Tests non paramétriques

Retour au plan du cours

1 Fonction de répartition empirique

1.1 Définition et premières propriétés

Soit $X_1, X_2, \dots X_n, \dots$ une suite de variables aléatoires réelles i.i.d. de fonction de répartition F. On rappelle que, pour tout $x \in \mathbb{R}$,

$$F(x) = \mathbb{P}(X_i \leq x).$$

DÉFINITION 1. — *On appelle fonction de répartition empirique associée au n échantillon $X_1, X_2, \dots X_n$ la fonction*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq x}.$$

- F_n est croissante, continue à droite, $\lim_{x \to -\infty} F_n(x) = 0$, $\lim_{x \to +\infty} F_n(x) = 1$.
- $nF_n(x)$ suit une loi binomiale de paramètre $(n, F(x))$.
- $E(F_n(x)) = F(x)$, pour tout x, $F_n(x)$ est un estimateur sans biais de $F(x)$.
- $\text{Var}(nF_n(x)) = nF(x)(1 - F(x))$,

$$\text{Var}(F_n(x)) = \frac{F(x)(1 - F(x))}{n} \to 0.$$

- Par l'inégalité de Tchebichev,

$$\forall \varepsilon > 0,\ \mathbb{P}\left(|F_n(x) - F(x)| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} \text{Var}(F_n(x)) \to 0.$$

$$|F_n(x) - F(x)| \xrightarrow{\text{Prob}} 0$$

C'est aussi une conséquence de la loi des grands nombres.

- On déduit du TLC que pour tout x tel que $F(x)(1 - F(x)) \neq 0$,

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{\text{Loi}} \mathcal{N}(0, F(x)(1 - F(x))).$$

1.2 Inverse généralisée

DÉFINITION 2. — *Soit F une fonction de répartition. On définit l'inverse généralisée F^{-1} de F par*

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

Remarque. — Si F est une bijection, F^{-1} est la bijection réciproque.

Exemple. — Calculer F^{-1} pour la loi de Bernoulli de paramètre p.

LEMME 3. — *Soit U une variable aléatoire de loi uniforme sur $[0, 1]$. Soit F une fonction de répartition et F^{-1} son inverse généralisée. La variable aléatoire $X = F^{-1}(U)$ a pour fonction de répartition F.*

Ceci permet de simuler des v.a. de loi donnée, dès lors que l'on sait calculer F^{-1}.
Exercice : comment simuler une variable de loi exponentielle de paramètre λ ? De loi de Bernoulli de paramètre p ?

Démonstration. — La preuve repose sur l'équivalence suivante :

$$F^{-1}(x) \leq y \Leftrightarrow x \leq F(y). \tag{1}$$

On déduit de (1) que, pour tout x,

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x),$$

donc $X = F^{-1}(U)$ a pour fonction de répartition F. ∎

LEMME 4. — *Soit X une variable aléatoire de fonction de répartition F, supposée continue. Alors $F(X)$ suit une loi uniforme sur $[0, 1]$.*

2 Tests basés sur la fonction de répartition empirique

2.1 Test d'adéquation de Kolmogorov

Soit $X_1, \ldots, X_n$ i.i.d. de fonction de répartition F. On se donne une fonction de répartition F_0, supposée continue. On veut tester l'hypothèse H_0 : $F = F_0$ contre $H_1 : F \neq F_0$.

DÉFINITION 5. — *Le test de Kolmogorov est défini par la statistique de test*

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|.$$

Il consiste à rejeter l'hypothèse H_0 si $D_n \geq d_{n,\alpha}$.

PROPOSITION 6. — *La loi de D_n sous l'hypothèse H_0 ($F = F_0$) est indépendante de F_0.*

Remarque. — Soit $X_{(1)} \leq \ldots \leq X_{(n)}$ l'échantillon ordonné. On pose $X_{(0)} = -\infty$ et $X_{(n+1)} = +\infty$.

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

ce qui permet de calculer facilement D_n.

