Statistiques de rang, de signe et permutation

Lionel Truquet,
lionel.truquet@ensai.fr

1. Définition et propriétés des statistiques de rang

2. Statistiques de rang signées

3. Tests de permutation

Outline

1. Définition et propriétés des statistiques de rang
2. Statistiques de rang signées
3. Tests de permutation

Statistiques d'ordre et de rang

- Si $X_1, X_2, \ldots, X_n$ sont i.i.d. et à valuers réelles, on note $X_{n,1}, X_{n,2}, \ldots, X_{n,n}$ leur réarrangement croissant. En particulier $X_{n,1} = \min_{1\leq i\leq n} X_i$ et $X_{n,n} = \max_{1\leq i\leq n} X_i$.
- On supposera que les X_i ont une fonction de répartition continue. Ceci assure que $\mathbb{P}\left(\cup_{i,j=1}^{n}\{X_i = X_j\}\right) = 0$.
- On peut alors définir sans ambiguité le rang de X_i dans l'échantillon (p.s). Formellement $R_{n,i} = \sum_{j=1}^{n} \mathbb{1}_{X_j \leq X_i}$.
- On a alors $X_i = X_{n,R_{n,i}}$.
- La forme générale d'une statistique (linéaire) de rang est $\sum_{i=1}^{n} c_{n,i} a_{n,R_{n,i}}$ où les $c_{n,i}$ (les coefficients) et $a_{n,i}$ (appelés les rangs) sont des nombres réels donnés.

Exemples

- **Problème de localisation à deux échantillons.** On cherche à savoir si deux sous-échantillons indépendants $X_1, \ldots, X_{n_1}$ et $X_{n_1+1}, \ldots, X_n$ sont de même loi ou si la loi du second est stochastiquement plus grande que le premier. $T_n = \sum_{i=n_1+1}^{n} R_{n,i}$ est appelée statistique de Wilcoxon (on rejette pour les grandes valeurs). Une autre statistique est celle de van der Waerden, $T_n = \sum_{i=n_1+1}^{n} \Phi^{-1}(R_{n,i})$ où Φ^{-1} est la fonction quantile de la loi gaussienne.

- **Le test de la médiane** est basé sur la statistique $T_n = \sum_{i=n_1+1}^{n} \mathbb{1}_{R_{n,i} \leq \frac{n+1}{2}}$. On compte le nombre d'observations X_j, $n_1 + 1 \leq j \leq n$ plus petites que la médiane. On rejette l'équidistribution des deux échantillons lorsque T_n est élevé (ce qui signifie la distribution du deuxième échantillon est décalée sur la gauche par rapport au premier).

- On dit que U est **stochastiquement plus petite** que V lorsque $F_V(x) \leq F_U(x)$ pour tout $x \in \mathbb{R}$. C'est le cas par exemple si les densités de probabilité vérifient $f_V(x) = f_U(x + h)$ pour $h > 0$.

Propriétés des statistiques de rang

Lemma 1

Soient $X_1, \ldots, X_n$ i.i.d. de loi de densité f. On a alors les propriétés suivantes.

1. *Les vecteurs $X_{n,\cdot}$ et $R_{n,\cdot}$ sont indépendants.*
2. *La densité h_n de $X_{n,\cdot}$ est donnée par $h_n(x_1, \ldots, x_n) = n! \prod_{i=1}^n f(x_i) \mathbb{1}_{x_1 < \cdots < x_n}$.*
3. *Le vecteur R_n est distribué uniformément sur l'ensemble $\mathcal{S}_n$ des permutations de $\{1, \ldots, n\}$.*
4. *Si $T : \mathbb{R}^n \to \mathbb{R}$ et $\sigma \in \mathcal{S}_n$, alors*
$$\mathbb{E}\left(T(X_1, \ldots, X_n) | R_n = \sigma\right) = \mathbb{E}\left(T\left(X_{n,\sigma(1)}, \ldots, X_{n,\sigma(n)}\right)\right).$$
5. *Si $T = \sum_{i=1}^n c_{n,i} \cdot a_{n,R_{n,i}}$, alors*
$$\mathbb{E}T = n\bar{c}_n \bar{a}_n, \quad Var(T) = \frac{1}{n-1} \sum_{i=1}^n (c_{n,i} - \bar{c}_n) \sum_{i=1}^n (a_{n,i} - \bar{a}_n)^2.$$

Conséquences

- Le fait que R_n est de loi uniforme sur $\mathcal{S}_n$ guarantit que toute statistique basée sur les rangs a une distribution de probabilité qui ne dépend pas de la loi des observations.
- Si n est grand, il peut être difficile de calculer numériquement les quantiles d'une statistique de rang. On a aussi des propriétés asymptotiques (voir plus loin).
- Il existe une méthode pour construire des statistiques linéaires des rangs afin de tester des problèmes spécifiques et qui ont une bonne puissance pour des alternatives données.

