

Tests basés sur la fonction de répartition empirique et sur les rangs

C. Maugis-Rabusseau
Bureau 116, GMM
cathy.maugis@insa-toulouse.fr

4MA, 2020-2021

1 Rappels

- Fonction de répartition et quantiles
- Fonction de répartition empirique

2 Test de Kolmogorov de comparaison ou d'adéquation

3 Tests de comparaison de deux échantillons

- Tests de Kolmogorov-Smirnov
- Test de Wilcoxon- Mann-Whitney
- Test de la médiane

4 Tests de normalité

- Méthode graphique : droite de Henry
- Test de normalité de Kolmogorov-Smirnov (de Lilliefors)
- Test de Shapiro-Wilk

1 Rappels

- Fonction de répartition et quantiles
- Fonction de répartition empirique

2 Test de Kolmogorov de comparaison ou d'adéquation

3 Tests de comparaison de deux échantillons

- Tests de Kolmogorov-Smirnov
- Test de Wilcoxon- Mann-Whitney
- Test de la médiane

4 Tests de normalité

- Méthode graphique : droite de Henry
- Test de normalité de Kolmogorov-Smirnov (de Lilliefors)
- Test de Shapiro-Wilk

- X v.a.r de **fonction de répartition** F :

$$\forall t \in \mathbb{R}, F(t) = \mathbb{P}(X \leq t).$$

- La **fonction quantile (ou inverse généralisée)** F^{-1} de F par

$$\forall p \in [0, 1], F^{-1}(p) = \inf \{t \in \mathbb{R}; F(t) \geq p\}.$$

- Si F est une bijection, F^{-1} est la bijection réciproque.
- Exo : Calculez F^{-1} pour la loi de Bernoulli de paramètre θ .

Fonction de répartition et quantiles

Soit $t \in \mathbb{R}$ et $p_0 \in]0, 1[$.

- 1 F croissante, continue à droite, $\lim_{t \rightarrow -\infty} F(t) = 0$, $\lim_{t \rightarrow +\infty} F(t) = 1$
- 2 $\{t \in \mathbb{R}; F(t) \geq p_0\} = [F^{-1}(p_0), +\infty[$
- 3 F^{-1} est croissante
- 4 $F \circ F^{-1}(p_0) \geq p_0$ avec égalité si $p_0 \in F(\mathbb{R})$
- 5 $F(t) \geq p_0 \Leftrightarrow t \geq F^{-1}(p_0)$
- 6 Si $U \sim \mathcal{U}([0, 1])$, $F^{-1}(U)$ a pour fonction de répartition F

1 Rappels

- Fonction de répartition et quantiles
- Fonction de répartition empirique

2 Test de Kolmogorov de comparaison ou d'adéquation

3 Tests de comparaison de deux échantillons

- Tests de Kolmogorov-Smirnov
- Test de Wilcoxon- Mann-Whitney
- Test de la médiane

4 Tests de normalité

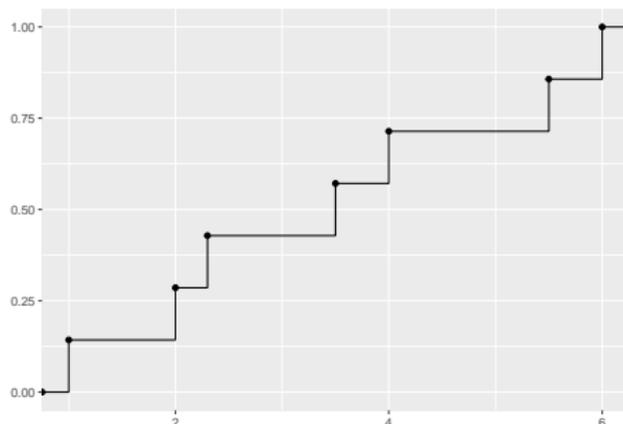
- Méthode graphique : droite de Henry
- Test de normalité de Kolmogorov-Smirnov (de Lilliefors)
- Test de Shapiro-Wilk

Fonction de répartition empirique

- (X_1, X_2, \dots, X_n) v.a.r i.i.d. de fonction de répartition F .
- La **fonction de répartition empirique** associée :

$$\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq t} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_{(i)} \leq t}$$

avec les statistiques d'ordre $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$



Fonction de répartition empirique

- \widehat{F}_n croissante, continue à droite, $\lim_{t \rightarrow -\infty} \widehat{F}_n(t) = 0$, $\lim_{t \rightarrow +\infty} \widehat{F}_n(t) = 1$
- $\forall t \in \mathbb{R}, n\widehat{F}_n(t) \sim \mathcal{B}(n, F(t))$
- $\forall t \in \mathbb{R}, \mathbb{E} [\widehat{F}_n(t)] = F(t)$
- $\forall t \in \mathbb{R}$

$$\text{Var}(\widehat{F}_n(t)) = \frac{F(t)(1 - F(t))}{n} \xrightarrow{n \rightarrow +\infty} 0$$

- $\forall t \in \mathbb{R}, \widehat{F}_n(t) \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} F(t)$
- On déduit du TLC que pour tout $t \in \mathbb{R}$ tel que $F(t)(1 - F(t)) \neq 0$,

$$\sqrt{n}(\widehat{F}_n(t) - F(t)) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, F(t)(1 - F(t)))$$

