

Statistique de Rang

Jean-Claude Régnier

Méthodes quantitatives 4PAKQNT1

1. Le coefficient de corrélation des rangs R_S de Spearman

A chaque observation, nous associons son rang k dans l'ordre du recueil des observations et son rang R_k dans l'ordre de l'échantillon.

Rangs des observations	1	2	...	k	...	n
Rangs dans l'échantillon ordonné	R_1	R_2	...	R_k	...	R_n

La statistique est

$$S_n = R_S = 1 - \frac{\sum_{k=1}^{k=n} [R_k - k]^2}{n(n^2 - 1)}$$

l'espérance de R_S est $E(R_S) = 0$

la variance de R_S est $V(R_S) = \frac{1}{n-1}$

Pour tester la significativité de la valeur obtenue, on prend en référence la situation théorique d'indépendance des deux rangements dans la population, c'est à dire que les $n!$ classements sont équiprobables. Dans ce cas de l'indépendance, la valeur est $\rho_S = 0$. Cependant il convient de rappeler que la réciproque est fautive.

Dans le cas d'une tendance monotone croissante parfaite, les classements sont identiques

$$R_k = k$$

$$\rho_S = 1 : \text{les deux classements sont identiques}$$

Dans le cas d'une tendance monotone décroissante parfaite, les classements sont inversés $R_k = n+1-k$

$$\rho_S = -1 : \text{les deux classements sont inverses}$$

Il s'agit alors de prendre une décision sur la base d'une valeur r_S , réalisation de la variable R_S sur un échantillon. A l'aide de la table du coefficient de corrélation de Spearman ou d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss, on peut fonder le rejet ou le non-rejet au seuil α de l'hypothèse nulle H_0 postulant l'indépendance des deux rangements. La région critique W , région de rejet de H_0 au profit de H_1 , est définie selon H_1 , l'une des trois hypothèses alternatives usuelles.

H1 tendance monotone croissante dépendance positive	H1 tendance monotone dépendance quelconque	H1 tendance monotone décroissante dépendance négative
$W = \{ r_s, r_s > c \}$	$W = \{ r_s, r_s > c \}$	$W = \{ r_s, r_s < -c \}$

La valeur critique c ($c \geq 0$) s'obtient en résolvant l'équation $\text{Prob}(W|H_0) = \alpha$

Asymptotiquement la variable $Z_s = R_s \sqrt{n-1}$ suit la loi de la variable de Laplace-Gauss LG(0,1). Cette approximation est jugée acceptable pour $n > 30$.

Asymptotiquement la variable $T_s = R_s \sqrt{\frac{n-2}{1-R_s^2}}$ suit la loi de la variable de Student à $n-2$ ddl. Cette approximation est jugée acceptable pour $n > 10$.

2. Test de concordance de p rangements de n objets de M.G. Kendall

Objets Critères	1	2	...	i	...	n	totaux
1	R_{11}	R_{21}		R_{i1}		R_{n1}	$R_{.1}$
2	R_{12}	R_{22}		R_{i2}		R_{n2}	$R_{.2}$
...							
j	R_{1j}	R_{2j}		R_{ij}		R_{nj}	$R_{.j}$
...							
p	R_{1p}	R_{2p}		R_{ip}		R_{np}	$R_{.p}$
Totaux	$R_{1.}$	$R_{2.}$		$R_{i.}$		$R_{n.}$	$r_{..}$

2.1. Méthode de calcul:

Chaque ligne est une permutation des nombres entiers de 1 à n dont la somme est constante et vaut $\frac{n(n+1)}{2}$. Ainsi $r_{..} = p \frac{n(n+1)}{2}$. Dans l'hypothèse d'une concordance parfaite, les totaux des colonnes seraient égaux respectivement à $p, 2p, 3p, \dots, np$ à une permutation près. On utilise alors la statistique $S_K = \sum_{i=1}^{i=n} (R_{i.} - \frac{r_{..}}{n})^2$ dont la valeur maximale

$$\text{est } S_{\max} = \frac{p^2(n^3-n)}{12}$$

Le coefficient de concordance de Kendall est:

$$W = \frac{S_K}{S_{\max}} = \frac{S_K}{\frac{1}{12} p^2 (n^3 - n)}$$

On peut aussi $W = \frac{1}{p} + \frac{2}{p^2} \sum_{i=1}^{i=p-1} \sum_{j=i+1}^j=p R_{ij}$ D'où

- l'espérance de W : $E(W) = \frac{1}{p}$

- la variance de W : $V(W) = \frac{2(p-1)}{p^3(n-1)}$

$w = 0$ correspond au cas où chaque colonne a même total. De faibles valeurs de W suggèrent l'indépendance des classements.