La loi de D_n sous H_0 est tabulée. On trouve dans les tables les quantiles $d_{n,1-\alpha}$ tels que

$$\mathbb{P}_{H_0}(D_n \geq d_{n,1-\alpha}) \leq \alpha,$$

(en étant le plus proche possible de α). Ces tables sont obtenues à partir de simulations de D_n, sous l'hypothèse que les X_i sont i.i.d. de loi uniforme sur $[0,1]$ ($F_0 = \mathbf{1}_{[0,1]}$). Si la loi de D_n dépendait de F_0, il faudrait construire une table pour chaque loi F_0.

Pour faire un test unilatéral, $H_0 : F = F_0$ contre $H_1 : F \geq F_0$ (respectivement $F \leq F_0$), on utilise la statistique de test

$$D_n^+ = \sup_{x \in \mathbb{R}} (F_n(x) - F_0(x))$$

respectivement

$$D_n^- = \sup_{x \in \mathbb{R}} (F_0(x) - F_n(x))$$

On rejette H_0 si $D_n^+ \geq d_{n,1-\alpha}^+$, respectivement $D_n^- \geq d_{n,1-\alpha}^-$. Les quantiles sont lus dans les tables.

PROPOSITION 7. — *(admise)*

$$\forall \lambda > 0,\ \mathbb{P}_{H_0}(\sqrt{n} D_n^+ \geq \lambda) \to \exp(-2\lambda^2) \textit{ Smirnov (1942)}$$

$$\forall \lambda > 0, \mathbb{P}_{H_0}(\sqrt{n} D_n \geq \lambda) \to 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2\lambda^2) \textit{ Kolmogorov (1933)}$$

$$\forall \lambda > 0,\ \mathbb{P}_{H_0}(\sqrt{n} D_n \geq \lambda) \leq 2 \exp(-2\lambda^2) \textit{ Massart (1990)}$$

Il existe d'autres tests basés sur la fonction de répartition empirique. Le test de Cramer Von Mises utilise la statistique

$$C_n = n \int_{-\infty}^{+\infty} (F_n(x) - F_0(x))^2 f_0(x) dx,$$

le test d'Anderson Darling utilise la statistique de test

$$D_n = n \int_{-\infty}^{+\infty} (F_n(x) - F_0(x))^2 \frac{f_0(x)}{F_0(x)(1 - F_0(x))} dx.$$

Comme pour le test de Kolmogorov, on montre que les lois de C_n et A_n sont indépendantes de F_0 sous H_0. Ces lois sont tabulées.

3 Tests de comparaison de deux échantillons

On considère deux échantillons indépendants $X_1, \ldots, X_n$ i.i.d. de fonction de répartition F_0 et $Y_1, \ldots, Y_m$ i.i.d. de fonction de répartition F_1. Dans le cas où F_0 correspond à une loi normale $\mathcal{N}(m_0, \sigma^2)$ et F_1 à la loi $\mathcal{N}(m_1, \sigma^2)$, on peut utiliser un test de Student pour tester $H_0 : F_0 = F_1$ contre $H_1 : F_0 \neq F_1$. (cf cours de 3ième année). Nous ne revenons pas sur ce test et nous nous plaçons ici dans un cadre non paramétrique. Les lois des variables X_i et Y_j ne sont pas supposées connues.

3.1 Tests de Kolmogorov-Smirnov de comparaison de deux échantillons

On considère deux échantillons indépendants : $X_1, \ldots, X_n$ i.i.d. de fonction de répartition F_0 et $Y_1, \ldots, Y_m$ i.i.d. de fonction de répartition F_1. On veut tester $H_0 : F_0 = F_1$ contre $H_1 : F_0 \neq F_1$.
Soit F_n la fonction de répartition empirique de l'échantillon $(X_1, \ldots, X_n)$ et G_m celle de l'échantillon $(Y_1, \ldots, Y_m)$.

DÉFINITION 8. — *Le test de Kolmogorov-Smirnov est défini par la statistique de test*

$$D_{n,m} = \sup_{x \in \mathbb{R}} |F_n(x) - G_m(x)|.$$

Il consiste à rejeter l'hypothèse H_0 si $D_{n,m} \geq d_{n,m,1-\alpha}$.