- Glivenko-Cantelli (admis)

$$\sup_{t \in \mathbb{R}} |\widehat{F}_n(t) - F(t)| \xrightarrow[n \rightarrow +\infty]{p.s.} 0$$

1 Rappels

- Fonction de répartition et quantiles
- Fonction de répartition empirique

2 Test de Kolmogorov de comparaison ou d'adéquation

3 Tests de comparaison de deux échantillons

- Tests de Kolmogorov-Smirnov
- Test de Wilcoxon- Mann-Whitney
- Test de la médiane

4 Tests de normalité

- Méthode graphique : droite de Henry
- Test de normalité de Kolmogorov-Smirnov (de Lilliefors)
- Test de Shapiro-Wilk

- X_1, \dots, X_n v.a.r i.i.d. de même loi que X , de fonction de répartition F supposée continue
- Une fonction de répartition F_0 donnée, supposée continue sur \mathbb{R} et Y_0 une v.a.r de fonction de répartition F_0
- On souhaite construire un test de \mathcal{H}_0 : " X et Y_0 ont la même loi ($F = F_0$)" contre :
 - 1 \mathcal{H}_1 : X et Y_0 ne suivent pas la même loi ($F \neq F_0$)
 - 2 \mathcal{H}_1^+ : X a tendance à prendre des valeurs plus petites que Y_0 ($F \geq F_0$)
 - 3 \mathcal{H}_1^- : X a tendance à prendre des valeurs plus grandes que Y_0 ($F \leq F_0$)

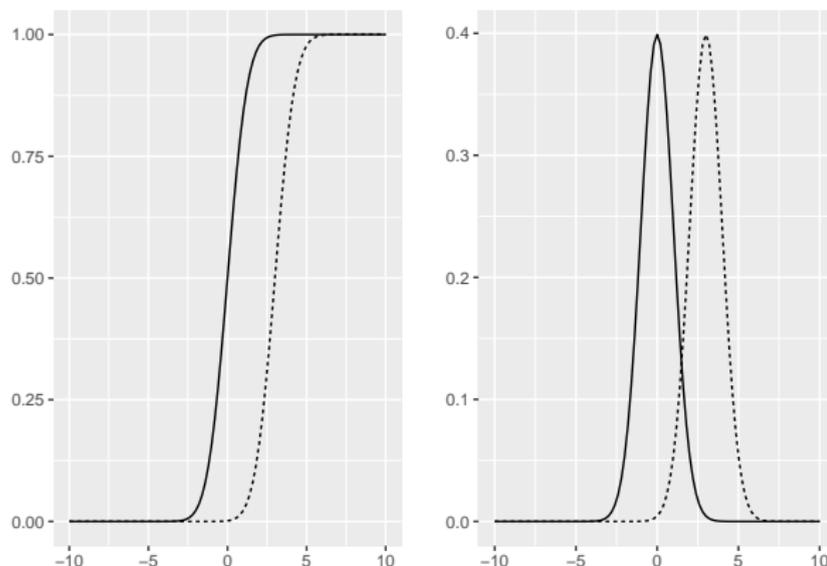


FIGURE – Fonction de répartition (à gauche) et densité (à droite) pour la loi $\mathcal{N}(0, 1)$ et $\mathcal{N}(3, 1)$

Exemple

On mesure les durées de vie de 20 ampoules d'un même type.
Les résultats, en heures, sont :

673, 389, 1832, 570, 522, 2694, 3683, 644, 1531, 2916

Est-ce que l'on peut affirmer, au risque 5%, que la durée de vie d'une ampoule de ce type ne suit pas la loi exponentielle $\mathcal{E}(1/1500)$?

On modélise donc la durée de vie de la i ème ampoule par X_i ,
 F est sa fonction de répartition inconnue
 F_0 est la fonction de répartition de la loi $\mathcal{E}(1/1500)$.

Test de Kolmogorov

- $\mathcal{H}_0 : F = F_0$ contre $\mathcal{H}_1 : F \neq F_0$
- Statistique de test :

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

- Région de rejet au niveau α

$$\mathcal{R}_\alpha = \{D_n \geq d_{n,1-\alpha}\}$$

- La loi de D_n sous l'hypothèse H_0 ($F = F_0$) est indépendante de F_0 .
- Autre écriture de la statistique de test :

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

Test de Kolmogorov

- $\mathcal{H}_0 : F = F_0$ contre $\mathcal{H}^+ : F \geq F_0$, on utilise

$$D_n^+ = \sup_{t \in \mathbb{R}} (\widehat{F}_n(t) - F_0(t))$$

et la région de rejet de niveau α est de la forme

$$\mathcal{R}_\alpha = \{D_n^+ > d_{n,1-\alpha}^+\}$$

- $\mathcal{H}_0 : F = F_0$ contre $\mathcal{H}^- : F \leq F_0$, on utilise

$$D_n^- = \sup_{t \in \mathbb{R}} (F_0(t) - \widehat{F}_n(t))$$

et la région de rejet de niveau α est de la forme

$$\mathcal{R}_\alpha = \{D_n^- > d_{n,1-\alpha}^-\}$$

- $\forall \lambda > 0, \mathbb{P}_{\mathcal{H}_0}(\sqrt{n}D_n^+ \geq \lambda) \xrightarrow{n \rightarrow +\infty} \exp(-2\lambda^2)$ (Smirnov, 1942)
- $\forall \lambda > 0, \mathbb{P}_{\mathcal{H}_0}(\sqrt{n}D_n \geq \lambda) \xrightarrow{n \rightarrow +\infty} 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2\lambda^2)$
(Kolmogorov, 1933)
- $\forall \lambda > 0, \mathbb{P}_{\mathcal{H}_0}(\sqrt{n}D_n \geq \lambda) \leq 2 \exp(-2\lambda^2)$ (Massart, 1990)

