Chp 4 / Tests non parametrique de comparaison d'échantillons.

I/ Test de Smirnov (à 2 échantillons)

Deux échantillons proviennent-ils de la même population ?

Exemples :

1/ Double correction : Deux correcteurs d'un examen, corrigent-ils de la même maniere?

2/ Efficacité de pieces : Les pieces produites par deux machines differentes, sont-elles de même consistance?

3) Deux champs d'oliviers produisent-ils la même quantité d'olives.

Le test consiste à verifier si deux échantillons sont issus de la même population. c-a-d qu'ils suivent la même loi de fonction de répartition F inconnue.

Soient deux echantillons $X_1, \dots, X_{n_1}$ et $Y_{n_1}, \dots, Y_{n_2}$ issus de deux v.a X et Y respectivement de fonctions de repartition F_1 et F_2 inconnues.

Notons par $F^*_{n_1}$ et $F^*_{n_2}$ les fonctions de répartitions empiriques des deux échantillons respectivement

Test : H_0 : " $F_1 = F_2$ " vs H_1 : " $F_1 \neq F_2$ "

ou H_0 : " $\forall x \; F_1(x) = F_2(x)$ " vs H_1 : " $\exists x \; F_1(x) \neq F_2(x)$ "

La statistique de décision définie par Smirnov est donnée par :

$$T = \sup_x \left| F^*_{n_1}(x) - F^*_{n_2}(x) \right| \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

Pour n_1 et n_2 assez grands, T suit asymptotiquement une loi tabulée. (voir La table de Smirnov à 2 échantillons)

La region critique du test est donnée par :

$$RC = [T > k_\alpha]$$

α un seuil fixé.

II / Test de wilcoxon.

Les deux echantillons $X_1, \dots, X_{n_1}$ et $Y_1, \dots, Y_{n_2}$ proviennent-ils de la même population (ont-ils la même loi)

1/ Considerons la statistique

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbb{1}_{]-\infty, x_i[}(y_j)$$

(Somme des y_j avant chaque x_i)

U varie de 0 à $n_1 n_2$.

Test : "H_0 $F_1 = F_2$" vs "H_1 : $F_1 \neq F_2$"

Sous H_0, $(F_1 = F_2)$, U est uniformement repondue sur $\{0, \dots, n_1 n_2\}$

On a $E(U) = \frac{n_1 n_2}{2}$. En effet,

$$E(U) = \sum_{k=0}^{n_1 n_2} k \frac{1}{n_1 n_2 + 1} = \frac{1}{n_1 n_2 + 1} \sum_{k=0}^{n_1 n_2} k = \frac{1}{n_1 n_2 + 1} \left(\frac{n_1 n_2 (n_1 n_2 + 1)}{2} \right)$$

(suite arithmetique de raison 1)

La région critique du test est donnée par :

$$RC = \left[\left| U - \frac{n_1 n_2}{2} \right| > k_\alpha \right]$$ à un seuil donné.

2/ Autre formulation du test de wilcoxon

Considerons la somme des rangs des X_i dans l'échantillon global notée W_X, on a

$\min W_X = \frac{n_1 (n_1 + 1)}{2}$ et $\max W_X = \frac{n_1 (n_1 + 1)}{2} + n_1 n_2$.

Exp :

X :	0	0,5	1	1,5	2
Y	0,1	0,6	1,1	3	

L'échantillon global ordonné : 0 0,1 0,5 0,6 1 1,1 1,5 2 3

$W_X = 1 + 3 + 5 + 7 + 8 = 24$

$W_Y = 2 + 4 + 6 + 9 = 21$

Rq: Il est facile de voir que $W_X = U + \frac{n_1(n_1+1)}{2}$

Sous H_0, W_X est uniformément repartie sur $\left\{\frac{n_1(n_1+1)}{2}, \ldots, \frac{n_1(n_1+1)}{2} + n_1 n_2\right\}$

$E(W_X) = \frac{n_1 n_2}{2} + \frac{n_1(n_1+1)}{2}$; $Var(W_X) = \frac{n_1 n_2 (n_1+n_2+1)}{12}$

La region critique du test de Wilcoxon est donnée par :

$RC = [|W_X| > k_\alpha]$ α seuil donné.

