

Tests non paramétriques

adapté du cours de S. Ducay

F. Wlazinski

Licence d'économie

Dans une population P , on étudie un caractère quantitatif représenté par une variable aléatoire X de moyenne μ et d'écart-type σ . De cette population, on extrait un échantillon $E = (x_1, x_2, \dots, x_n)$ de taille n de X .

Nos estimateurs usuels sont la moyenne d'échantillon et la variance corrigée d'échantillon données respectivement par $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ et $s_c^2 = \frac{n}{n-1} s^2$ avec $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \overline{x^2} - (\bar{x})^2$.

Dans les précédents chapitres, nous avons présenté des formules d'intervalles de confiance et des procédures de tests statistiques dont la mise en œuvre nécessitait souvent de supposer que la variable aléatoire X suivait une loi normale $\mathcal{N}(\mu; \sigma)$.

Cette hypothèse de normalité n'est pas gratuite et doit être vérifiée par un test statistique permettant de la valider ou de la réfuter.

Un test d'adéquation permet de statuer sur la compatibilité d'une distribution observée avec une distribution théorique associée à une loi de probabilité. Il s'agit de modélisation. Nous avons déjà présenté le test de χ^2 d'adéquation à une loi théorique et, en particulier, il peut s'appliquer à la loi normale. Mais il n'est pas toujours adapté, en particulier au cas de petits échantillons.

1 Le test de Shapiro-Wilk

Principe

Etant donné un échantillon $E = (x_1, x_2, \dots, x_n)$, on cherche à déterminer si E peut être issu d'une population gaussienne.

Le test de Shapiro-Wilk est très populaire et est particulièrement puissant pour les petits effectifs ($n \leq 50$).

Méthode

On teste H_0 contre H_1 où H_0 est "X suit une loi normale" et $H_1 = \overline{H_0}$ est "X ne suit pas une loi normale".

A partir des observations $(x_1, x_2, \dots, x_n)$, la mise en œuvre du test se fait de la manière suivante :

- i. On trie les données x_i dans l'ordre croissant. On obtient une nouvelle série que l'on note $x_{(i)}$.
- ii. On calcule la moyenne $\bar{x}$ et la variance corrigée s_c^2 de la série (idem x_i ou $x_{(i)}$).
- iii. On calcule les écarts $d_i = (x_{(n+1-i)} - x_{(i)})$ pour i allant de 1 à la moitié stricte de l'effectif, autrement dit jusqu'à $\left\lfloor \frac{n}{2} \right\rfloor$ qui est la partie entière inférieure du rapport $\frac{n}{2}$.
- iv. On récupère les valeurs des coefficients a_i dans la table de Shapiro-Wilk pour l'effectif n (il y a $\left\lfloor \frac{n}{2} \right\rfloor$ valeurs).

v. On calcule $t^2 = \frac{1}{n-1} \left(\sum_{i=1}^{\left[\frac{n}{2}\right]} a_i (x_{(n+1-i)} - x_{(i)}) \right)^2$.

vi. On calcule $w = \frac{t^2}{s_c^2}$.

vii. Pour un risque α , on lit le **seuil** critique $w_{\min}$ dans la table de Shapiro-Wilk pour l'effectif n .

On décide que :

- si $w \geq w_{\min}$, on ne peut pas rejeter H_0 au risque α c'est-à-dire on ne peut pas rejeter la normalité.
- si $w < w_{\min}$, on rejette H_0 au risque α c'est-à-dire on rejette la normalité.

Exemple 1.1

On a demandé à dix étudiants de l'U.F.R. d'économie-gestion d'Amiens, le temps hebdomadaire qu'ils consacraient au travail personnel. On a obtenu les résultats suivants :

Etudiant	A	B	C	D	E	F	G	H	I	J
x_i	3	4	2	5	5	7	4	6	3	4

Où x_i est le temps de travail personnel hebdomadaire (en heures).

On souhaite savoir si cet échantillon est issu d'une population gaussienne.

2 Les tests non paramétriques

Lorsque l'on considère des échantillons de tailles faibles et que l'hypothèse de normalité ne peut être retenue, les tests paramétriques, appuyés sur les estimateurs usuels et décrits dans les précédents chapitres, ne peuvent être mis en œuvre. Une alternative est d'utiliser des tests non paramétriques.