Exemple : Durée de vie des ampoules

$$\mathcal{H}_0 : F = F_0 \text{ contre } \mathcal{H}_1 : F \neq F_0,$$

où F_0 est la fonction de répartition de $\mathcal{E}(1/1500)$

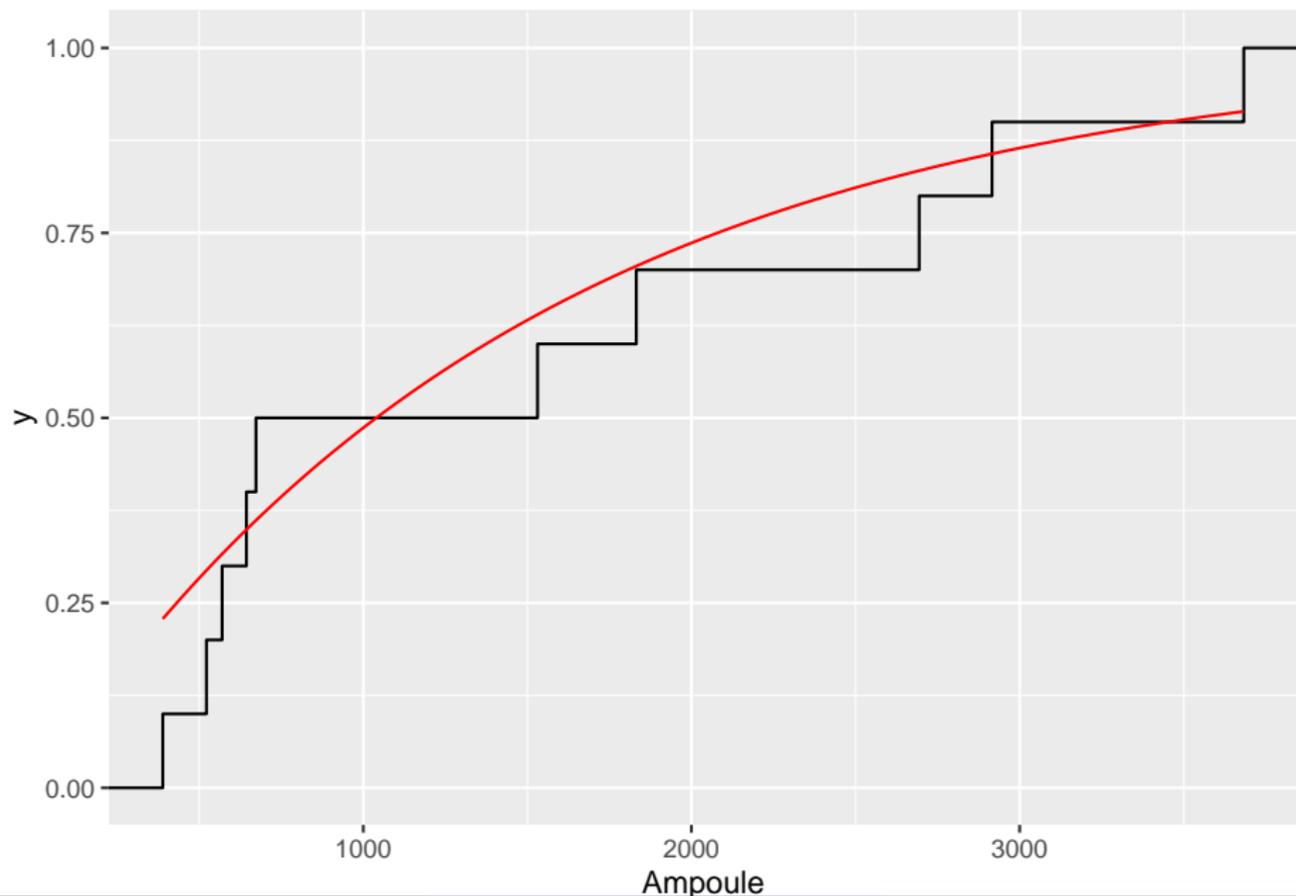
```
> Ampoule = c(673, 389, 1832, 570, 522, 2694, 3683, 644, 1531, 2916)
```

```
> ggplot(data.frame(Ampoule), aes(Ampoule)) +  
  stat_ecdf(geom = "step") +  
  stat_function(fun = pexp, args = list(rate = 1/1500), col="red")
```

```
> ks.test(Ampoule,pexp,1/1500,alternative="two.sided")  
One-sample Kolmogorov-Smirnov test
```

```
data: Ampoule  
D = 0.22843, p-value = 0.597  
alternative hypothesis: two-sided
```

Exemple : Durée de vie des ampoules



1 Rappels

- Fonction de répartition et quantiles
- Fonction de répartition empirique

2 Test de Kolmogorov de comparaison ou d'adéquation

3 Tests de comparaison de deux échantillons

- Tests de Kolmogorov-Smirnov
- Test de Wilcoxon- Mann-Whitney
- Test de la médiane

