Notions fondamentales en statistique

Frédéric Bertrand et Myriam Maumy

IRMA, UMR 7501, Université de Strasbourg

08 septembre 2011

$5^{\text{ème}}$ partie

Les tests non-paramétriques II

Sommaire

Sommaire

Tests non libres de distribution

Certains tests statistiques ne sont valables que sous certaines conditions concernant la distribution de la ou les variable(s).

Exemples

Le test de Student impose que les deux variables sont issues d'une distribution normale, l'analyse de la variance également.

Tests libres de distribution

D'autres tests au contraire sont valables indépendamment de toute distribution. Nous les appelons les tests « libres de distribution » (distribution-free tests).

Exemples

C'est le cas du test du Khi-deux, du test des signes, ou du test du coefficient de Spearman.

Tests paramétriques

Certains tests ont pour but de montrer une égalité sur certains paramètres : ce sont les tests paramétriques.

Exemples de paramètres

1. La moyenne (test de comparaison de deux moyennes ou plus),
2. la variance (test de comparaison de deux variances ou plus),
3. etc.

Tests non paramétriques

D'autres tests testent des hypothèses plus générales : ce sont les tests non paramétriques.

Exemples

1. Une égalité de lois,
2. l'indépendance entre deux variables qualitatives,
3. etc.

Une question naturelle : quel test choisir ?

Habituellement, les tests paramétriques sont plus puissants. Par conséquent, ils seront choisis plutôt que les tests non paramétriques.
De même les tests non libres sont généralement plus efficaces que les tests libres. Cependant, ils sont aussi plus contraignants, car il faut vérifier les conditions d'application qui sont plus nombreuses dans ce cas.
On choisira généralement un test libre ou non paramétrique lorsque

1. les conditions d'application du test ne sont pas vérifiées
2. ou il est impossible de vérifier ces conditions.

Sommaire

Introduction

Nous observons, de manière indépendante, une variable aléatoire X de loi continue, sur $k \geqslant 3$ populations, ou sur une population divisée en $k \geqslant 3$ sous-populations.

Nous supposons ainsi que nous disposons de k échantillons aléatoires **indépendants** $(X_{1,1}, \ldots, X_{1,n_1})$, ..., $(X_{k,1}, \ldots, X_{k,n_k})$ et de $k \geqslant 3$ séries d'observations $(x_{1,1}, \ldots, x_{1,n_1})$ pour la première, ..., $(x_{k,1}, \ldots, x_{k,n_k})$ pour la dernière. Nous notons $\mathcal{L}_i(X)$ la loi de la variable aléatoire X sur la (sous-)population d'ordre i avec $1 \leqslant i \leqslant k$.

Sans faire d'hypothèses spécifiques, le test de Kruskal-Wallis ne permet pas de tester l'égalité des moyennes ni celle des médianes.

Hypothèses du test

Le **test de Kruskal-Wallis** est utilisé pour tester les hypothèses suivantes :

$$\boxed{\mathcal{H}_0 : \mathcal{L}_1(X) = \cdots = \mathcal{L}_i(X) = \cdots = \mathcal{L}_k(X)}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Les lois } \mathcal{L}_1(X), \ldots, \mathcal{L}_k(X) \text{ ne sont pas toutes identiques.}}$$

Sommaire

Statistique de test

Calculons :

1. le rang $R_{i,j}$ de $X_{i,j}$ parmi les $n_\bullet$ valeurs ;
2. puis la somme des rangs associée à chaque échantillon :

$$R_{i,\bullet} = \sum_{j=1}^{n_i} R_{i,j}\,;$$

3. la moyenne des rangs de chaque échantillon :

$$\overline{R_{i,\bullet}} = R_{i,\bullet}/n_i.$$

Statistique de test (suite)

La statistique de Kruskal-Wallis $KW_{n_\bullet}$ prend en compte l'écart entre la moyenne des rangs de chaque échantillon et la moyenne de tous les rangs, qui vaut $(n_\bullet + 1)/2$:

$$\begin{aligned} KW_{n_\bullet} &= \frac{12}{n_\bullet(n_\bullet+1)} \sum_{i=1}^{k} n_i \left(\overline{R_{i,\bullet}} - \frac{n_\bullet + 1}{2} \right)^2 \\ &= \frac{12}{n_\bullet(n_\bullet+1)} \sum_{i=1}^{k} \frac{{R_{i,\bullet}}^2}{n_i} - 3(n_\bullet + 1). \end{aligned}$$

Propriétés

Lorsque l'hypothèse nulle $\mathcal{H}_0$ est vraie, la variable aléatoire $KW_{n_\bullet}$ a les trois propriétés suivantes.

