_2 est la somme des rangs du second échantillon.

Test de la somme des rangs de Wilcoxon (ou test de Mann-Whitney) III

Proposition

Sous l'hypothèse que les variables sont diffuses, on a les résultats suivants :

i) W_{XY} *est égal au nombre de paires* (X_i, Y_j) *(parmi les* nm *paires possibles) telles que* $X_i < Y_j$*.*

ii) $W_{XY} + W_{YX} = nm$*, p.s..*

iii) *Sous l'hypothèse* $H_0 : F = G$*, la loi de* Σ_1 *est symétrique par rapport à* $n(N+1)/2$*. Autrement dit, sous* H_0*, la loi de* W_{YX} *est symétrique par rapport à* $nm/2$*.*

iv) *Sous l'hypothèse* $H_0 : F = G$*, les variables* W_{XY} *et* W_{YX} *ont la même loi.*

Test de la somme des rangs de Wilcoxon (ou test de Mann-Whitney) IV

Théorème

La loi de W_{XY} (ou W_{YX}) est libre sous H_0. Cette loi ne dépend que de n et m. De plus,

$$\mathbb{E}_{H_0}(W_{XY}) = \frac{nm}{2}, \quad \mathbb{V}\mathrm{ar}_{H_0}(W_{XY}) = \frac{nm(N+1)}{12}$$

$$\text{et} \quad \frac{W_{XY} - \mathbb{E}_{H_0}(W_{XY})}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(W_{XY})}} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0,1) \text{ sous } H_0.$$

Test exact ou test asymptotique

- Le test rejette H_0 : "$F = G$" pour les grandes valeurs de $|W_{YX} - n(N+1)/2|$.
- Loi tabulée pour les petites valeurs de n et m (≤ 10).
- Pour les grandes valeurs, on utilise l'approximation gaussienne.

Lien entre Mann-Whitney et Wilcoxon

La stat. signe et rang de Wilcoxon peut être vue comme un cas particulier de la stat. somme des rangs de Wilcoxon. En effet,

- Soit $Z_1 \ldots, Z_N$ un échantillon,
- $U_1, \ldots, U_n$ sous-échantillon correspondant aux valeurs de Z_i telles que $Z_i > 0$,
- $V_1, \ldots, V_m$ sous-échantilon correspondant aux valeurs $-Z_i$ pour les $Z_i < 0$.
- Ordonner les $\{U_i, V_j\}$ revient à ordonner $\{|Z_i|\}$.
- La somme des rangs de l'échantillon des U_i est donc égale à la somme des rangs des $Z_i > 0$.
- Sous H_0, chacun des deux échantillons devrait être de taille environ $N/2$, mais il faut tenir compte de l'aléa dans la répartition des signes pour pouvoir faire un parallèle exact.

Test de Mann-Whitney : exemple I

Souris infectées par des larves (suite)

On reprend l'exemple précédent.

```
> wilcox.test(x,y,paired=F,alternative="g")

        Wilcoxon rank sum test

data:  x and y
W = 51, p-value = 0.02494
alternative hypothesis: true location shift is greater than
```

On conclut encore à l'efficacité du traitement, avec un degré de significativité plus faible que précédemment.

Test de Mann-Whitney : exemple II

Remarques sur R

- *wilcox.test* réalise le test de Mann-Whithney sur 2 échantillons lorsque l'option *paired=FALSE* est spécifiée.
- Noter que les alternatives unilatères ont des conventions opposées pour les tests *ks.test* et *wilcox.test*.

Remarques sur les tests de comparaison I

Remarques

- Le test de Mann-Whitney est très général et n'utilise que les valeurs relatives des variables entre elles.
- Le test d'homogénéité de KS pour 2 échantillons est assez différent car il prend en compte la forme des distributions et pas seulement des phénomènes de translation.