Tests de rang localement le plus puissant

- Imaginons que le problème de test se ramène à tester la nullité d'un paramètre. Le lemme de Neyman-Pearson assure que pour tester un problème du type $H_0 : \theta = 0$ contre $H_1 : \theta = \theta'$, la statistique basée sur R_n et qui conduit au test le plus puissant est $T(R_n)$ où

$$T(\sigma) = \frac{P_{\theta'}(R_n = \sigma)}{P_0(R_n = \sigma)} = n! P_{\theta'}(R_n = \sigma).$$

- Lorsque le problème de test considéré peut se ramener à tester la nullité d'un certain paramètre, l'idée est alors de faire un développement limité de $\theta \mapsto P_\theta(R_n = \sigma)$ au voisinage de 0 et garder la partie linéaire $\frac{d}{d\theta} P_\theta(R_n = \sigma)|_{\theta=0}$.
- Cette méthode permet de déterminer les scores $a_{n,i}$. Le test obtenu est alors très puissant (parmi les statistiques de tests basées sur les rangs) lorsque l'alternative $H_1 : \theta = \theta'$ pour θ' petit.
- Cette approche est intéressante au sens suivant. Lorsqu'on est loin de l'hypothèse nulle et n est grand, tout test pertinent aura une puissance raisonnable. La différence de puissance entre plusieurs tests est intéressante lorsque $\theta' = \theta'_n$ tend vers 0.

Tests de rang localement le plus puissant

- Supposons que sous l'hypothèse nulle, $X_1, \ldots, X_n$ soient i.i.d. de loi à densité f_0 et que sous l'alternative X_i ait une loi à densité $f_{c_{n,i}\varepsilon}$.

- Par exemple, dans le problème de test de localisation à deux échantillons, on a $c_{n,i} = 0$ pour $1 \leq i \leq n_1$ et $c_{n,i} = 1$ pour $n_1 + 1 \leq i \leq n = n_1 + n_2$.

- Un calcul montre que (sous certaines conditions de régularité), les scores "optimaux" sont données par

$$a_{n,i} = E_0\left[\frac{\dot{f}_0}{f_0}(X_{n,i})\right] = \mathbb{E}\left[\frac{\dot{f}_0}{f_0}\left(F_0^{-1}(U_{n,i})\right)\right].$$

- On dit alors que la fonction $\phi = \frac{\dot{f}_0}{f_0} \circ F_0^{-1}$ est la fonction génératrice des scores.

Exemple 1 : problème de localisation à deux échantillons

- Pour trouver un test localement le plus puissant pour se problème, on regarde des alternatives du type $f_{X_i}(x) = f(x-\theta)$ pour $n_1 + 1 \leq i \leq n$ (f désigne la densité de X_1 et θ est positif).
- En appliquant les calculs précédents, on trouve que $a_{n,i} = -\mathbb{E}\left[\frac{f'}{f}\left(F^{-1}(U_{n,i})\right)\right]$.
- Lorsque f est la densité Gaussienne (resp. logistique), on retrouve la statistique de van der Waerden (resp. Wilcoxon).

Exemple 2 : test du log-rank

- En survie, la fonction de hasard cumulée d'une fonction de répartition F est $\Lambda = -\log(1-F)$.
- On souhaite tester si deux échantillons indépendants on la même fonction de hasard cumulée.
- Sous l'hypothèse des risques proportionnels, on considère que sous l'alternative, $\Lambda_\theta = (1+\theta)\Lambda_0$.
- On peut montrer que la fonction génératrice des scores est donnée par $\phi(u) = -\log(1-u)$. On parle alors du test de log-rank.
- On a aussi le test de Savage qui utilise $a_{n,i} = \sum_{j=n-i+1}^{n} \frac{1}{j} \approx -\log\left(1 - \frac{i}{n+1}\right)$.

Propriétés asymptotiques

- On a vu que les scores associés à un test de rang localement le plus puissant sont de la forme $a_{n,i} = \mathbb{E}\phi(U_{n,i})$, où $U_1, \ldots, U_n$ sont i.i.d. et de loi uniforme sur $[0, 1]$.
- Les autres tests vus précédemment sont de la forme $a_{n,i} = \phi\left(\frac{i}{n+1}\right)$. Noter que $\mathbb{E}U_{n,i} = \frac{i}{n+1}$.
- On va montrer l'équivalence asymptotique entre $T_n = \sum_{i=1}^{n} c_{n,i} a_{n,R_{n,i}}$ et une somme pondérée de variables aléatoires i.i.d.