1. Pour $i = 1, \ldots, k$, $W_i = n_i \overline{R_{i,\bullet}}$ est la statistique de Wilcoxon qui compare le $i-$ème traitement aux $k-1$ autres traitements.

   Sous l'hypothèse nulle $\mathcal{H}_0$, nous en déduisons que $\mathbb{E}\left(W_i\right) = n_i(n_\bullet + 1)/2$ et $\mathrm{Var}\left(W_i\right) = n_i(n_\bullet - n_i)(n_\bullet + 1)/12$.

   Ainsi par conséquent, nous avons :

$$KW_{n_\bullet} = \frac{1}{n_\bullet} \sum_{i=1}^{k} (n_\bullet - n_i) \frac{\left(W_i - \mathbb{E}\left(W_i\right)\right)^2}{\mathrm{Var}\left(W_i\right)}.$$

Propriétés (suite)

2. **(Suite)** Nous calculons alors l'espérance et la variance de $KW_{n_\bullet}$ sous l'hypothèse nulle $\mathcal{H}_0$:

$$\begin{aligned}
\mathbb{E}\left(KW_{n_\bullet}\right) &= k-1, \\
\mathrm{Var}\left(KW_{n_\bullet}\right) &= 2(k-1) - \frac{2\left[3k^2-6k+n_\bullet(2k^2-6k+1)\right]}{5n_\bullet(n_\bullet+1)} \\
&\quad -\frac{6}{5}\sum_{i=1}^{k}\frac{1}{n_i}.
\end{aligned}$$

3. Il est possible de déterminer la distribution de $KW_{n_\bullet}$ bien que le calcul soit complexe. Elle est donc tabulée pour les faibles valeurs des n_i.

Règle de décision et conclusion du test

- **Premier cas :** L'un des effectifs n_i, $1 \leqslant i \leqslant k$, est inférieur ou égal à 4, . Pour un seuil donné α, des tables de la loi de Kuskal-Wallis nous fournissent une valeur critique c_α. Alors nous décidons :

$$\begin{cases} si\ KW_{n_\bullet}(obs) \geqslant c_\alpha & \mathcal{H}_1\ est\ vraie, \\ si\ KW_{n_\bullet}(obs) < c_\alpha & \mathcal{H}_0\ est\ vraie. \end{cases}$$

  Le niveau de signification réel du test est généralement strictement inférieur à α.

Règle de décision et conclusion du test (suite)

- **Second cas :** Si $n_i \geqslant 5$, pour tout $1 \leqslant i \leqslant k$, nous utilisons l'approximation $KW_{n_\bullet} \approx \chi^2(k-1)$. Pour un seuil donné α, des tables de la loi du χ^2 nous fournissent une valeur critique c_α telle que $\mathbb{P}_{\mathcal{H}_0}(-c_\alpha < Z_{n_1,n_2} < c_\alpha) = 1 - \alpha$. Alors nous décidons :

$$\begin{cases} si\ KW_{n_\bullet}(obs) \geqslant c_\alpha & \mathcal{H}_1\ est\ vraie, \\ si\ KW_{n_\bullet}(obs) < c_\alpha & \mathcal{H}_0\ est\ vraie. \end{cases}$$

  Lorsque nous rejetons $\mathcal{H}_0$, nous décidons que $\mathcal{H}_1$ est vraie avec un risque d'erreur de première espèce α.
  Lorsque nous conservons $\mathcal{H}_0$, c'est avec un risque d'erreur de deuxième espèce β.

Présence d'ex æquo dans les observations : la méthode des rangs moyens

À chaque nombre appartenant à un groupe d'ex æquo nous attribuons le rang moyen du groupe auquel il appartient puis nous déterminons la somme $T = \sum_{l=1}^{h}(t_l^3 - t_l)$ où t_l désigne le nombre d'éléments du $l-$ème groupe d'ex æquo. Il est d'usage de substituer à $KW_{n_\bullet}$ la variable $KW_{n_\bullet}^\star$ définie par :

$$KW_{n_\bullet}^\star = \frac{KW_{n_\bullet}}{1 - \dfrac{T}{n_\bullet^3 - n_\bullet}}.$$

Sommaire

Test de Steel-Dwass-Critchlow-Fligner

$W_{n_i,n_{i'}}$ est la statistique de Wilcoxon qui compare le $i-$ème traitement au $i'-$ème traitement.