Remarques sur les tests de comparaison II

Détection d'une différence de forme

```
> x <- rnorm(500,0,1)
> y <- rnorm(700,0,1.2)
> ks.test(x,y)

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.0657, p-value = 0.161
alternative hypothesis: two-sided

> wilcox.test(x,y,paired=F)

        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 180731, p-value = 0.3329
alternative hypothesis: true location shift is not equal to
```

Plan partie 2

Contexte des tests de corrélation

Contexte

- On dispose de deux échantillons $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_n$ de v.a. réelles et appariées.
- Exemple : on mesure deux quantités X et Y sur un ensemble d'individus.
- On veut tester H_0 : "X et Y sont non corrélées" contre H_1 : "X et Y sont corrélées".
- NB : si les variables ne sont pas gaussiennes, "non corrélation" $\neq$ "indépendance".
- NB : Un test de permutation va tester H_0' : "X et Y sont indépendantes" et pas H_0 : "X et Y sont non corrélées".

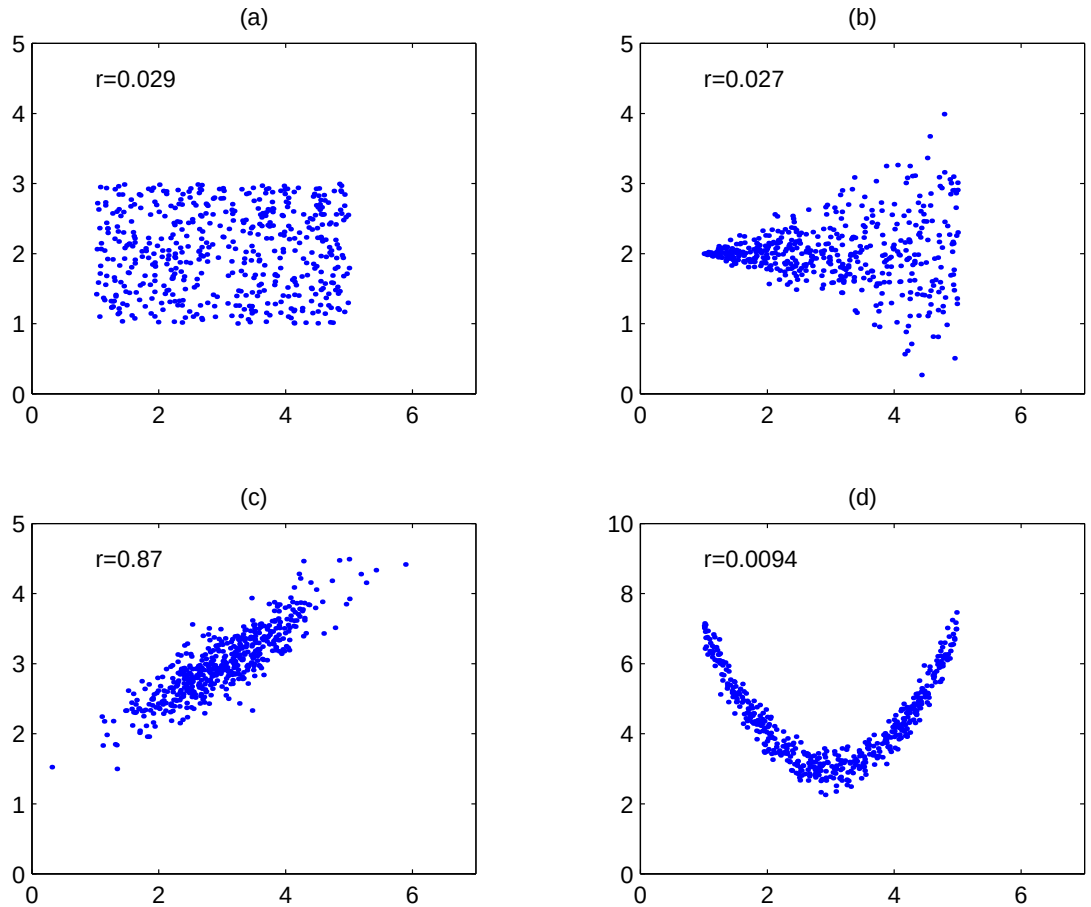
Remarque

Le test du χ^2 d'indépendance pour variables discrètes est un test paramétrique car les supports des variables sont finis !