4 Tests de normalité

- Méthode graphique : droite de Henry
- Test de normalité de Kolmogorov-Smirnov (de Lilliefors)
- Test de Shapiro-Wilk

- On considère deux échantillons indépendants
 - X_1, \dots, X_n i.i.d. de fonction de répartition F
 - Y_1, \dots, Y_m i.i.d. de fonction de répartition G
- $N = n + m$
- Dans le cas de deux échantillons gaussiens -> cf cours MIC3
- Ici, cadre non paramétrique : les lois des variables X_i et Y_j ne sont pas supposées connues

1 Rappels

- Fonction de répartition et quantiles
- Fonction de répartition empirique

2 Test de Kolmogorov de comparaison ou d'adéquation

3 Tests de comparaison de deux échantillons

- Tests de Kolmogorov-Smirnov
- Test de Wilcoxon- Mann-Whitney
- Test de la médiane

4 Tests de normalité

- Méthode graphique : droite de Henry
- Test de normalité de Kolmogorov-Smirnov (de Lilliefors)
- Test de Shapiro-Wilk

Test de Kolmogorov-Smirnov

- $\mathcal{H}_0 : F = G$ contre $\mathcal{H}_1 : F \neq G$
- \widehat{F}_n fonction de répartition empirique de (X_1, \dots, X_n)
- \widehat{G}_m fonction de répartition empirique de (Y_1, \dots, Y_m)
- Statistique de test

$$D_{n,m} = \sup_{t \in \mathbb{R}} |\widehat{F}_n(t) - \widehat{G}_m(t)|.$$

- Région de rejet au niveau α

$$\mathcal{R}_\alpha = \{D_{n,m} \geq d_{n,m,1-\alpha}\}$$

- Si F est continue, la loi de $D_{n,m}$ sous \mathcal{H}_0 est indépendante de F .

Test de Kolmogorov-Smirnov

Pour faire un test unilatéral

$$\mathcal{H}_0 : F = G \text{ contre } \mathcal{H}_1 : F \geq G$$

on utilise la statistique de test

$$D_{n,m}^+ = \sup_{t \in \mathbb{R}} (\hat{F}_n(t) - \hat{G}_m(t)).$$

La région de rejet au niveau α est de la forme

$$\mathcal{R}_\alpha = \{D_{n,m}^+ \geq d_{n,m,1-\alpha}^+\}$$

Exemple : Comparaison de deux médicaments

On souhaite comparer deux médicaments pour soulager la douleur post-opératoire.

On a observé 16 patients :

- 8 ont pris le médicament A habituel
- les 8 autres un médicament B expérimental

Dans le tableau suivant sont reportés les temps (en heures) entre la prise du médicament et la sensation de soulagement.

médicament A	6,8	3,1	5,8	4,5	3,3	4,7	4,2	4,9
médicament B	4,4	2,5	2,8	2,1	6,6	1,5	4,8	2,3

Exemple : Comparaison de deux médicaments

Test d'une différence d'efficacité entre les deux médicaments

$\mathcal{H}_0 : F_A = F_B$ contre $\mathcal{H}_1 : F_B \neq F_A$

```
> ks.test(mB, mA,  
         alternative="two.sided")
```

Two-sample Kolmogorov-Smirnov test

data: mB and mA

D = 0.625, p-value = 0.08702

alternative hypothesis: two-sided

Tester si le médicament B est plus efficace que A

$\mathcal{H}_0 : F_A = F_B$ contre $\mathcal{H}_1 : F_B \geq F_A$:

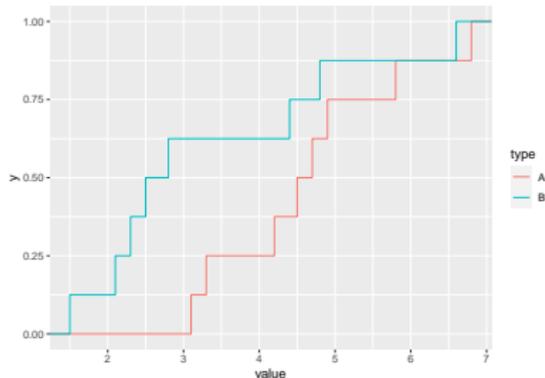
```
> ks.test(mB, mA,  
         alternative="greater")
```

Two-sample Kolmogorov-Smirnov test

data: mB and mA

D⁺ = 0.625, p-value = 0.04394

alternative hypothesis: the CDF of x lies above that of y



1 Rappels

- Fonction de répartition et quantiles
- Fonction de répartition empirique

2 Test de Kolmogorov de comparaison ou d'adéquation

3 Tests de comparaison de deux échantillons

- Tests de Kolmogorov-Smirnov
- Test de Wilcoxon- Mann-Whitney
- Test de la médiane

4 Tests de normalité

- Méthode graphique : droite de Henry
- Test de normalité de Kolmogorov-Smirnov (de Lilliefors)
- Test de Shapiro-Wilk

