Tests non-paramétriques

Rappels

➤ En statistique, qu'appelle-t-on un paramètre ?

➤ Dans quelles conditions utilise-t-on des tests non paramétriques ?

1. Tests de fréquences et loi du χ^2

L'idée du test consiste à sommer les différences entre des effectifs observés et des effectifs théoriques. Si cette somme de différences est « suffisamment » grande, on considère que les effectifs observés diffèrent des effectifs théoriques.

Soit X_1, X_2, …, X_k k variables aléatoires indépendantes de même loi normale centrée et réduite, alors la variable X telle que $X = \sum_{i=1}^{k} X_i^2$ suit une loi du χ^2 à k degrés de liberté.

➤ Quelles sont les conditions d'applications (recommandées) pour utiliser un test du χ^2 ?

1.1. Test du χ^2 d'indépendance de deux variables qualitatives

Le tableau suivant résulte d'une expérience demandant à des sujets (experts ou novices en dégustation de bières) de déguster des bières, puis d'exprimer leur préférence soit pour une bière blonde, soit pour une bière brune. Les effectifs obtenus sont répartis de la manière suivante :

	Expert	Novice
Bière blanche	8	22
Bière Brune	15	13

➤ Préciser les nombres de modalités p et q prises respectivement par les variables « Type de bière » et « Niveau d'expertise ».

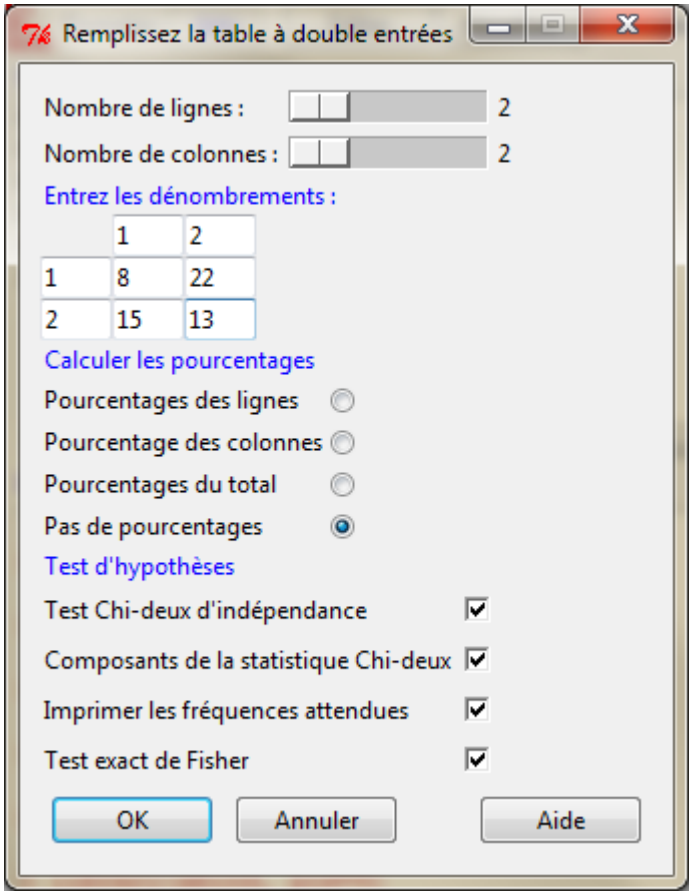
➤ Calculer les effectifs marginaux $N_{i,j}$ pour l'échantillon.

➤ Calculer les effectifs théoriques attendus pour chaque cellule (sous l'hypothèse d'indépendance des évènements).

➤ Calculer la statistique servant à tester l'indépendance des évènements et trouver la valeur de la probabilité associé (au risque 5%).

➤ Tracer la distribution du χ^2 à 1, 10, puis 20 ddl R : Menu **Distributions > Distributions continues > Distribution Chi-deux > graphe de la distribution Chi-deux.**

➤ Vérifier le résultat dans R : Menu **Statistiques > Tables de contingence > Remplir et analyser un tableau à double entrée** :



Fisher exact test : un autre test similaire.

1.2. Test du χ^2 d'ajustement d'effectifs observés à des effectifs théoriques

On a demandé à 87 dégustateurs de classer 5 chocolats par ordre de préférence. Le tableau suivant contient, pour chaque produit, le nombre de sujets qui ont classé ce produit au rang i :

RANGS	choc1	choc2	choc3	choc4	choc5
1	16	20	16	12	24
2	13	20	29	10	14
3	17	11	26	10	23
4	29	17	11	15	15
5	12	19	5	40	11
	87	87	87	87	87

➤ En ne considérant les données que pour un seul produit (une seule colonne), à quelle valeur moyenne aurait-on pu s'attendre pour chaque cellule (valeur théorique) si les produits avaient été classés aléatoirement par les dégustateurs ?