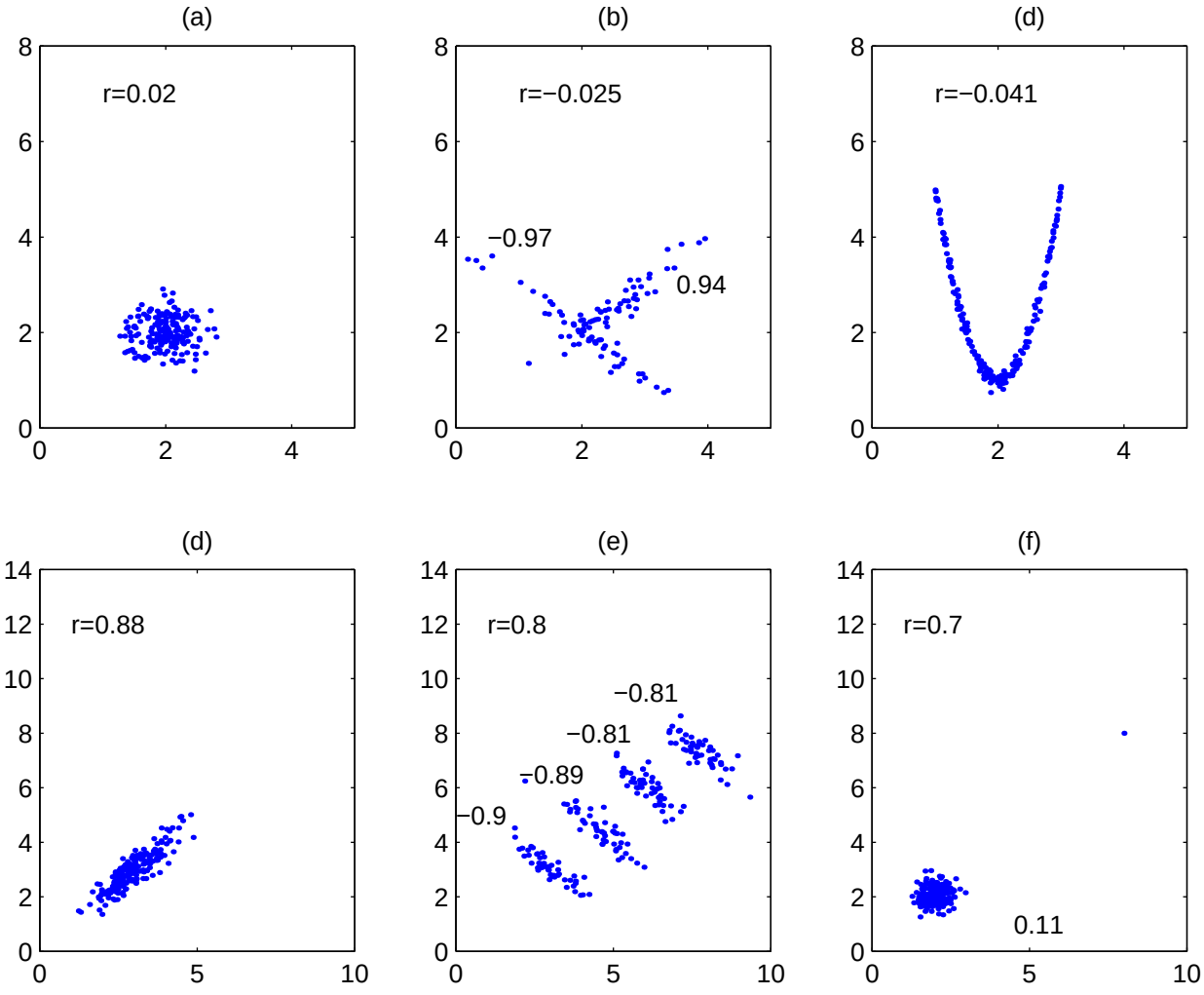
Corrélation de Pearson I

Rappels

Le coefficient de corrélation de Pearson mesure la dépendance linéaire entre deux variables réelles.



