Biostatistique

Tests d'hypothèses: cas 2 échantillons

Anicet Ebou, Institut National Polytechnique Félix Houphouët-Boigny,
ediman.ebou@inphb.ci

Plan

Introduction

Introduction

Nous avons vu les tests paramétriques sur un échantillon. Nous allons finir les tests paramétriques avec le cas “2 échantillons”, puis nous allons voir certains tests non paramétriques:

- Tests sur la normalité (Shapiro-Wilk);
- Test d’ajustement du Khi-deux;
- Test d’indépendance entre deux variables (test du Khi-deux).

L’analyse de la variance sera abordée séparemment vu son importance et sa grande ubiquité en biologie et expérimentation biologique.

Tests d'hypothèses avec 2 échantillons

Tests d'hypothèses avec 2 échantillons

Le tableau en pièce-jointe du cours résume les tests d'hypothèses pour différentes situations, et en particulier pour deux échantillons.

Exercice 1:

Une analyse statistique descriptive des données sur des mesures du diamètre (en cm) des roses produites par deux exploitations horticoles A et B a donné les résultats suivants:

Exploitation	A	B
nombre d'observations	51	51
$\overline{x}$	1,375147	1,374982
s	0,000167	0,000172

Au seuil critique 1% peut-on affirmer que le diamètre des pièces de l'exploitation A est supérieur en moyenne à celui des pièces de l'exploitation B?

Tests sur la normalité

Tests sur la normalité

But: vérifier si les données d'un échantillon proviennent d'une population normale.

Méthode graphique

Soit $X_{(1)}, X_{(2)}, ..., X_{(n)}$ les données de l'échantillon rangées dans l'ordre croissant.

On compare ces données aux centiles d'une loi normale en traçant dans le plan les point $(X_{(j)}, z_j)$, où

$$z_j = F_Z^{-1}\left(\frac{j-\frac{1}{2}}{n}\right) \text{ avec } \quad j = 1, 2, ..., n \quad \text{et} \quad Z \sim N\left(\overline{X}, S^2\right)$$

Si les données proviennent d'une loi normale alors $X_{(j)} \simeq z_j$ et les points obtenus s'alignent approximativement sur une droite.

Le graphe ainsi obtenu est un *graphe de probabilité normal*.

Tests sur la normalité (ii)

Tests de normalité:

Il s'agit de tester

$$H_0 : X \sim \text{Normale}$$

$$H_1 : X \text{ ne suit pas une loi normale}$$

Il existe plusieurs types de tests qui sont généralement réalisés à l'aide d'un logiciel. Les logiciels calculent la statistique du test et la P-value correspondante.

On distingue, entre autres: le test d'Anderson-Darling, le test d'Agostino-Pearson, le test de Geary, le test du Khi-deux, le test de Kolmogorov-Smirnov (statistique D) et le test de Shapiro-Wilk (statistique W).

Nous nous intéressons en particulier au test de Kolmogorov-Smirnov et Shapiro-Wilk.

Tests sur la normalité: test de Shapiro-Wilk

Ce test calcule une statistique W représentant le carré d'un coefficient de corrélation entre les $X_{(j)}$ observés et les centiles théoriques z_i d'une loi normale $N(0,1)$.

- On peut montrer que $c \leq W \leq 1$, où $c \simeq 0,70$.
- En pratique, on rejette $H_0 : X \sim$ Normale lorsque la valeur de W est petite (proche de 0,70).

Remarque: Lorsque l'hypothèse de normalité est acceptée (une grande P-value) il est important de confirmer l'hypothèse à l'aide des différents graphiques (quantile-quantile, etc.). Car , comme dans tout test statistique, l'acceptation de H_0 n'est pas une preuve que l'hypothèse soit vraie.

Test d'ajustement du Khi-deux de Pearson

Test d'ajustement du Khi-deux

- On cherche à vérifier si les données $x_1, ..., x_n$ dont on dispose proviennent d'une population distribuée selon une loi particulière $F(x, \theta)$.
- A partir d'un échantiloon aléatoire $X_1, ..., X_n$ de taille n d'une variable X, on va tester les hypothèses:

$$H_0 : X \sim F(x, \theta)$$

$$H_1 : X \neq F(x, \theta)$$

Test d'ajustement du Khi-deux: méthode

- On procède à un regroupement des observations selon k valeurs (ou intervalles). On obtient ainsi un tableau dont la forme générale est:

Valeurs (x_i)	V_1	V_2	...	V_i	...	Total
Effectifs observés (O_i)	O_1	O_2	...	O_i	...	n
Effectifs attendus (E_i)	E_1	E_2	...	E_i	...	n

Les O_i sont les effectifs observés, tandis que les E_i sont les effectifs attendus lorsque H_0 est vraie.