Test de Mann-Whitney

- $\mathcal{H}_0 : F = G$ contre $\mathcal{H}_1 : F \geq G$
- On suppose que F et G sont continues
- Le principe du test consiste à déterminer le nombre de couples (X_i, Y_j) pour lesquels $Y_j > X_i$
- Statistique de test :

$$MW_{X < Y} = \sum_{i=1}^n \sum_{j=1}^m \mathbb{1}_{Y_j > X_i}$$

- La région de rejet au niveau α est de la forme

$$\mathcal{R}_\alpha = \{MW_{X < Y} \geq u_{(n,m), 1-\alpha}\}$$

Test de Mann-Whitney

- La loi de $MW_{X<Y}$ sous \mathcal{H}_0 peut être établie par récurrence (cf Caperaa Van Cutsem, p 126)
- On peut aussi utiliser un résultat asymptotique (Hajek, 1968) :
Sous \mathcal{H}_0 , quand $n \rightarrow +\infty$, $\frac{n}{n+m} \rightarrow \lambda \in]0, 1[$,

$$\frac{MW_{X<Y} - \mathbb{E}_{\mathcal{H}_0}[MW_{X<Y}]}{\sqrt{\text{Var}_{\mathcal{H}_0}(MW_{X<Y})}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

On utilise ce résultat en pratique si $n, m \geq 8$.

On a, sous l'hypothèse nulle, que

$$\mathbb{E}_{\mathcal{H}_0}[MW_{X<Y}] = \frac{mn}{2} \text{ et } \text{Var}_{\mathcal{H}_0}(MW_{X<Y}) = mn \left(\frac{n+m+1}{12} \right).$$

Test de Mann-Whitney

- Tester $\mathcal{H}_0 : F = G$ contre $\mathcal{H}_1 : F \leq G$:

$$MW_{X>Y} = \sum_{i=1}^n \sum_{j=1}^m \mathbb{1}_{X_i > Y_j}.$$

La région de rejet au niveau α est de la forme

$$\mathcal{R}_\alpha = \{MW_{X>Y} \geq \tilde{u}_{(n,m),1-\alpha}\}.$$

- Tester $\mathcal{H}_0 : F = G$ contre $\mathcal{H}_1 : F \neq G$:

$$MW_{X,Y} = \min(MW_{X<Y}, MW_{X>Y}).$$

La région de rejet au niveau α est de la forme

$$\mathcal{R}_\alpha = \{MW_{X,Y} \geq v_{(n,m),1-\alpha}\}.$$

Test des rangs de Wilcoxon

- Ech. complet : $Z = (Z_1, \dots, Z_N) = (X_1, \dots, X_n, Y_1, \dots, Y_m)$
- Tester $\mathcal{H}_0 : F = G$ contre $\mathcal{H}_1 : F \geq G$
 - R_j le rang de Y_j dans $Z_{(\cdot)}$: $R_j = \sum_{k=1}^N \mathbb{1}_{Z_k < Y_j} + 1$.
 - La statistique de Wilcoxon :

$$W_Y = \sum_{j=1}^m R_j$$

- Relation entre Wilcoxon et Mann-Whitney :

$$MW_{X < Y} = W_Y - \frac{m(m+1)}{2} \text{ donc ils conduisent au m\^eme test.}$$

- Tester $\mathcal{H}_0 : F = G$ contre $\mathcal{H}_1 : F \leq G$
 - R_i le rang de X_i dans $Z_{(\cdot)}$
 - La statistique de Wilcoxon : $W_X = \sum_{i=1}^n R_i$.
- Remarque : $W_X + W_Y = \sum_{k=1}^N k = \frac{N(N+1)}{2}$

Exemple

On reprend l'exemple des médicaments.

On veut tester $\mathcal{H}_0 : F_A = F_B$ contre $\mathcal{H}_1 : F_B \geq F_A$

$$\begin{aligned}z_{(.)} &= (1.5, 2.1, 2.3, 2.5, 2.8, 3.1, 3.3, 4.2, 4.4, 4.5, 4.7, 4.8, 4.9, 5.8, 6.6, 6.8) \\ &= (mB_6, mB_4, mB_8, mB_2, mB_3, mA_2, mA_5, mA_7, mB_1, mA_4, mA_6, mB_7, mA_8, mA_3, mB_5, mA_1)\end{aligned}$$

Les rangs observés pour les valeurs de B valent donc

$$R_1 = 9, R_2 = 4, R_3 = 5, R_4 = 2, R_5 = 15, R_6 = 1, R_7 = 12, R_8 = 3$$

ce qui donne $W_B = 51$ et $W_A = (16 \times 17)/2 - 51 = 85$.

On a également que

$$MW_{B < A} = \sum_{i=1}^8 \sum_{j=1}^8 \mathbb{1}_{B_i < A_j} = 5+5+5+6+6+7+7+8 = 49 = W_A - (8 \times 9)/2$$

et

$$MW_{B > A} = \sum_{i=1}^8 \sum_{j=1}^8 \mathbb{1}_{B_i > A_j} = 3+3+3+2+2+1+1+0 = 15 = W_B - (8 \times 9)/2.$$

Exemple

On reprend l'exemple des médicaments.

On veut tester $\mathcal{H}_0 : F_A = F_B$ contre $\mathcal{H}_1 : F_B \geq F_A$

```
> wilcox.test(mB,mA,alternative="less")
```

Wilcoxon rank sum test

data: mB and mA

W = 15, p-value = 0.04149

alternative hypothesis: true location shift is less than 0

Traitement des ex-aequos

Nous avons supposé les lois continues, donc la probabilité d'avoir des ex-aequos est nulle. En pratique, soit parce que les lois ne sont pas continues, soit parce qu'on a des mesures arrondies, on peut avoir des ex-aequos.