Corrélation de Pearson II



Corrélation de Pearson III

Coefficient de corrélation

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

Propriétés

- $-1 \leq r \leq 1$ avec égalité lorsque la relation entre X et Y est linéaire.
- La distribution exacte de r sous H_0 dépend des distributions de X et Y. Ce n'est pas une statistique libre en loi sous H_0.
- On peut utiliser un test de permutation dans le cas de H_0' : "X et Y sont indépendantes".
- Asymptotiquement, sous H_0, r suit une loi gaussienne centrée.

Corrélation de Pearson IV

Remarques

- La fonction *cor.test* de R implémente différents calculs et tests de corrélation. La méthode "pearson" de cette fonction implémente le test du coefficient de corrélation de Pearson lorsque les variables sont gaussiennes uniquement. À manipuler avec précaution dans le cas de petits échantillons !
- Pour obtenir une statistique libre en loi sous H_0, il faut se "débarrasser" des valeurs prises par les variables. Une possibilité est de passer par les rangs des variables.

Test de corrélation des rangs de Spearman I

Contexte

- Le coefficient de corrélation des rangs de Spearman mesure la dépendance monotone entre deux variables réelles et diffuses.
- C'est le coefficient de corrélation de Pearson entre les rangs des variables de deux échantillons.
- Soient $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_n$ deux échantillons appariés et $R_1, \ldots, R_n$, $S_1, \ldots, S_n$ les rangs respectifs des variables.
- On teste donc H_0 "X, Y non corrélées" contre H_1 : "X, Y sont en relation monotone".

Statistique de corrélation des rangs de Spearman

$$\rho = \frac{\sum_i (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_i (r_i - \bar{r})^2 \sum_i (s_i - \bar{s})^2}}.$$

On rejette H_0 pour des grandes valeurs de $|\rho|$.

Test de corrélation des rangs de Spearman II

Propriétés

- $-1 \leq \rho \leq 1$ et si $X = f(Y)$ avec f croissante (resp. décroissante) alors $\rho = +1$ (resp. -1).
- Sous H_0, la stat ρ est libre en loi.
- Test exact : en utilisant un test de permutation pour l'hyp H_0' (et pas H_0). Possible si n pas trop grand.
- Tests asymptotiques de H_0 : à partir de transformations de ρ.

Test de corrélation des rangs de Spearman III

Cas des ex-æquos

- Normalement le test s'applique à des variables diffuses, donc pas d'ex-æquos en théorie.
- En pratique, on affecte des rangs moyens en cas d'égalité.
- S'il n'y a pas d'ex-æquos, l'expression de ρ se simplifie et devient

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)},$$

où $d_i = r_i - s_i$ est la différence des rangs de l'individu i.

Test de corrélation des rangs de Kendall I

Contexte

- Le coefficient de corrélation des rangs de Kendall mesure la dépendance entre deux variables réelles et diffuses.
- Pour tout couple d'individus (i, j), on dit que les paires (x_i, y_i) et (x_j, y_j) sont concordantes si
  - soit $x_i < x_j$ et $y_i < y_j$,
  - soit $x_i > x_j$ et $y_i > y_j$,

  et discordante sinon.
- Cette fois encore, on teste H_0 "X, Y non corrélées" contre H_1 : "X, Y sont en relation monotone".

Test de corrélation des rangs de Kendall II

Statistique de corrélation des rangs de Kendall

$$\tau_n = \frac{(\text{Nb de paires concordantes}) - (\text{Nb de paires discordantes})}{n(n-1)/2}.$$

Cas des ex-æquos

- Normalement le test s'applique à des variables diffuses, donc pas d'ex-æquos en théorie.
- Il existe des variantes de la définition de τ_n qui prennent en compte le cas des ex-æquos.