PROPOSITION 9. — *Si F_0 est continue, la loi de $D_{n,m}$ sous l'hypothèse $F_0 = F_1$ est indépendante de F_0. Cette loi est tabulée.*

Remarque. — Pour faire un test unilatéral ($H_0 : F_0 = F_1$ contre $H_1 : F_0 \geq F_1$) , on utilise la statistique de test

$$D^+_{n,m} = \sup_{x \in \mathbb{R}} (F_n(x) - G_m(x)).$$

3.2 Test de Wilcoxon- Mann-Whitney

On considère deux échantillons indépendants : $X_1, \ldots, X_n$ i.i.d. de fonction de répartition F_0 et $Y_1, \ldots, Y_m$ i.i.d. de fonction de répartition F_1. On veut tester $H_0 : F_0 = F_1$ contre $H_1 : F_0 \geq F_1$. on suppose que F_0 et F_1 sont continues.
Le principe du test consiste à détermniner le nombre de couples (X_i, Y_j) pour lesquels $Y_j \geq X_i$. Sous H_1, pour tout x, $\mathbb{P}(Y \leq x) \leq P(X \leq x)$ (avec parfois l'inégalité stricte), par conséquent pour tout x, $\mathbb{P}(Y > x) \geq P(X > x)$ et le nombre de couples (X_i, Y_j) pour lesquels $Y_j \geq X_i$ prend des valeurs plus grandes sous H_1 que sous H_0.

DÉFINITION 10. — *On appelle test de Mann-Whitney le test défini à partir de la statistique*

$$U_{(n,m)} = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{1}_{Y_j > X_i}.$$

Le test consiste à rejeter H_0 si $U_{(n,m)} \geq u_{(n,m),1-\alpha}$.

Remarque. — La loi de $U_{(n,m)}$ sous H_0 peut être établie par récurrence (cf Caperaa Van Cutsem p 126). On note

$$p_{n,m}(k) = \mathbb{P}_{H_0}(U_{(n,m)} = k) \text{ pour } k = 0, 1, \ldots mn$$

$$p_{n,0}(k) = p_{0,m}(k) = 1 \text{ pour } k = 0; = 0 \text{ pour } k \neq 0.$$

Alors pour tout k,

$$(n+m)p_{n,m}(k) = mp_{n-1,m}(k) + np_{n,m-1}(k-1).$$

Cette formule de récurrence permet de calculer la loi de $U_{(n,m)}$ sous H_0.

On peut aussi utiliser un résultat asymptotique :

THÉORÈME 11. — *(Hajek (1968)) (admis)*
Sous H_0,

$$\frac{U_{(n,m)} - \mathbb{E}_{H_0}(U_{(n,m)})}{\sqrt{Var_{H_0}(U_{(n,m)})}} \xrightarrow{\text{Loi}} \mathcal{N}(0,1) \text{ quand } n \to \infty, n/(n+m) \to \lambda \in]0,1[.$$

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

$$\mathbb{E}_{H_0}(U_{(n,m)}) = \frac{mn}{2},$$

$$\text{Var}_{H_0}(U_{(n,m)}) = mn\left(\frac{n+m+1}{12}\right).$$

Il existe une autre forme équivalement de ce test, appelé test de la somme des rangs de Wilcoxon, qui consiste à calculer la somme des rangs des individus du deuxième échantillon :

$$W_{n,m} = \sum_{j=1}^{m} R_j$$

où R_j représente le rang de Y_j dans l'échantillon complet ordonné : on note $(Z_1, \ldots, Z_n, Z_{n+1}, \ldots, Z_N) = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$. On pose pour tout j de 1 à m

$$R_j = \sum_{i=1}^{N} \mathbf{1}_{Z_i < Y_j} + 1.$$

On a la relation

$$U_{(n,m)} = W_{n,m} - \frac{m(m+1)}{2}.$$

Les deux statistiques conduisent donc au même test.

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. Dans ce cas, la solution la plus couramment employée dans les logiciels est la méthode des rangs moyens. Elle consiste à affecter à tous les élements d'un groupe d'ex-aequos la moyenne des rangs des élements du groupe.