Les observations des deux groupes i et i' sont ordonnées puis regroupées en h classes d'ex æquo C_j, $1 \leqslant j \leqslant h$.

Notons d_j le nombre d'ex æquo de la classe C_j et $m_{i,i'} = n_i + n_{i'}$.

Test de Steel-Dwass-Critchlow-Fligner (suite)

Nous décidons qu'**au seuil global** α deux lois $\mathcal{L}_i(X)$ et $\mathcal{L}_{i'}(X)$, parmi les $\mathbf{k(k-1)}$ **comparaisons** que nous allons faire, sont significativement différentes si :

$$\begin{aligned}\left|W_{n_i,n_{i'}} - \frac{n_i(m_{i,i'}+1)}{2}\right| &\geqslant q'(k;+\infty;1-\alpha)\sqrt{\frac{n_i n_{i'}(m_{i,i'}+1)}{24}} \\ &\quad \times \sqrt{\left(1 - \frac{\sum_{j=1}^{h}\left(d_j^3 - d_j\right)}{m_{i,i'}^3 - m_{i,i'}}\right)}\end{aligned}$$

où $q'(k;+\infty;1-\alpha)$ est le quantile d'ordre $1-\alpha$ pour la loi du maximum du module studentisé pour k moyennes et $+\infty$ degrés de liberté.

Test de Steel-Dwass-Critchlow-Fligner (suite)

Contrairement aux trois autres approches présentées ci-après, le **test Steel-Dwass-Critchlow-Fligner** n'est pas qu'une procédure de comparaisons multiples : c'est une **alternative complète** au test de **Kruskal-Wallis**.

Autres tests de comparaisons multiples

Si nous rejetons l'hypothèse nulle $\mathcal{H}_0$, nous nous demandons quelles sont les lois $\mathcal{L}_i(X)$ qui diffèrent.

Les formules ci-après sont valables en absence ou en présence d'ex æquo. En absence d'ex æquo, le terme $1 - T/(n_\bullet^3 - n_\bullet)$ est égal à 1.

Application de la méthode de Sheffé

Nous décidons qu'**au seuil α deux lois** $\mathcal{L}_i(X)$ et $\mathcal{L}_{i'}(X)$ sont significativement différentes si :

$$\left|\overline{R_{i,\bullet}} - \overline{R_{i',\bullet}}\right| \geqslant \sqrt{\chi^2(k-1;1-\alpha)}\sqrt{\frac{n_\bullet(n_\bullet+1)}{12}\left(1-\frac{T}{n_\bullet^3-n_\bullet}\right)} \times\sqrt{\frac{1}{n_i}+\frac{1}{n_{i'}}},$$

où $\chi^2(k-1;1-\alpha)$ est le quantile d'ordre $1-\alpha$ pour la loi du χ^2 à $k-1$ degrés de liberté.

Application de l'inégalité de Bonferroni

Nous décidons qu'**au seuil global α** deux lois $\mathcal{L}_i(X)$ et $\mathcal{L}_{i'}(X)$, parmi les **$\mathbf{k(k-1)}$ comparaisons** que nous allons faire, sont significativement différentes si :

$$\begin{aligned}\left|\overline{R_{i,\bullet}} - \overline{R_{i',\bullet}}\right| \quad \geqslant \quad & u\left(1-\frac{\alpha}{k(k-1)}\right)\sqrt{\frac{n_\bullet(n_\bullet+1)}{12}\left(1-\frac{T}{n_\bullet^3-n_\bullet}\right)} \\ & \times\sqrt{\frac{1}{n_i}+\frac{1}{n_{i'}}},\end{aligned}$$

où $u(1-\alpha/(k(k-1)))$ est le quantile d'ordre $1-\alpha/(k(k-1))$ pour la loi normale centrée-réduite.

Remarque

Il s'agit d'une application des inégalités de Bonferroni. Cette procédure est plus puissante que la précédente.

Il est également possible d'utiliser une approche séquentielle comme la procédure de Holm-Bonferroni (test de Dunn) ou de Holm-Sidak.