La loi W_X est tabulée (Voir la table de wilcoxon)

Rq:

Si n_1 et n_2 sont supérieurs à 18, alors la région critique :

$$RC = \left[\left| W_X - \frac{n_1(n_1+n_2+1)}{2} \right| > U_{\alpha/2} \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}} \right)$$

où $U_{\alpha/2} = \phi^{-1}_{N(0,1)}(\alpha)$ (approximation par la loi normale $N(0,1)$)

III/ Comparaison de plusieurs échantillons issus d'une v.a qualitative

Exemple

Considérons les 4 sections d'étudiants de la 1ère année MI.

La variable qualitative est la mention obtenue au BAC

les résultats sont présentés dans le tableau suivant.

	S.M	AB	B	TB
Section A	n_{11}	n_{12}	n_{13}	n_{14}
Section B	n_{21}	n_{22}	n_{23}	n_{24}
Section C	n_{31}	n_{32}	n_{33}	n_{34}
Section D	n_{41}	n_{42}	n_{43}	n_{44}

$i = 1, \ldots, 4$ $\quad j = 1, \ldots, 4$

n_{ij} est le n^{bre} d'étudiants de la section i ayant la mention j

Les 4 échantillons proviennent-ils de la même population ?

Soient k échantillons $ech_1, \ldots, ech_k$, et X une v.a observée sur les k échantillons.

On note par $Mod_1, \ldots, Mod_r$ les r modalités de X

Le tableau des résultats

Ech \ Mod	Mod 1	Mod 2	...	Mod r	Somme des lignes (Colonnes)
ech 1	n_{11}	n_{12}		n_{1r}	$n_{1\cdot}$
ech 2					
⋮					
ech k	n_{k1}	n_{k2}		n_{kr}	$n_{\cdot r}$
somme ligne	$n_{1\cdot}$	$n_{\cdot 2}$		$n_{k\cdot}$	n

Notations

$n_{i\cdot} = \sum_{j=1}^{r} n_{ij}$ est la taille du $i^{ème}$ ech

$n_{\cdot j} = \sum_{i=1}^{k} n_{ij}$ est ~~la taille~~ l'effectif de la $j^{ème}$ Mod

$n = \sum_{i=1}^{k} n_{i\cdot} = \sum_{j=1}^{r} n_{\cdot j} = \sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}$ = taille globale des k échantillons

Test : H_0 : "les k échantillons ont la même loi."

vs H_1 : "les k échantillons sont significativement différents."

- 1er Cas : On suppose que l'on connait sous H_0, les proportions $p_1, \dots, p_r$ de posséder les modalités Mod1, ..., Modr.

La statistique de décision est donnée par :

$$d^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} \left(\frac{n_{i\cdot} p_j - n_{ij}}{n_{i\cdot} p_j} \right)^2$$

d^2 suit la distribution de χ^2_{kr-k} sous H_0

La région critique $RC = [d^2 > k_\alpha]$ α seuil fixé

- 2ème Cas : Les proportions $p_1, \dots, p_r$ sont inconnues.

Souvent, en pratique $p_1, \dots, p_r$ sont inconnues, alors on estime p_j pour $j = 1, \dots, r$ par : $\hat{p}_j = \frac{n_{\cdot j}}{n}$

On obtient ainsi $(r-1)$ estimateurs indépendants $\sum_{j=1}^{r} p_{\cdot j} = 1$

La statistique de décision devient :

$$d_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \left(\frac{n_{i\cdot} \hat{p}_{\cdot j} - n_{ij}}{n_{i\cdot} \hat{p}_{\cdot j}} \right)^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{r} \left(\frac{n_{i\cdot} n_{\cdot j} - n n_{ij}}{n_{i\cdot} n_{\cdot j}} \right)^2$$

Et $d_0^2 \sim \chi^2_{kr-k-(r-1)} = \chi^2_{(r-1)(k-1)}$

Rq : Si la variable X est quantitative, les modalités sont remplacées par des classes.