3.3 Test de la médiane

On considère deux échantillons indépendants : $X_1, \ldots, X_n$ i.i.d. de fonction de répartition F_0 et $Y_1, \ldots, Y_m$ i.i.d. de fonction de répartition F_1. On veut tester $H_0 : F_0 = F_1$ contre $H_1 : F_0 \geq F_1$. on suppose que F_0 et F_1 sont continues.

Le principe du test consiste à détermniner le nombre de variables du deuxième échantillon qui sont supérieures à la médiane de l'ensemble des observations. On note $N = n + m$.

DÉFINITION 12. — *Le test de la médiane est défini à partir de la statistique*

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

Pour tester $H_0 : F_0 = F_1$ contre $H_1 : F_0 \geq F_1$, on rejette H_0 si $M_{n,m} \geq m_{n,m,1-\alpha}$.

Exemple d'application : Test de localisation. $X_1, \ldots, X_n$ sont i.i.d. de fonction de répartition F_0 et $Y_1, \ldots, Y_m$ sont i.i.d. de fonction de répartition $F_1 = F_0(. - \mu)$. Par exemple, on étudie la pression artérielle de patients soumis à un traitement contre l'hypertension (Y_j), et on les compare à des patients non traités (X_i). Supposons qu'après traitement, la loi de la pression artérielle est translatée de μ. Le traitement est efficace si $\mu < 0$, il est inefficace si $\mu = 0$.

Loi de $M_{n,m}$ sous H_0 : Supposons N pair.

$$\forall k \in \{0, \ldots, m\}, \mathbb{P}(mM_{n,m} = k) = \frac{C_m^k C_n^{N/2-k}}{C_N^{N/2}}.$$

Il s'agit d'une loi hypergéométrique de paramètre $(N, N/2, m)$. La connaissance de la loi de $M_{n,m}$ sous H_0 permet de déterminer la zone de rejet du test.

$$* \mathbb{E}_{H_0}(M_{(n,m)}) = \frac{1}{2} \text{ si } N \text{ pair} \quad (2)$$

$$= \frac{N-1}{2N} \text{ si } N \text{ impair} \quad (3)$$

*

$$* \text{Var}_{H_0}(M_{(n,m)}) = \frac{n}{4m(N-1)} \text{ si } N \text{ pair} \quad (4)$$

$$= \frac{n(N+1)}{4mN^2} \text{ si } N \text{ impair} \quad (5)$$

* Pour $n, m \geq 30$, on peut approximer la loi de $M_{(n,m)}$ sous H_0 par la loi $\mathcal{N}(\mathbb{E}_{H_0}(M_{(n,m)}), \text{Var}_{H_0}(M_{(n,m)}))$.

Remarque. — Les tests de Wilcoxon, Mann-Withney et de la médiane ne permettent pas de tester des alternatives bilatérales.

4 Tests de normalité

4.1 Méthode graphique : droite de Henry

La méthode est aussi appelée "Normal Probability Plot" ou "Q-Q Plot". On représente le graphe des points $(X_{(i)}, F^{-1} \circ F_n(X_{(i)}))$, où $X_{(1)} \leq \ldots \leq X_{(n)}$ est l'échantillon ordonné, F_n la fonction de répartition empirique de l'échantillon $(X_1, \ldots, X_n$ (notons que $F_n(X_{(i)}) = i/n$) et F représente la fonction de répartition de la loi $\mathcal{N}(0, 1)$. Sous l'hypothèse que les X_i sont i.i.d. de loi normale, les points $(X_{(i)}, F^{-1} \circ F_n(X_{(i)}))$ sont pratiquement alignés.

4.2 Test de normalité de Kolmogorov

Soit $X_1, \ldots, X_n$ i.i.d. de fonction de répartition F. On souhaite tester l'hypothèse H_0 : "les X_i suivent une loi normale", contre l'hypothèse H_1 : "les X_i ne suivent pas une loi normale". On note

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \; S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Le test de normalité de Kolmogorov utilise la statistique de test

$$T_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_{(\bar{X}, S^2)}(x)|$$

où $F_{(\bar{X}, S^2)}$ est la fonction de répartition de la loi normale $\mathcal{N}(\bar{X}, S^2)$. Le test consiste à rejeter l'hypothèse de normalité pour de grandes valeurs de T_n.