Pour tester l'hypothèse nulle H_0 d'indépendance des p rangements, on procède selon les procédures suivantes:

- Pour $n \leq 7$ on utilise une table,

- Pour $n \leq 7$ et $2 < p \leq 20$ la variable $\frac{(p-1)W}{1-W}$ est distribuée comme la variable de Fisher-Snédecor $F(n_1 = n-1-\frac{2}{p}; n_2 = (p-1)(n-1-\frac{2}{p}))$

- Pour $n > 7$ on utilise la variable $p(n-1)W$ qui est une variable de Pearson χ^2_{n-1}

Dans le cas où l'on est conduit à rejeter l'hypothèse d'indépendance des classements, on utilise souvent la règle de classement suivante:

les objets sont classés dans l'ordre défini par les totaux des colonnes.

Lorsqu'il y a des ex æquo on remplace le rang de ceux-ci par la moyenne arithmétique des rangs qu'ils auraient eu sans ex æquo.

$$W = \frac{S_K}{S_{\max}} = \frac{12S_K}{p^2(n^3-n) - p \sum_{j=1}^{j=p} (t_j^3 - t_j)}$$

avec t_j = nombre d'ex æquo au $j^{\text{ème}}$ classement

Cette table fournit les valeurs critiques k telles que $P(W \geq k) = \alpha = 0,05$

3. Le test de Mann et Whitney

Rangs des observations X dans l'échantillon global ordonné de taille $N = m + n$	R_1	R_2	...	R_k	...	R_m	
Rangs des observations Y dans l'échantillon global ordonné de taille $N = m + n$	Q_1	Q_2	...	Q_k	Q_n

Ce test est fondé sur la statistique suivante dérivée de celle de Wilcoxon :

$$S_N = U_N = W_N - \frac{m(m+1)}{2} = \sum_{k=1}^{k=m} R_k - \frac{m(m+1)}{2}$$

on peut en déduire que :

- U_N varie entre 0 et mn

- l'espérance est $E(U_N) = \frac{mn}{2}$

- la variance est $V(U_N) = V(W_N) = \frac{mn(m+n+1)}{12}$

3.1. Conditions d'utilisation:

Etant donnés les deux échantillons indépendants $(X_1, X_2, X_3, \dots, X_m)$ et $(Y_1, Y_2, Y_3, \dots, Y_n)$ issus de deux populations P_1 et P_2 .

On mélange ces deux échantillons et on réordonne les valeurs.

On dénombre les couples (X_p, Y_q) tels que X_p a un rang plus grand que Y_q

U_N est la variable qui à chaque situation associe ce nombre. Elle varie entre 0 et nm selon les deux cas extrêmes:

$$x_1, x_2, x_3, \dots, x_m, y_1, y_2, y_3, \dots, y_n \quad \text{et} \quad y_1, y_2, y_3, \dots, y_n, x_1, x_2, x_3, \dots, x_m,$$

Sous l'hypothèse de l'identité des distributions des deux variables X et Y , la loi exacte de U_N peut être calculée pour de faibles valeurs de n et de m . Toutefois dès que $n > 8$ et $m > 8$ on peut l'approcher par une loi de Laplace-Gauss.

$$\frac{U_N - \frac{nm}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \approx \text{LG}(0;1)$$

3.2. Statistique et variable de décision

Pour plus de faciliter, on utilise de façon intermédiaire la variable W_X , somme des rangs

de la variable X , puis on calcule $U_N = W_X - \frac{m(m+1)}{2}$

3.3. Test bilatéral :

H_0 (identité des deux distributions $F = G$) contre H_1 (Les deux distributions sont différentes $F \neq G$)

3.4. Test unilatéral :

H_0 (identité des deux distributions $F = G$) contre H_1 (Les deux distributions sont différentes soit $F < G$, soit $F > G$)

Il s'agit alors de prendre une décision sur la base d'une valeur u_N , réalisation de la variable U_N sur un échantillon. A l'aide d'une table ou d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss, on peut fonder le rejet ou le non-rejet au seuil α de l'hypothèse nulle H_0 postulant l'identité des deux distributions. La région critique W , région de rejet de H_0 au profit de H_1 , est définie selon H_1 , l'une des trois hypothèses alternatives usuelles.