2.1 Comparaison de deux moyennes

Dans deux populations P_1 et P_2 , on étudie un même caractère, représenté par des variables aléatoires X_1 et X_2 , de moyennes respectives μ_1 et μ_2 , d'écart-types respectifs σ_1 et σ_2 . Des populations P_1 et P_2 , on dispose d'un échantillon $E_1 = (x_{1,1}, x_{1,2}, \dots, x_{1,n_1})$ de taille n_1 de X_1 et un échantillon $E_2 = (x_{2,1}, x_{2,2}, \dots, x_{2,n_2})$ de taille n_2 de X_2 .

Les moyennes des échantillons sont alors respectivement $\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1,i}$ et $\bar{x}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2,i}$.

Les variances corrigées des échantillons sont respectivement $s_{c,1}^2 = \frac{n_1}{n_1-1} s_1^2$ et $s_{c,2}^2 = \frac{n_2}{n_2-1} s_2^2$ avec $s_1^2 = \overline{x_1^2} - (\bar{x}_1)^2$ et $s_2^2 = \overline{x_2^2} - (\bar{x}_2)^2$.

2.1.1 Cas de petits échantillons indépendants : test de Mann et Whitney

Méthode

On teste H_0 contre H_1 où H_0 est "les moyennes sont égales", i.e., $\mu_1 = \mu_2$ et $H_1 = \overline{H_0} : \mu_1 \neq \mu_2$.

On suppose que $n_1 \leq 30$ ou $n_2 \leq 30$, que les échantillons E_1 et E_2 sont indépendants et que les lois de X_1 et de X_2 sont inconnues.

On classe par ordre croissant l'ensemble des valeurs observées $x_{1,1}, x_{1,2}, \dots, x_{1,n_1}$ et $x_{2,1}, x_{2,2}, \dots, x_{2,n_2}$ des deux échantillons.

On affecte à chaque valeur son rang dans ce classement. S'il y a des ex-æquo, on attribue à chacun d'eux un rang égal à la moyenne des rangs qu'ils occupent.

Si r_1 et r_2 désigne la somme des rangs des valeurs de chacun des deux échantillons, alors on pose $u_1 = \frac{n_1(2n_2 + n_1 + 1)}{2} - r_1$ et $u_2 = \frac{n_2(2n_1 + n_2 + 1)}{2} - r_2$.

On peut vérifier que $r_1 + r_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$ et que $u_1 + u_2 = n_1 n_2$ (qui peut être une méthode de contrôle pour les calculs à la main).

Enfin, on définit u comme la plus petite des deux valeurs u_1 et u_2 .

Grâce aux tables de Mann et Whitney, on détermine le réel seuil $u_{\min}$.

On décide que :

- Si $u > u_{\min}$, on ne peut pas rejeter H_0 au risque α .
- Si $u \leq u_{\min}$, on rejette H_0 avec une probabilité α de se tromper.

Cas particulier : $n_1 > 20$ et $n_2 > 20$

Dans ce cas, on peut calculer $z = \frac{u - m}{s}$ où $m = \frac{n_1 n_2}{2}$ et $s = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$.

On détermine alors $z_{\max} = \varphi^{-1}\left(1 - \frac{\alpha}{2}\right)$ (table de la loi normale centrée réduite).

On décide que :

- si $|z| < z_{\max}$, alors on ne peut pas rejeter H_0 .
- si $|z| \geq z_{\max}$, alors on rejette H_0 avec une probabilité α de se tromper.

Exemple 2.1

Sur deux échantillons indépendants issus de deux populations, l'observation d'une variable quantitative a donné les résultats suivants :

Echantillon 1 : 3,0 9,8 2,0 5,2 3,6 5,9 8,5 9,4

Echantillon 2 : 9,3 12,5 11,3 7,6 3,2 8,6 7,2 14,2 9,6 3,8

Ne faisant aucune hypothèse sur la loi suivie par la variable étudiée, on veut comparer les moyennes des deux populations.

Remarque 2.2

Ce test est dit *non-paramétrique* car il ne nécessite pas d'estimation des paramètres (moyenne et variance). Il est également appelé un *test de rangs* car les valeurs observées sont remplacées par leur rang au sein des échantillons.

2.1.2 Cas de petits échantillons appariés : test de Wilcoxon

Les notations sont les mêmes que dans la partie 2.1.1.