Procédure:

- Si on constate des E_i petits, il faut regrouper des classes. On recommande généralement de choisir les intervalles de sorte que $E_i \geq 5, \forall i$.
- On calcule les effectifs attendus $E_i = n \times p_i^{(0)}$ où $P_i^{(0)} = P(X \in V_i \mid H_0 \text{ est vraie}), i = 1, 2, ..., k$ et $\sum_{i=1}^{k} p_i^{(0)} = 1$.

Test d'ajustement du Khi-deux: méthode (ii)

- On calcule la statistique du test

$$\chi_0^2 = \sum_{i=1}^{k} \left(\frac{(O_i - E_i)^2}{E_i} \right)$$

- La statistique X_0^2 représente une sorte de "distance" globale entre les effectifs observés et les effectifs attendus. Plus elle est grande moins l'hypothèse H_0 est plausible.
- Lorsque H_0 est vraie, χ_0^2 est distribuée selon une loi khi-deux à $\nu = k - p - 1$ degrés de liberté, où:
  - k est le nombre de classes retenues.
  - p est le nombre de paramètres estimés.
- Pour un niveau critique α donné, le test consiste à rejeter H_0 si $\chi_0^2 > \chi_{\alpha;\nu}^2$.

Note: Pour un même jeu de données, il est courant que plusieurs distributions ne puissent être rejetées par ce test.

Test d'ajustement du Khi-deux: méthode (iii)

Exercice 2:

On dispose des données suivantes sur une variable X:

Valeurs (x_i)	1	2	3	Total
Effectifs observés (O_i)	28	18	12	58

Tester l'hypothèse selon laquelle les données proviennent d'une population distribuée selon une loi géométrique, c'est-à-dire: $H_0 : X \sim G(p)$. Utiliser $\alpha = 0,05$.

Test d'ajustement du Khi-deux: méthode (iv)

Exercice 3: On dispose des données suivantes sur une variable X:

Intervalle	$[0;0,5[$	$[0,5;1[$	$[1;1,5[$	$[1,5;2[$	$[2;2,5[$	$[2,5;3[$	$[3;\infty[$
Nombre observé	2	23	17	4	2	0	2

Tester l'hypothèse selon laquelle les données proviennent d'une population distribuée selon une loi normale, i.e. $H_0 : X \sim N(\mu, \sigma^2)$. Utiliser $\alpha = 0,05$. La moyenne et l'écart-type de l'échantillon sont $\overline{X} = 1,168$ et $S = 0,591$.

Test d'indépendance entre deux variables

Test d'indépendance entre deux variables

Il arrive en pratique que l'on étudie plusieurs variables simultanément. Dans le cas particulier de deux variables, on peut être amené à vérifier s'il existe un lien entre les deux. La méthode du khi-deux permet d'effectuer ce test.

Exemples:

- On aimerait vérifier si, dans une population donnée, les hommes et les femmes ont la même opinion au sujet du tabagisme. On dit alors qu'on effectue le test de l'indépendance entre le sexe (X) et l'opinion (Y).
- On veut vérifier si le type de pneu (X) est dépendant du kilométrage parcouru avant usure (Y).

Test d'indépendance: méthode

Il s'agit dans ces cas d'un test non paramétrique des hypothèses:

H_0: X et Y sont indépendantes.
H_1: X et Y sont dépendantes.

Afin d'effectuer un tel test, on prélève un échantillon de taille n de la population que l'on classe conjointement selon les r modalités de X et les c modalités de Y. On obtient alors un **tableau de contingence**.

Tout comme le cas du test d'ajustement, le principe du test du Khi6deux consiste à comparer les effectifs observés O_{ij} aux effectifs attendus E_{ij} si H_0 est vraie. Si les deux variables sont indépendantes, les effectifs attendus E_{ij} (avec $i = 1, ..., r$ et $j = 1, ..., c$) sont calculés à partir du tableau de contingence:

$$E_{ij} = \frac{1}{n} \left(\sum_{k=1}^{c} O_{ik} \right) \times \left(\sum_{l=1}^{r} O_{lj} \right)$$

Test d'indépendance: méthode (ii)

La statistique du test est

$$X_0^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

.

- Lorsque H_0 est vraie, la statistique χ_0^2 suit une loi Khi-deux à $\nu = (r - 1) \times (c - 1)$ degrés de liberté.
- Pour un niveau critique α donné, le test consiste à rejeter H_0 si $\chi_0^2 > X_{\alpha;\nu}^2$.

Test d'indépendance: méthode (iii)

Exercice 4:

Une flotte d'autobus est équipée de quatre types de pneus (A, B, C et D). On mesure le kilométrage (en milliers) pour lesquelles on a obtenu les résultats suivants:

Observé	A	B	C	D	Total
< 20	26	23	15	32	96
$[20; 30]$	118	93	116	121	448
> 30	56	84	69	47	256
Total	200	200	200	200	800

Tester si les deux variables sont indépendantes au seuil critique $\alpha = 0,05$.