H1 $F > G$	H1 $F \neq G$	H1 $F < G$
$W = \{ u_N, u_N \leq c_1 \}$	$W = \{ u_N, u_N \leq c_1 \text{ ou } u_N \geq c_2 \}$	$W = \{ u_N, u_N \geq c_2 \}$

La valeur critique c ($c \geq 0$) s'obtient en résolvant l'équation $\text{Prob}(W|H_0) = \alpha$

A l'aide d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss $\frac{U_N - E(U_N)}{\sigma(U_N)}$ tend vers $LG(0,1)$

Dans ce cas, il convient de tenir du passage d'une variable discrète à une variable continue en réalisant une correction (*correction de continuité*) selon la démarche suivante :

$\text{Prob}\{U_N = s\} = \text{Prob}\left\{s - \frac{1}{2} < U_N < s + \frac{1}{2}\right\}$ La valeur critique au seuil α s'obtient alors par la résolution d'équation

4. Test d'homogénéité

4.1. Comparaison de k échantillons indépendants Test H de Kruskal-Wallis

Rangs des observations X_1 dans l'échantillon global ordonné de taille $N = \sum_{p=1}^{p=k} n_p$	R_{11}	R_{12}	...	R_{1k}	...	R_{1n_1}			
Rangs des observations X_2 dans l'échantillon global ordonné de taille $N = \sum_{p=1}^{p=k} n_p$	R_{21}	R_{22}	...	R_{2k}		...	R_{2n_1}		
Rangs des observations X_p dans l'échantillon global ordonné de taille $N = \sum_{p=1}^{p=k} n_p$	R_{p1}	R_{p2}	...	R_{pk}			...	R_{pn_p}	
Rangs des observations X_k dans l'échantillon global ordonné de taille $N = \sum_{p=1}^{p=k} n_p$	R_{k1}	R_{k2}	...	R_{kk}				...	R_{kn_k}

4.2. Conditions d'utilisation:

Etant donnés les k échantillons indépendants respectivement de taille n_1, n_2, \dots, n_k issus de k populations $P_1, P_2 \dots P_k$. La variable étudiée est une variable ordinale

4.3. Statistique et variable de décision

On mélange ces k échantillons et on réordonne les $N = \sum_{p=1}^{p=k} n_p$ valeurs.

On prend en compte le rang de chaque observation dans le classement global

Sous l'hypothèse H_0 de l'identité des k distributions de la variable ordinale, les rangs sont distribués au hasard dans chaque échantillon. Considérons:

- la variable S_p = somme des rangs des n_p observations de l'échantillon $n^o p$

- la variable "rang moyen" est alors: $\frac{S_p}{n_p}$

- l'espérance de cette variable "rang moyen" sous H_0 :

$$E\left(\frac{S_p}{n_p}\right) = \frac{1}{n_p} E(S_p) = \frac{1}{n_p} \left(n_p \frac{N+1}{2} \right) = \frac{N+1}{2}$$

On mesure l'écart entre les résultats attendus sous l'hypothèse H_0 et les observations par la variable H :

$$H = \frac{12}{N(N+1)} \sum_{p=1}^{p=k} n_p \left(\frac{S_p}{n_p} - \frac{N+1}{2} \right)^2$$

La distribution exacte de H s'obtient par le dénombrement des m configurations équiprobables

$$m = \frac{n!}{n_1! n_2! \dots n_k!} \quad \text{où } n = \sum_{p=1}^{p=k} n_p$$

L'espérance de la variable H : $E(H) = k-1$

La variance de la variable H : $V(H) = 2(k-1) - \frac{2[3k^2-6k+N(2k^2-6k+1)]}{5N(N+1)} - \frac{6}{5} \sum_{p=1}^{p=k} \frac{1}{n_p}$

Cependant la distribution de H peut être approchée par la distribution de la variable de Pearson à $k-1$ ddl. Pour calculer la valeur expérimentale h , on peut aussi utiliser une

expression équivalente de la variable H : $H = \frac{12}{n(n+1)} \left\{ \sum_{p=1}^{p=k} \frac{S_p^2}{n_p} \right\} - 3(n+1)$

4.4. Test unilatéral : H_0 (identité des K distributions) contre H_1 (deux distributions au moins sont différentes)

On choisit un niveau de risque de 1ère espèce α . A ce seuil, on détermine la valeur critique c :

- soit à l'aide d'une table du H de Kruskal et Wallis
- soit à l'aide de la table de la variable de Pearson à $ddl = k-1$

Dans ce cas, on détermine la valeur c telle que $\text{Prob} \{ \chi^2(k-1) < c \} = 1 - \alpha$

On détermine la valeur expérimentale h de H par l'une des deux possibilités décrites ci-dessus: **si** $h > c$ alors **on rejette H_0** en prenant un risque de première espèce de niveau α **sinon on ne rejette pas H_0** , ce qui revient à accepter l'hypothèse H_1 en prenant un risque de seconde espèce β .

4.5. En cas d'existence d'ex æquo

En cas d'ex æµo chaque observation reçoit un rang égal à la moyenne des rangs qu'elles occupent. Pour chaque groupe de q observations ex æµo , on pose $Q = (q-$

$1)q(q+1)$. On calcule la somme $\sum_{i=1}^{i=r} Q_i$ des valeurs Q obtenues pour les r groupes d'ex æµo

Et on utilise la variable $H^* = \frac{H}{\sum_{i=1}^{i=r} Q_i} \cdot \frac{1}{1 - \frac{1}{n(n-1)}}$