Méthode

On teste H_0 contre H_1 où H_0 est "les moyennes sont égales", i.e., $\mu_1 = \mu_2$ et $H_1 = \overline{H_0} : \mu_1 \neq \mu_2$.

Mais, dans cette partie, on suppose que $n_1 = n_2 = n \leq 30$, que les échantillons E_1 et E_2 sont appariés et que les lois de X_1 et de X_2 sont inconnues.

On calcule les différences $d_i = x_{1,i} - x_{2,i}$ entre les valeurs observées des deux échantillons.

On supprime les différences nulles et on note N le nombre de différences non nulles.

On classe ces différences par ordre croissant des valeurs absolues. On affecte à chaque différence son rang dans ce classement. S'il y a des ex-æquo, on attribue à chacun d'eux un rang égal à la moyenne des rangs qu'ils occupent.

On calcule alors la somme w_{pos} des rangs des différences positives et la somme w_{neg} des rangs des différences négatives. On définit w comme la plus petite des deux valeurs w_{pos} et w_{neg} .

On peut vérifier que $w_{pos} + w_{neg} = \frac{N(N+1)}{2}$.

Grâce à la table de Wilcoxon, on trouve la valeur seuil $w_{\min}$.

On décide que :

- Si $w > w_{\min}$, on ne peut pas rejeter H_0 au risque α .
- Si $w \leq w_{\min}$, alors on rejette H_0 avec une probabilité α de se tromper.

Cas particulier : $N > 25$

Dans ce cas, on calcule $z = \frac{w - \mu}{\sigma}$ où $\mu = \frac{N(N+1)}{2}$ et $\sigma = \sqrt{\frac{N(N+1)(2N+1)}{24}}$.

On détermine $z_{\max} = \varphi^{-1}\left(1 - \frac{\alpha}{2}\right)$ (table de la loi normale centrée réduite).

On décide que :

- Si $|z| < z_{\max}$, alors on ne peut pas rejeter H_0 .
- Si $|z| \geq z_{\max}$, alors on rejette H_0 avec une probabilité α de se tromper.

Exemple 2.3

Dans le cadre d'un test gustatif sur un nouveau produit alimentaire, on a demandé à 10 personnes d'évaluer par une note sur 10, deux recettes différentes de ce produit. Les résultats sont les suivants :

Personne	a	b	c	d	e	f	g	h	i	j
Recette A	5	4	2	3	4	3	8	5	4	5
Recette B	6	3	3	1	1	3	4	2	5	7

On cherche à savoir si les deux recettes sont également appréciées au seuil de 5%.

Remarque 2.4

Remarque 2.5

Comme le test de Mann et Whitney, le test de Wilcoxon est un test de rangs non-paramétrique.

2.2 Comparaison de $k(\geq 2)$ moyennes : test de Kruskal et Wallis

Principe

Le test de Kruskal et Wallis est l'analogue non paramétrique de l'analyse de la variance à un facteur. On dispose de k échantillons E_i de taille n_i qui sont supposés indépendants. Mais on ne fait aucune hypothèse sur les lois des X_i .

Méthode

On teste H_0 contre H_1 où l'hypothèse nulle H_0 est "toutes les moyennes sont égales" c'est-à-dire $\mu_1 = \mu_2 = \dots = \mu_k$ et $H_1 = \overline{H_0}$ est "au moins deux de ces moyennes ne sont pas égales" c'est-à-dire il existe un couple (i, j) de $\llbracket 1; k \rrbracket^2$ tel que $\mu_i \neq \mu_j$.

On classe par ordre croissant l'ensemble des valeurs observées $x_{1,1}, x_{1,2}, \dots, x_{1,n_1}, x_{2,1}, x_{2,2}, \dots, x_{2,n_2}$ et $x_{k,1}, x_{k,2}, \dots, x_{k,n_k}$ des k échantillons. On affecte à chaque valeur son rang dans ce classement. S'il y a des ex-æquo, on attribue à chacun d'eux un rang égal à la moyenne des rangs qu'ils occupent.

A chaque échantillon E_i on associe la somme r_i des rangs des valeurs de cet échantillon.

On calcule $h = \frac{12}{n(n+1)} \left(\sum_{i=1}^k \frac{r_i^2}{n_i} \right) - 3(n+1)$ avec $n = \sum_{i=1}^k n_i$.