➤ Calculer la statistique permettant de comparer la répartition obtenue à une répartition due au hasard (loi uniforme), au risque 5%, pour le chocolat « choc1 ».

➤ Vérifier le résultat dans R pour tous les chocolats : importer le jeu de données "chocolats.txt", puis dans la fenêtre de script saisir *apply(chocolats,2,chisq.test)*

2. Tests non paramétriques de comparaison de deux groupes

2.1. Test de Mann et Whitney (échantillons indépendants)

Le principe du test de Mann Whitney est simple. On calcule :

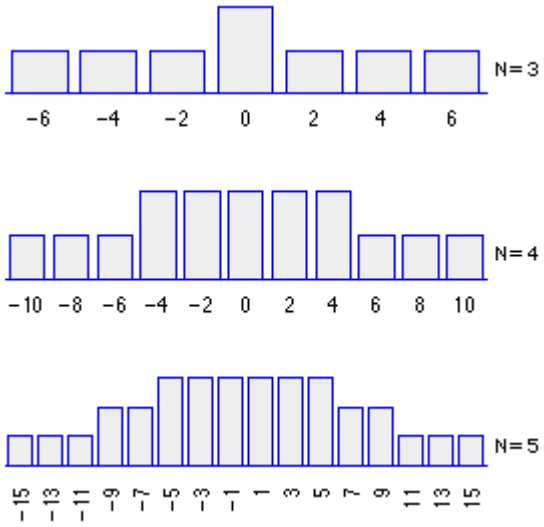
- U1 le nombre de fois où une note du 1er groupe est supérieure à une note du 2e groupe (on compte 0.5 si les notes sont égales)
- U2 le nombre de fois où une note du 2e groupe est supérieure à une note du 1er groupe (on compte 0.5 si les notes sont égales)
- U = inf(U1, U2)
- U1+U2=n1n2

L'hypothèse nulle consiste à affirmer qu'il y a le même nombre de notes du 1er groupe supérieures aux notes du 2e groupe que l'inverse :

H0 : P(U1 > U2) = 0.5, $m_U = \frac{n_1 n_2}{2}$, les distributions sont identiques

H1 : P(U1 > U2) ≠ 0.5, les distributions sont différentes

On peut remarquer que U est approximativement normalement distribué :



$z = \frac{U - m_U}{\sigma_U} \sim \mathcal{N}(0,1)$, avec $m_U = \frac{n_1 n_2}{2}$ et $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$

Par exemple, voici une série de notes données à des cakes A et B, par des dégustateurs différents.

A	5	7	5	4	7	3	4	5	5	4	2	
B	7	6	7	5	8	6	7	5	6	8	5	6

➤ Calculer la statistique de Mann Whitney. Conclure.

➤ Vérifier le résultat dans R : importer le jeu de données "non_parametric_sample.csv" (l'appeler data), puis saisir dans la fenêtre de script : *wilcox.test(data$N1, data$N2, alternative='less', paired=FALSE)*.

2.2. Test de Wilcoxon (échantillons appariés)

Le test de Wilcoxon permet de se débarrasser de l'effet sujet en travaillant sur les différences d_i de notes entre deux produits :

- On classe les d_i par ordre de valeurs absolues croissantes,
- On calcule W_- la somme des rangs négatifs,
- On calcule W_+ la somme des rangs positifs,
- W = inf(W_-, W_+)

L'hypothèse nulle consiste à affirmer que la somme des rangs négatifs est égale à la somme des rangs positifs :

H0 : $W_+ = W_- = n(n+1)/4$, $m_W = 0$, les distributions sont identiques

H1, $W_+ \neq W_-$, les distributions sont différentes

Si n est suffisamment grand (n > 20) :

$$Z = \frac{W_+ - m_W}{\sigma_W} \sim \mathcal{N}(0,1), \text{ avec } m_W = n(n+1)/4 \text{ et } \sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

Par exemple, voici une série de notes données à des cakes B et C, par les mêmes dégustateurs :

B	7	6	7	5	8	6	7	5	6	8	5
C	3	3	7	6	6	3	7	1	2	5	7

➤ Calculer la statistique de Wilcoxon. Conclure.