- Correction des statistiques de test de Mann-Whitney :

$$\tilde{M}W_{X<Y} = \sum_{i=1}^n \sum_{j=1}^m \left\{ \mathbb{1}_{X_i < Y_j} + \frac{1}{2} \mathbb{1}_{X_i = Y_j} \right\}$$

et

$$\tilde{M}W_{X>Y} = \sum_{i=1}^n \sum_{j=1}^m \left\{ \mathbb{1}_{X_i > Y_j} + \frac{1}{2} \mathbb{1}_{X_i = Y_j} \right\}$$

Rem : $\tilde{M}W_{X<Y} + \tilde{M}W_{X>Y} = nm$.

- Test de Wilcoxon : On corrige les rangs R_j en utilisant les rangs moyens. Le rang de tous les éléments d'un groupe d'ex-aequos est la moyenne des rangs des éléments du groupe.

Exemple

On considère les valeurs observées suivantes pour les deux échantillons :

$\underline{x} = (5, 3, 6, 8, 1, 6)$ avec $n = 6$ et $\underline{y} = (5, 7, 9, 5, 2)$ avec $m = 5$.

$x_{(.)}$	1	3	5	6	6	8
$y_{(.)}$	2	5	5		7	9
\tilde{R}_i	1	3	5	7.5	7.5	10
\tilde{R}_j	2	5	5		9	11

Ainsi

$$\left(\tilde{M}W_{X<Y}\right)^{obs} = 1 + \left(2 + \frac{1}{2}\right) + \left(2 + \frac{1}{2}\right) + 5 + 6 = 17,$$

$$\left(\tilde{W}_Y\right)^{obs} = \sum_{j=1}^5 \tilde{R}_j = 2 + 5 + 5 + 9 + 11 = 32$$

et on retrouve bien que $\left(\tilde{M}W_{X<Y}\right)^{obs} = \tilde{W}_Y^{obs} - \frac{5 \times 6}{2}$.

1 Rappels

- Fonction de répartition et quantiles
- Fonction de répartition empirique

2 Test de Kolmogorov de comparaison ou d'adéquation

3 Tests de comparaison de deux échantillons

- Tests de Kolmogorov-Smirnov
- Test de Wilcoxon- Mann-Whitney
- Test de la médiane

4 Tests de normalité

- Méthode graphique : droite de Henry
- Test de normalité de Kolmogorov-Smirnov (de Lilliefors)
- Test de Shapiro-Wilk

Test de la médiane

- $\mathcal{H}_0 : F = G$ contre $\mathcal{H}_1 : F \geq G$, F et G sont continues.
- Statistique de test :

$$M_{X,Y} = \frac{1}{m} \sum_{j=1}^m \mathbb{1}_{R_j > \frac{N+1}{2}}.$$

- Région de rejet au niveau α : $\mathcal{R}_\alpha = \{M_{X,Y} \geq m_{n,m,1-\alpha}\}$.
- Loi de $M_{X,Y}$ sous \mathcal{H}_0 :
 - Si N pair, $nM_{X,Y} \sim \mathcal{H}(N, \frac{N}{2}, m)$
 $\implies \mathbb{E}_{\mathcal{H}_0}[M_{X,Y}] = \frac{1}{2}$ et $\text{Var}_{\mathcal{H}_0}(M_{X,Y}) = \frac{n}{4m(N-1)}$
 - Si N est impair, $nM_{X,Y} \sim \mathcal{H}(N, \frac{N-1}{2}, m)$
 $\implies \mathbb{E}_{\mathcal{H}_0}[M_{X,Y}] = \frac{N-1}{2N}$ et $\text{Var}_{\mathcal{H}_0}(M_{X,Y}) = \frac{n(N+1)}{4mN^2}$
- Pour $n, m \geq 30$, sous \mathcal{H}_0 ,

$$M_{X,Y} \stackrel{\mathcal{L}}{\simeq} \mathcal{N}(\mathbb{E}_{\mathcal{H}_0}[M_{X,Y}], \text{Var}_{\mathcal{H}_0}(M_{X,Y}))$$

- 1 **Rappels**
 - Fonction de répartition et quantiles
 - Fonction de répartition empirique

- 2 **Test de Kolmogorov de comparaison ou d'adéquation**

- 3 **Tests de comparaison de deux échantillons**

- Tests de Kolmogorov-Smirnov
- Test de Wilcoxon- Mann-Whitney
- Test de la médiane

- 4 **Tests de normalité**

- Méthode graphique : droite de Henry
- Test de normalité de Kolmogorov-Smirnov (de Lilliefors)
- Test de Shapiro-Wilk

- (X_1, \dots, X_n) de fonction de répartition F
- \hat{F}_n la fonction de répartition empirique associée à cet échantillon
- Pour illustration des méthodes :
trois échantillons de taille $n = 200$:

Ech1 : un échantillon simulé selon la loi $\mathcal{N}(2, 1)$

Ech2 : un échantillon simulé selon la loi uniforme sur l'intervalle $[2, 4]$

Ech3 : un échantillon simulé selon la loi de Cauchy de paramètre 1

1 Rappels

- Fonction de répartition et quantiles
- Fonction de répartition empirique

2 Test de Kolmogorov de comparaison ou d'adéquation

3 Tests de comparaison de deux échantillons

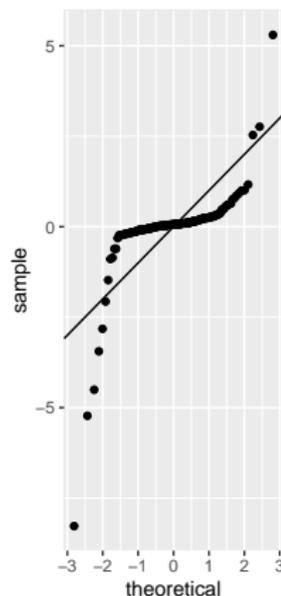
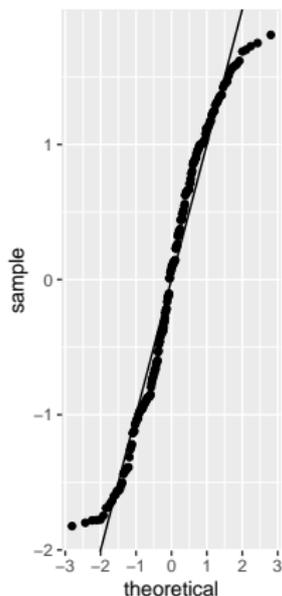
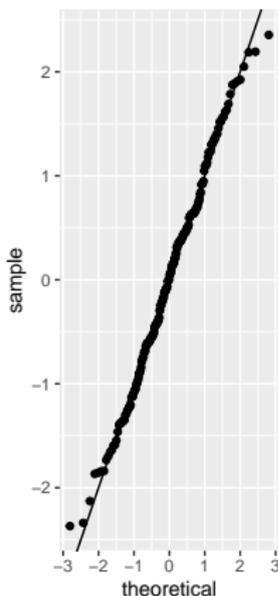
- Tests de Kolmogorov-Smirnov
- Test de Wilcoxon- Mann-Whitney
- Test de la médiane

4 Tests de normalité

- Méthode graphique : droite de Henry
- Test de normalité de Kolmogorov-Smirnov (de Lilliefors)
- Test de Shapiro-Wilk

Méthode graphique : droite de Henry

- Droite de Henry = "Normal Probability Plot" = "Q-Q Plot"
- On représente les points $(X_{(i)}, \Phi^{-1} \circ \hat{F}_n(X_{(i)}))$, où Φ fonction de répartition de $\mathcal{N}(0, 1)$.
- Sous l'hypothèse que les X_i sont i.i.d. de loi normale, les points $(X_{(i)}, \Phi^{-1} \circ \hat{F}_n(X_{(i)}))$ sont pratiquement alignés.



1 Rappels

- Fonction de répartition et quantiles
- Fonction de répartition empirique

2 Test de Kolmogorov de comparaison ou d'adéquation

3 Tests de comparaison de deux échantillons

- Tests de Kolmogorov-Smirnov
- Test de Wilcoxon- Mann-Whitney
- Test de la médiane

4 Tests de normalité

- Méthode graphique : droite de Henry
- Test de normalité de Kolmogorov-Smirnov (de Lilliefors)
- Test de Shapiro-Wilk

Test de normalité de K.-S. (de Lilliefors)

- \mathcal{H}_0 : “les X_i suivent une loi normale”
 \mathcal{H}_1 : “les X_i ne suivent pas une loi normale”.
- Statistique de test

$$DN_n = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - \Phi(t; \bar{X}, S^2)|$$

où $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ et $\Phi(\cdot; \bar{X}, S^2)$ f.d.r. de $\mathcal{N}(\bar{X}, S^2)$

- La région de rejet au niveau α : $\mathcal{R}_\alpha = \{DN_n > d_{n,1-\alpha}\}$
- Sous \mathcal{H}_0 (i.e $X_i \sim \mathcal{N}(\mu, \sigma^2)$), la loi de DN_n ne dépend pas des paramètres inconnus (μ, σ^2) . Il s'agit de la loi de

$$KSN_n = \sup_{t \in \mathbb{R}} \left| \hat{\Phi}_n(t) - \Phi\left(\frac{t - \hat{\mu}_{\mathcal{Z}}}{S_{\mathcal{Z}}}\right) \right|$$

où $\mathcal{Z} = (Z_1, \dots, Z_n)$ i.i.d de loi $\mathcal{N}(0, 1)$, $\hat{\Phi}_n$ la fonction de répartition empirique de \mathcal{Z} , $\hat{\mu}_{\mathcal{Z}}$ la moyenne empirique et $S_{\mathcal{Z}}^2$ la variance empirique de \mathcal{Z} .

Exemples

On applique le test de normalité de Kolmogorov-Smirnov sur les trois échantillons simulés :

```
> library(nortest)
> lillie.test(Ech1)
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: Ech1
D = 0.041892, p-value = 0.532
```

```
> lillie.test(Ech2)
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: Ech2
D = 0.083624, p-value = 0.001691
```

```
> lillie.test(Ech3)
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: Ech3
D = 0.34735, p-value < 2.2e-16
```

1 Rappels

- Fonction de répartition et quantiles
- Fonction de répartition empirique

2 Test de Kolmogorov de comparaison ou d'adéquation

3 Tests de comparaison de deux échantillons

- Tests de Kolmogorov-Smirnov
- Test de Wilcoxon- Mann-Whitney
- Test de la médiane

4 Tests de normalité

- Méthode graphique : droite de Henry
- Test de normalité de Kolmogorov-Smirnov (de Lilliefors)
- Test de Shapiro-Wilk

Test de Shapiro-Wilk

- Il s'agit d'un test basé sur les L -statistiques (combinaison linéaire des statistiques d'ordre), qui se base sur une comparaison de la variance empirique avec un estimateur de la variance des X_i qui a de bonnes propriétés sous l'hypothèse de normalité.
- Il se base sur une mesure de corrélation (au carré) entre les quantiles empiriques et les quantiles théoriques d'une loi normale centrée réduite.