Analogue de la méthode de Tukey-Kramer

Nous décidons qu'**au seuil global α** deux lois $\mathcal{L}_i(X)$ et $\mathcal{L}_{i'}(X)$, parmi les **$\mathbf{k(k-1)}$ comparaisons** que nous allons faire, sont significativement différentes si :

$$\left|\overline{R_{i,\bullet}} - \overline{R_{i',\bullet}}\right| \geqslant q(k;+\infty;1-\alpha)\sqrt{\frac{n_\bullet(n_\bullet+1)}{12}\left(1-\frac{T}{n_\bullet^3-n_\bullet}\right)} \times\sqrt{\frac{1}{2}\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)},$$

où $q(k;+\infty;1-\alpha)$ est le quantile d'ordre $1-\alpha$ pour la loi de l'étendue studentisée pour k moyennes et $+\infty$ degrés de liberté.

Remarque

Il s'agit d'une procédure analogue à celle de Tukey-Kramer dans le cas paramétrique et valide asymptotiquement. Elle est généralement plus puissante que les deux approches précédentes.

Sommaire

Application

Des forestiers ont réalisé des plantations d'arbres en trois endroits.

Plusieurs années plus tard, ils souhaitent savoir si la hauteur moyenne des arbres est identique dans les trois forêts. Chacune des forêts constitue une population et dans chacune d'entre elles, un échantillon d'arbres est tiré au sort. Puis la hauteur de chaque arbre est mesurée en mètres. Le tableau ci-après donne les valeurs des mesures de hauteur ainsi que les moyennes et les variances corrigées estimées dans chacun des groupes.

Données

	Forêt 1	Forêt 2	Forêt 3
Données expérimentales (en mètres)	23,4	18,9	22,5
	24,4	21,1	22,9
	24,6	21,1	23,7
	24,9	22,1	24,0
	25,0	22,5	24,0
	26,2	23,5	24,5
Moyennes	24,750	21,533	23,600
Variances corrigées	0,831	2,487	0,568

La Figure 1 correspond aux boîtes à moustaches de la hauteur en fonction des trois forêts.

Représentation graphique

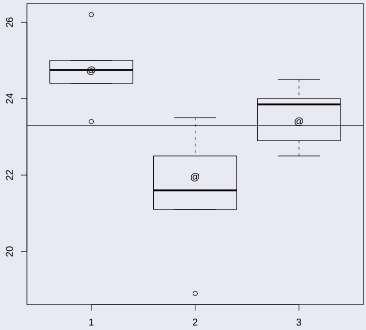

Figure: Boîtes à moustaches (à gauche) – Comparaisons multiples (à droite)

Statistique du test

Nous sommes en présence d'ex æquo, nous devons donc utiliser la statistique de test $KW^{\star}_{n_\bullet}$ à la place de la statistique de test $KW_{n_\bullet}$.

Nous calculons la valeur de $KW^{\star}_{n_\bullet}$ sur l'échantillon : $KW^{\star}_{n_\bullet}(obs) = 11,51$.

Règle de décision à l'aide d'une valeur critique

Pour un seuil $\alpha = 5\%$ la valeur critique d'un Khi-deux à 2 degrés de liberté, est $c_{0,05} = 5,99$. Comme $KW^{\star}_{n_\bullet}(obs) \leqslant c_{0,05}$, nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$, et que l'hypothèse alternative $\mathcal{H}_1$ est vraie. Il y a une influence significative, au seuil $\alpha = 5\%$, de la forêt sur la hauteur des arbres. Le risque associé à cette décision est un risque de première espèce qui vaut $\alpha = 5\%$.

Remarque

La valeur de $KW_{n_\bullet}(obs)$ est égale à $11,48$. Nous remarquons la différence apportée par la correction pour prendre en compte les ex æquo.

Règle de décision à l'aide d'une *p*-valeur

En utilisant un logiciel de statistique nous calculons la *p*-valeur du test de Kruskal-Wallis. Il faut bien vérifier qu'elle tient compte des ex æquo. Elle vaut dans cas $0,003$.

Comme la *p*-valeur est $\leqslant 0,05$, nous décidons, au seuil $\alpha = 5\%$, de rejeter l'hypothèse nulle $\mathcal{H}_0$, et que l'hypothèse alternative $\mathcal{H}_1$ est vraie. Il y a une influence significative, au seuil $\alpha = 5\%$, de la forêt sur la hauteur des arbres. Le risque associé à cette décision est un risque de première espèce qui vaut $\alpha = 5\%$.