Propriétés asymptotiques

On pose $T'_n = n\bar{c}_n\bar{a}_n + \sum_{i=1}^n (c_{n,i} - \bar{c}_n)\,\phi\,(F(X_i))$.

Theorem 1

On suppose que la fonction de répartition F des observations est continue et strictement croissante.

1. *On suppose que $a_{n,i} = \mathbb{E}\phi(U_{n,i})$ avec ϕ non constante presque partout et $\int_0^1 \phi(u)^2 du < \infty$. On a alors $\frac{T_n - \mathbb{E}T_n}{sd(T_n)} - \frac{T'_n - \mathbb{E}T'_n}{sd(T'_n)} = o_P(1)$.*

2. *Lorsque $a_{n,i} = \phi\left(\frac{i}{n+1}\right)$, on suppose en plus que ϕ est continue presque partout et que $n^{-1}\sum_{i=1}^n \phi^2(i/(n+1)) \to \int_0^1 \phi^2(u)du < \infty$. Alors la conclusion du point précédent reste vraie.*

Corollary 1

Sous l'un ou l'autre des jeux d'hypothèses ci-dessus et si en plus $\lim_{n\to\infty} \frac{\max_{1\le i\le n}(c_{n,i}-\bar{c}_n)^2}{\sum_{i=1}^n (c_{n,i}-\bar{c}_n)^2} = 0$, alors $(T_n - \mathbb{E}T_n)/sd(T_n) \Rightarrow \mathcal{N}(0,1)$.

Autre exemple. Test d'indépendance à partir des rangs

- Lorsque $X_i = (Y_i, Z_i)$, on se demande si Y_i est indépendant de Z_i. On utilise des statistiques du type $T_n = \sum_{i=1}^n a_{n,R_{n,i}} b_{n,S_{n,i}}$ où a_n et b_n sont croissants, R_n sont les rangs des Y_i (dans $\{Y_1, \ldots, Y_n\}$) et S_n sont les rangs des Z_i.

- On définit alors les antirangs $R_n^0(\omega)$ qui sont les rangs de $\{X_{\sigma(1)}(\omega), \ldots, X_{\sigma(n)}(\omega)\}$ lorsque $Y_{\sigma(1)}(\omega) < \cdots < Y_{\sigma(n)}(\omega)$.

- Sous l'hypothèse d'indépendance, R_n^0 est distribué uniformément sur l'ensemble des permutations de $\{1, \ldots, n\}$. De plus

$$\sum_{i=1}^n a_{n,R_{n,i}} b_{n,S_{n,i}} = \sum_{i=1}^n a_{n,R^0_{n,i}} b_{n,i}.$$

- Correlation des rangs de Spearman. On pose $a_{n,i} = b_{n,i} = i$. On a alors une équivalence avec le le test basé sur le coefficient de corrélation des rangs.

$$\rho_n = \frac{\sum_{i=1}^n \left(R_{n,i} - \overline{R}_n\right)\left(S_{n,i} - \overline{S}_n\right)}{\sqrt{\sum_{i=1}^n \left(R_{n,i} - \overline{R}_n\right)^2 \sum_{i=1}^n \left(S_{n,i} - \overline{S}_n\right)^2}} = \frac{12 \sum_{i=1}^n R_{n,i} S_{n,i}}{n(n-1)(n+1)} - 3\frac{n+1}{n-1}.$$

Outline

1. Définition et propriétés des statistiques de rang
2. Statistiques de rang signées
3. Tests de permutation

Quelques définitions

- $sign(x)$ vaut -1, 0 ou 1 si $x < 0$, $x = 0$ et $x > 0$ respectivement.
- On note R_n^+ les rangs des observations $|X_i|$. On parle de rang absolu.
- Une statistique de rang signée est de la forme $T_n = \sum_{i=1}^n a_{n,R_{n,i}^+} sign(X_i)$.
- On peut retrouver les rangs ordinaires à partir des rangs absolus et des signes. On a donc plus d'information. Cette approche est intéressante lorsque la loi des donnée est supposée symétrique.

Propriétés

Lemma 2

On suppose la distribution des données continue et symétrique par rapport à 0. On a alors les propriétés suivantes.