Si les n_i sont tous inférieurs ou égaux à 5, grâce à la table de Kruskal et Wallis avec α et les n_i on trouve la valeur plafond $h_{\max}$.

On décide que :

- Si $h < h_{\max}$, on ne peut pas rejeter H_0 au risque α .
- Si $h \leq h_{\max}$, alors on rejette H_0 avec une probabilité α de se tromper.

Cas particulier : Si certains des n_i sont strictement plus grands que 5, grâce à la table du χ^2 avec d.d.l. = $k - 1$ et α , on trouve la valeur plafond $h_{\max}$.

On décide que :

- Si $h < b_{\max}$, on ne peut pas rejeter H_0 au risque α .
- Si $h \leq b_{\max}$, alors on rejette H_0 avec une probabilité α de se tromper.

Exemple 2.6

On a dosé la teneur en calcium de trois types d'eaux issues d'origines géographiques différentes. Chaque type d'eau a fait l'objet de quatre prélèvements. Les résultats des dosages (en mg de calcium par litre d'eau) sont :

Eau 1 :	18	20	22	25
Eau 2 :	15	16	17	21
Eau 3 :	15	20	21	25

On souhaite savoir si l'origine géographique a une influence significative sur la teneur en calcium des eaux considérées.

2.3 Coefficient de corrélation de rang : test de Spearman.

Principe

Sur une population, on considère deux variables aléatoires X et Y . On veut tester l'absence de corrélation entre X et Y . Sans hypothèse sur les lois de X et Y , on ne peut pas utiliser les résultats précédents.

Méthode

On teste $H_0 : \rho_S = 0$ c'est-à-dire les variables X et Y ne sont pas corrélées contre $H_1 : \rho_S \neq 0$.

A partir des n couples d'observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. On remplace les x_i par leur rang x'_i dans un ordre croissant des valeurs. De même, on remplace les y_i par leur rang y'_i . En cas d'ex-æquo, le rang est la moyenne des rangs occupés par ces ex-æquos. Soient alors X' et Y' les variables aléatoires ainsi déterminées et (x', y') l'échantillon obtenu.

On appelle *coefficient de corrélation de rang de Spearman* le nombre ρ_S égal au coefficient de corrélation entre X' et Y' .

Une estimation de ρ_S est donnée par le coefficient de corrélation r_S calculé à partir des n couples (x'_i, y'_i) de l'échantillon (x', y') .

$$\text{Rappel : } \text{cov}(x'; y') = E(x'y') - E(x')E(y') \text{ et } r_S = \frac{\text{cov}(x'; y')}{\sqrt{V(x') \times V(y')}}.$$

$$6 \sum_{i=1}^n d_i^2$$

Remarque : S'il n'y a pas d'ex-æquo, on a $r_S = 1 - \frac{\sum_{i=1}^n d_i^2}{n(n^2 - 1)}$ où $d_i = x'_i - y'_i$.

Soit R_S la variable aléatoire qui prend la valeur r_S au cours de chaque échantillonnage.

On détermine la valeur plafond $r_{\max}$ grâce à la table de Spearman avec α et n .

On décide que :

- si $|r_S| < r_{\max}$, alors on ne peut pas rejeter H_0 .
- si $|r_S| \geq r_{\max}$, alors on rejette H_0 avec une probabilité α de se tromper.

Cas particulier : Si $n \geq 10$, on peut calculer $t = \frac{r_S \sqrt{n-2}}{\sqrt{1-r_S^2}}$.

On détermine alors la valeur plafond $t_{\max}$ grâce à la table bilatérale de Student avec α et $d.d.l. = n-2$.

On décide que :

- si $|t| < t_{\max}$, alors on ne peut pas rejeter H_0 .
- si $|t| \geq t_{\max}$, alors on rejette H_0 avec une probabilité α de se tromper, i.e. que l'on considère que X et Y sont liées, sans savoir de quelle façon.

Exemple 2.7

On se demande s'il existe une relation entre la longueur des oreilles et celle du corps d'un souriceau élevé dans le noir. L'expérience a donné les résultats suivants :

Corps (cm)	11,6	12,4	10,9	11,2	12,1	11,8	13,1	12,5
Oreilles (mm)	10,4	10,1	9,7	9,9	10,8	11,0	12,1	11,7

Peut-on considérer, au vu de ces données, que les oreilles sont d'autant plus grande que la souris est plus grande?