Test de corrélation des rangs de Kendall III

Propriétés

- $-1 \leq \tau_n \leq 1$, et $\tau_n = 1$ lorsque la concordance entre les paires est parfaite, -1 lorsque la discordance est parfaite.
- Sous H_0, τ est une stat libre en loi, de moyenne nulle ($\mathbb{E}_{H_0}(\tau_n) = 0$). Loi tabulée pour n petit, qui donne un test exact.
- Test asymptotique pour n grand, fondé sur l'approximation gaussienne

$$\mathbb{E}_{H_0}(\tau_n) = 0, \quad \mathbb{V}\mathrm{ar}_{H_0}(\tau_n) = \frac{2(2n+5)}{9n(n-1)}$$

$$\frac{\tau_n - \mathbb{E}_{H_0}(\tau_n)}{\sqrt{\mathbb{V}\mathrm{ar}(\tau_n)}} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0;1) \text{ sous } H_0.$$

Tests de corrélation des rangs : exemple I

Relation linéaire, cas gaussien

```
> x <- rnorm(100,5,1) # var. gauss. moy. 5 et ecart-type 1
> perturb <- rnorm(100,0,0.5) # perturbation
> y <- 2*x+2+perturb # relation lineaire a perturbation
>                    # pres
> cor.test(x,y,method="pearson")

        Pearson's product-moment correlation

data:  x and y
t = 39.4183, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9554861 0.9796704
sample estimates:
      cor
0.9698819
```

Tests de corrélation des rangs : exemple II

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

        Spearman's rank correlation rho

data:  x and y
S = 5416, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9675008

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

Tests de corrélation des rangs : exemple III

```
        Kendall's rank correlation tau

data:  x and y
z = 12.4664, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.8456566
```

Les trois méthodes détectent la relation linéaire entre les variables.

Tests de corrélation des rangs : exemple IV

Relation linéaire, cas non gaussien

```
> x <- runif(5) # var. non gaussiennes
> perturb <- rnorm(5,0,0.1) # perturbation
> y <- 2*x+2+perturb # relation lineaire a perturbation
>                    # pres
> cor.test(x,y,method="pearson")

        Pearson's product-moment correlation

data:  x and y
t = 17.3, df = 3, p-value = 0.0004209
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9233280 0.9996881
sample estimates:
      cor
0.9950255
```

Tests de corrélation des rangs : exemple V

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

        Spearman's rank correlation rho

data:  x and y
S = 0, p-value = 0.01667
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
  1

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

Tests de corrélation des rangs : exemple VI

```
        Kendall's rank correlation tau

data:  x and y
T = 10, p-value = 0.01667
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
  1
```

- La méthode "pearson" détecte la relation linéaire, mais on ne doit pas s'y fier (p-value fausse *a priori* car échantillon petit et non gaussien).
- Les deux autres méthodes s'appliquent sans problèmes.

Tests de corrélation des rangs : exemple VII

Relation monotone, non linéaire, cas non gaussien

```
> x <- runif(10,100,200) # var. non gaussiennes
> perturb <- rnorm(10,0,1) # perturbation
> y <- x^3+perturb # relation croissante a
>                          # perturbation pres
> cor.test(x,y,method="pearson")

        Pearson's product-moment correlation

data:  x and y
t = 14.738, df = 8, p-value = 4.417e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9234763 0.9958984
sample estimates:
      cor
0.9820782
```

Tests de corrélation des rangs : exemple VIII

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

        Spearman's rank correlation rho

data:  x and y
S = 0, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
  1

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

Tests de corrélation des rangs : exemple IX

```
        Kendall's rank correlation tau

data:  x and y
T = 45, p-value = 5.511e-07
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
  1
```

- La méthode "pearson" ne devrait pas être appliquée dans ce cas,
- La méthode "spearman" fonctionne très bien,
- La méthode "kendall" donne un degré de significativité plus grande que "spearman" (mais quand même très faible).

Références


[De Wet and Randles 87] T. De Wet and R. Randles.
On the effect of substituting parameter estimators in limiting χ^2, u and v statistics.
Annals of Statistics, 15 :398–412, 1987.


[Morris 75] C. Morris.
Central limit theorems for multinomial sums.
The Annals of Statistics, 3 :165–188, 1975.


[Nikitin 95] Y. Nikitin.
Asymptotic efficiency of nonparametric tests.
Cambridge University Press, 1995.