➤ Vérifier le résultat dans R : importer le jeu de données "non_parametric_sample.csv", puis **Menu statistiques > Tests non paramètriques > Test Wilcoxon apparié, Première variable = N1, Seconde variable = N2, Hypothèse alternative = Différence > 0, Type de test = Approximation normale.**

3. Tests non paramétriques pour la comparaison de plus de deux groupes

3.1. Test de Kruskall et Wallis (échantillons indépendants)

Pour effectuer ce test, on remplace les notes des p produits par leur rang, puis on calcule la somme R_i de ces rangs par produit. Ensuite, on compare à l'aide d'un test du χ^2 (à p-1 ddl) les R_i aux sommes de rangs théoriques $\frac{n_i(n+1)}{2}$ si tous les produits étaient identiques. Pour cela, on utilise la statistique suivante :

$$Q = \frac{12}{n(n+1)}\left[\sum_i \frac{1}{n_i}\left(R_i - n_i\frac{n+1}{2}\right)^2\right] = \frac{12}{n(n+1)}\left(\sum_i \frac{R_i^2}{n_i}\right)\text{-}3(n+1)$$

Si on veut déterminer quelles moyennes diffèrent significativement :

- on calcule les différences entre les rangs moyens des produits $d_{i,j} = \frac{\frac{R_i}{n_i} - \frac{R_j}{n_j}}{\sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}$
- on compare à la valeur de z de la table de la loi normale au niveau $\frac{2\alpha}{p(p-1)}$

Par exemple, voici une série de notes données à des cakes A, B et C, par des dégustateurs différents :

A	5	7	5	4	7	3	4	5	5	4	2	
B	7	6	7	5	8	6	7	5	6	8	5	6
C	3	3	7	6	6	3	7	1	2	5	7	6

➤ Calculer la statistique de Kruskall et Wallis. Conclure. Préciser les moyennes significativement différentes.

➤ Vérifier le résultat dans R : importer le jeu de données "non_parametric_sample.csv" (l'appeler data), puis saisir dans la fenêtre de script :

N1=data$N1 # Vecteur des notes de 11 juges pour le produit A

N2=c(data$N2,6) # Vecteur des notes de 12 juges pour le produit B

N3=c(data$N3,6) # Vecteur des notes de 12 juges pour le produit C

kruskal.test(list(N1,N2,N3))

library(pgirmess)

mat=as.matrix(data[,-1])

vec=as.vector(t(mat))

categ<-as.factor(rep(c("P1","P2","P3"),times=11))

kruskalmc(vec, categ, cont="one-tailed")

3.2. Test de Friedman (échantillons appariés)

Le test de Friedman correspond à une analyse de la variance effectuée sur les rangs. Elle permet de prendre en compte les données de tous les produits afin de déterminer si les produits ont été classés de manière (significativement) différente ou non. Ce test est fréquemment employé en analyse sensorielle, car il correspond à une expérience équilibrée où n sujets ont noté chacun les p produits.

L'idée du test est globalement la même que celle du test de Kruskall Wallis. On utilise la statistique suivante :

$$F = \frac{12}{np(p+1)}\left(\sum_i R_i^2\right) - 3n(p+1)$$

Si on veut déterminer quelles moyennes diffèrent significativement, on compare chaque différence ($R_i - R_j$) à $\delta = z\sqrt{\frac{np(p+1)}{6}}$, où z est la valeur lue dans la table normale au niveau $\frac{2\alpha}{p(p-1)}$.

Par exemple, voici une série de notes données à des cakes A, B et C, par les mêmes dégustateurs:

	S1	S2	S3	S4	S5	S6	S7	S8	S8	S10	S11
A	5	7	5	4	7	3	4	5	5	4	2
B	7	6	7	5	8	6	7	5	6	8	5
C	3	3	7	6	6	3	7	1	2	5	7

➤ Calculer la statistique de Friedman. Conclure. Préciser les moyennes significativement différentes.

➤ Vérifier le résultat dans R : importer le jeu de données "non_parametric_sample.csv", puis **Menu statistiques > Tests non paramétriques > Test de somme de rangs de Friedman, Variables pour les mesures répétées = N1 + N2 + N3** ou saisir puis exécuter dans la fenêtre de script :

mat=as.matrix(data[,-1])

```
notes=as.vector(t(mat))

juges=rep(c("J1","J2","J3","J4","J5","J6","J7","J8","J9","J10","J11"),times=11,each=3)

produits=rep(c("P1","P2","P3"),times=11)

table=as.data.frame(cbind(juges,produits,as.numeric(notes)))

colnames(table)=c("juges","produits","notes")

table$notes=as.numeric(as.character(table$notes))

comparison=friedman(table$juges,table$produits, table$notes,alpha=0.05, group=TRUE,main="Friedman test")

bar.group(comparison,density=3,border="red",col="blue")
```

4. Analyse des données collectées en TP.