1. *Les vecteurs $(|X|, R_n^+)$ et $sign_n(X)$ sont indépendants. R_n^+ est uniformément distribué sur $\mathcal{S}_n$. $sign_n(X)$ est uniformément distribué sur $\{-1, 1\}^n$.*
2. *Une statistique de rang signée est centrée et de variance $Var(T_n) = \sum_{i=1}^n a_{n,i}^2$.*

Par exemple, la statistique de rang signée de Wilcoxon est définie par $T_n = \sum_{i=1}^n R_{n,i}^+ sign(X_i)$ est obtenu avec une fonction génératrice des scores $\phi(u) = u$. Une large valeur pour T_n indique que la présence de large valeurs positives. On peut l'utiliser pour tester $H_0 : \theta = 0$ lorsque la loi des donnée est de densité $f(\cdot - \theta)$, avec f symétrique.

Convergence

On pose $\overline{T}_n = \sum_{i=1}^{n} \phi\left(F^+(|X_i|)\right) sign(X_i)$ où ϕ est la fonction génératrice des scores et F^+ la fonction de répartition de $|X_1|$.

Theorem 2

Soient $X_1, \ldots, X_n$ i.i.d. avec une distribution continue et symétrique autour de 0. On suppose $\int_0^1 \phi(u)^2 du < \infty$.

1. *Si $a_{n,i} = \mathbb{E}\phi(U_{n,i})$, T_n est asymptotiquement équivalente à $\overline{T}_n$ et $n^{-1/2}T_n$converge en loi vers une gaussienne de moyenne 0 et de variance $\int_0^1 \phi(u)^2 du$.*
2. *Si $a_{n,i} = \phi(i/(n+1))$ et qu'en plus ϕ est continue presque partout et que $n^{-1}\sum_{i=1}^{n} \phi^2(i/(n+1)) \to \int_0^1 \phi^2(u)du$, alors on a la même conclusion que pour le point précédent.*

Test localement le plus puissant pour la localisation

- Lorsqu'on veut tester la symétrie et qu'une alternative est $f_{X_i}(x) = f(x-\theta)$ avec f symétrique autour de 0, on peut faire un développement limité de $\theta \mapsto P_\theta\left(sign_n(X) = s, R_n^+ = r\right)$ en 0.
- En examinant le premier ordre, on trouve une fonction génératrice des scores $\phi = -(f'/f) \circ (F^+)^{-1}$. On a sous l'hypothèse nulle, $(F^+)^{-1}(u) = F^{-1}((u+1)/2)$.
- Pour la gaussienne, on a $a_{n,i} = \Phi^{-1}\left(\left(U_{n,i}+1\right)/2\right)$.

Outline

1 Définition et propriétés des statistiques de rang

2 Statistiques de rang signées

3 Tests de permutation

Tests de permutation

- Un test de permutation est obtenu en tirant aléatoirement des permutations des données. Les tests de rang sont des exemples de ce type.

- On peut par exemple tester si deux échantillons ont la même loi à partir de

$$T_n = T_n(X_1, \ldots, X_n) = \frac{1}{n_1} \sum_{i=1}^{n_1} f(X_i) - \frac{1}{n - n_1} \sum_{i=n_1+1}^{n} f(X_i),$$

  pour une fonction f bien choisie. Si $n_1/n \to \lambda \in (0,1)$, le TLC assure que sous H_0,

$$\sqrt{n} T_n \Rightarrow \mathcal{N}(0, \sigma^2), \quad \sigma^2 = \frac{\mathrm{Var} f(X_1)}{\lambda(1-\lambda)}.$$

- On pourrait utiliser la loi limite précédente. L'idée des tests de permutation est d'utiliser plutôt la loi de $T_n\left(X_{\sigma_{n,1}}, \ldots, X_{\sigma_{n,n}}\right)$ conditionellement à $X_1, \ldots, X_n$ et où σ_n est une permutation aléatoire de $\{1, \ldots, n\}$ (indépendante des données). On peut alors simuler un quantile qui dépend des données.

Tests de permutation

Theorem 3

Supposons que $\mathbb{E}\left[f^2(X_1)+f^2(Y_1)\right]<\infty$ *et* $n, n_1 \to \infty$ *avec* $n_1/n \to \lambda \in (0,1)$. *Alors, presque sûrement,* $\sqrt{n}T_n\left(X_{\phi_{n,1}},\ldots,X_{\phi_{n,n}}\right)$ *est asymptotiquement Gaussienne et centrée. Sous* H_0*, la variance vaut* Var$f(X_1)/(\lambda(1-\lambda))$.

- Le résultat précédent justifie le tirage aléatoire d'une permutation (même loi sous H_0). On peut donc simuler un grand nombre de réalisations de la statistique de test.
- On notera que le quantile est propre au jeu de données et que le coût numérique est important.