Estimation de la moyenne et de la variance à l'aide des statistiques d'ordre pour des lois symétriques

- X_1, \dots, X_n i.i.d, $\mu = \mathbb{E}[X_i]$ et $\sigma^2 = \text{Var}(X_i)$
- La loi de $Y_i = (X_i - \mu)/\sigma$ est supposée symétrique
- $(Y_{(1)}, \dots, Y_{(n)})$ l'échantillon des Y_i ordonné : $Y_{(i)} = (X_{(i)} - \mu)/\sigma$
- On cherche "le meilleur estimateur linéaire" de μ et σ

Estimation de la moyenne et de la variance à l'aide des statistiques d'ordre pour des lois symétriques

- Pour $i, j \in \{1, \dots, n\}$, on pose $\alpha_j = \mathbb{E}[Y_{(j)}]$ et $B_{i,j} = \text{Cov}(Y_{(i)}, Y_{(j)})$
- Soit

$$X_{(i)} = \mu + \alpha_i \sigma + \varepsilon_i \text{ avec } \mathbb{E}[\varepsilon_i] = 0, \text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 B_{i,j}$$

- Ecriture matricielle :

$$\underbrace{\begin{pmatrix} X_{(1)} \\ \vdots \\ X_{(n)} \end{pmatrix}}_{X_{(\cdot)}} = \underbrace{\begin{pmatrix} 1 & \alpha_1 \\ 1 & \alpha_2 \\ \vdots & \vdots \\ 1 & \alpha_n \end{pmatrix}}_A \underbrace{\begin{pmatrix} \mu \\ \sigma \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon} \text{ avec } \varepsilon \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2 B)$$

Estimation de la moyenne et de la variance à l'aide des statistiques d'ordre pour des lois symétriques

- Ecriture matricielle :

$$X_{(\cdot)} = A\theta + \varepsilon \text{ avec } \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 B)$$

donc

$$B^{-1/2}X_{(\cdot)} = B^{-1/2}A\theta + B^{-1/2}\varepsilon \text{ avec } B^{-1/2}\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$$

- L'estimateur des moindres carrés pondérés de (μ, σ) :

$$\begin{pmatrix} \hat{\mu}_n \\ \hat{\sigma}_n \end{pmatrix} = (A'B^{-1}A)^{-1}A'B^{-1}X_{(\cdot)}$$

avec

$$A'B^{-1}A = \begin{pmatrix} 1'_n B^{-1} 1_n & 1'_n B^{-1} \alpha \\ \alpha' B^{-1} 1_n & \alpha' B^{-1} \alpha \end{pmatrix}.$$

Estimation de la moyenne et de la variance à l'aide des statistiques d'ordre pour des lois symétriques

- Lorsque la loi des Y_i est symétrique, $1'_n B^{-1} \alpha = 0$, la matrice $A' B^{-1} A$ est donc diagonale.
- Il en résulte que

$$\hat{\mu}_n = \frac{1'_n B^{-1} X_{(.)}}{1'_n B^{-1} 1_n}, \quad \hat{\sigma}_n = \frac{\alpha' B^{-1} X_{(.)}}{\alpha' B^{-1} \alpha}.$$

- Idée du test de Shapiro-Wilk : comparer l'estimateur $\hat{\sigma}_n^2$ (bon estimateur de la variance quand loi normale donc symétrique) avec l'estimateur usuel $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$.

Test de Shapiro-Wilk

- Y_1, \dots, Y_n i.i.d. de loi $\mathcal{N}(0, 1)$
- $\alpha = (\mathbb{E}[Y_{(1)}], \dots, \mathbb{E}[Y_{(n)}])'$ et B la matrice de covariance de $(Y_{(1)}, \dots, Y_{(n)})$.
- Statistique de test :

$$SW_n = \frac{(\alpha' B^{-1} \alpha)^2}{(\alpha' B^{-2} \alpha)} \frac{\hat{\sigma}_n^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = \frac{(\sum_{i=1}^n a_i X_{(i)})^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2},$$

avec $(a_1, \dots, a_n) = \frac{\alpha' B^{-1}}{(\alpha' B^{-1} B^{-1} \alpha)^{1/2}}$.

- La région de rejet est de la forme $\{SW_n \leq c_{n,\alpha}\}$.

Exemples

```
> shapiro.test(Ech1)
```

```
Shapiro-Wilk normality test
```

```
data: Ech1
```

```
W = 0.99268, p-value = 0.4201
```

```
> shapiro.test(Ech2)
```

```
Shapiro-Wilk normality test
```

```
data: Ech2
```

```
W = 0.95826, p-value = 1.289e-05
```

```
> shapiro.test(Ech3)
```

```
Shapiro-Wilk normality test
```

```
data: Ech3
```

```
W = 0.45775, p-value < 2.2e-16
```