PROPOSITION 13. — *Sous l'hypothèse H_0, (les X_i suivent une loi normale $\mathcal{N}(m, \sigma^2)$), la loi de T_n ne dépend pas de (m, σ^2).*

La loi de T_n est tabulée (on peut par exemple la simuler avec $m = 0$ et $\sigma = 1$ pour en estimer les quantiles).

4.3 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.

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

Soit $X_1, \ldots, X_n$ i.i.d.. On note $\mu = \mathbb{E}(X_i)$ et $\sigma^2 = \text{Var}(X_i)$. La loi de $Y_i = (X_i - \mu)/\sigma$ est supposée symétrique (ce qui signifie que $-Y_i$ a même loi que Y_i). On note $(X_{(1)}, \ldots, X_{(n)})$ l'échantillon des X_i ordonné : $X_{(1)} \leq \ldots \leq X_{(n)}$. On note $(Y_{(1)}, \ldots, Y_{(n)})$ l'échantillon des Y_i ordonné. On a

$$Y_{(i)} = (X_{(i)} - \mu)/\sigma.$$

Pour $i = 1, \ldots, n$, on note

$$\alpha_i = \mathbb{E}(Y_{(i)}), \quad B_{i,j} = \text{Cov}(Y_{(i)}, Y_{(j)}).$$

On a alors

$$X_{(i)} = \mu + \alpha_i \sigma + \varepsilon_i,$$

avec $\mathbb{E}(\varepsilon_i) = 0$. Les ε_i ne sont pas indépendantes. La matrice de variance-covariance du vecteur $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ est $\sigma^2 B$. On note $\mathbf{1}$ et α les vecteurs de $\mathbb{R}^n$ definis par

$$\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_n \end{pmatrix}.$$

On note A la matrice de taille $(n, 2)$ définie par $A = (\mathbf{1}, \alpha)$. Enfin, on note $X_{(.)} = (X_{(1)}, \ldots, X_{(n)})'$ et $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$. On a la relation

$$X_{(.)} = A \begin{pmatrix} \mu \\ \sigma \end{pmatrix} + \varepsilon.$$

L'estimateur des moindres carrés pondérés de (μ, σ) est obtenu en minimisant en les paramètres (μ, σ) le critère :

$$\left(X_{(.)} - A\begin{pmatrix}\mu\\ \sigma\end{pmatrix}\right)' B^{-1}\left(X_{(.)} - A\begin{pmatrix}\mu\\ \sigma\end{pmatrix}\right).$$

On obtient comme solution de ce système

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

(cf Cours sur le modèle linéaire)

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

LEMME 14. — *Lorque la loi des Y_i est symétrique, $\mathbf{1}'B^{-1}\alpha = 0$, la matrice $A'B^{-1}A$ est donc diagonale.*

Il en résulte que

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

On peut montrer que, si la loi des Y_i n'est pas symétrique, $\hat{\sigma}_n$ sous-estime σ.

4.3.2 Test de Shapiro-Wilk

DÉFINITION 15. — *Soit $Y_1, \ldots Y_n$ i.i.d. de loi $\mathcal{N}(0,1)$ et $Y_{(1)} \leq \ldots \leq Y_{(n)}$ l'échantillon ordonné.*
Soit $\alpha = (\mathbb{E}(Y_{(1)}), \ldots \mathbb{E}(Y_{(n)}))'$. Soit B la matrice de covariance du vecteur $(Y_{(1)}, \ldots, Y_{(n)})$.
Le test de Shapiro-Wilk pour tester l'hypothèse de normalité des X_i est basé sur la statistique de test :

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

On peut l'écrire sous la forme

$$SW_n = \frac{\left(\sum_{i=1}^n a_i X_{(i)}\right)^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2},$$

avec

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

La zone de rejet est de la forme $(SW_n \leq c_{n,1-\alpha})$.

Les a_i sont tabulés, ce qui permet de calculer facilement SW_n, les quantiles $(c_{n,1-\alpha})$ sont également tabulés.