Khi_2 et Tests non-Paramétriques

1 Statistiques paramétriques vs non-paramétriques

1.1 Contraintes des tests paramétriques

- Normalité des distributions
- Homogénéité des variances (Homosédastisité)
- Peu applicable aux effectifs réduits - approx. $n < 30$

1.2 Caractéristiques principales des tests non-paramétriques

- Pas de contrainte sur la population dont est extrait l´échantillon.
- Seuls tests applicable pour un échantillon de taille inférieure à 6.
- Seuls tests permettant de comparer des échantillons issus de populations ayant des distributions différentes.
- Seuls tests traitant des données qualitatives exprimées soit en variables nominales soit par la comparaison de rangs.
- **Inconvénient** : moins puissants que les tests paramétriques lorsqu'ils sont utilisables.

2 Vocabulaire initial - Petit rappel

2.1 Les hypothèses

Une hypothèse est une assertion sur les caractéristiques des données analysées et du modèle qu'on souhaite leur appliquer.
L'hypothèse H_0 est toujours la première hypothèse testée. Lorsque l'on compare deux échantillons, deux populations, deux groupes de données . . . , H_0 est l'hypothèse nulle selon laquelle il n'existe pas de différence entre ces deux ensembles. On cherche donc souvent à rejeter H_0.
Lorsque l'on a rejeté H_0, il est ensuite possible de tester une hypothèse alternative, appelée H_1.

Exemple : pour tester l'efficacité d'un médicament, on constitue 2 groupes de patients, l'un prenant le médicament, l'autre un placebo. H_0 est l'hypothèse selon laquelle il n'y a pas de différence observée entre ces deux groupes. H_1 pourrait être l'hypothèse selon laquelle les patients ayant pris le médicament sont moins malades que ceux ayant pris le placebo.

2.2 Tableau de contingence

Un tableau de contingence permet de représenter la répartition d'effectifs d'un échantillon en fonction de la valeur d'une observation, Cette valeur peut être discrète ou continue, dans ce cas elle sera modélisée en classes.

Exemple : 137 patients atteints de cirrhose sont divisés en 2 groupes. Le premier groupe C_1 recoit un médicament, le secoond C_2 un placebo. Le stade d´évolution de la maladie constitut la variable observée et est définie en 3 modalités pour 3 stades d´évolution : $d_1 = 1$ $d_2 = 2$ $d_3 = 3$.

	Stade			
	1	2	3	Total
placebo	13	29	26	68
traitement	16	37	16	69
Total	29	66	42	137

	Stade			
	1	2	3	Total
placebo	0.191	0.426	0.382	1
traitement	0.232	0.536	0.232	1
Total	0.212	0.482	0.306	1

On peut voir que la proportion de patients en stade 3 pour les patients sous traitement est moins forte que celle des patients sous placebo.
Question : cette différence est-elle significative ?

3 Test du Khi-2

Le test du Khi-2 aussi appelé test du Khi-2 de Pearson (qui a établi la théorie générale de la corrélation) consiste à mesurer l'écart qui existe entre la distribution des effectifs théoriques et la distribution des effectifs observés d'un échantillon, et à tester si cet écart est suffisamment faible pour être imputable aux fluctuations d'échantillonage.

Principe :

3.1 Calcul des effectifs théoriques

Si nous disposons d'un T1 tableau de contingence des effectifs observés, on définit un tableau T2 d'effectifs théoriques :

T1	Xa	Xb	
Ya	a	b	L1
Yb	c	d	L2
	C1	C2	N

T1	Xa	Xb	
Ya	a'	b'	L1
Yb	c'	d'	L2
	C1	C2	N

Les valeurs a',b',c'et d' sont calculés suivant les formules suivantes :

$$a' = \frac{C1 * L1}{N} \quad b' = \frac{C2 * L1}{N} \quad c' = \frac{C1 * L2}{N} \quad d' = \frac{C2 * L2}{N}$$

Attention : dans un tableau à 4 cases, chaque effectif théorique doit être au moins égale à 5.

3.2 Calcul du Khi-2 et interprétation

La formule générale du Khi-2 est : $Somme = \frac{(Observes - Theoriques)^2}{Theoriques}$

Soit : $X^2 = \frac{(a-a')^2}{a'} + \frac{(b-b')^2}{b'} + \frac{(c-c')^2}{c'} + \frac{(d-d')^2}{d'}$

La valeur obtenue est ensuite comparée avec un seuil lu dans la table du Khi-2 pour un degré de liberté et pour un risque d'erreur fixé.

- **Degré de liberté** Pour interpréter la valeur de Khi-2 obtenue, on doit connaitre le degré de liberté (d.d.l.) du modèle.
  Degrés de liberté = (nb lignes - 1) X (nb colonnes - 1)
- **Risque d'erreur** En général on accepte un risque d'erreur de 0.5%, si l'on désire un test très stringeant on peut aussi choisir un risque de 0.1%.

Si la valeur du Khi_2 est supérieure à celle fournie dans la table, on rejete H_0 et on dira qu'il n'y a pas de lien entre les variables.

3.3 Exercice

1. Calculer le tableau des effectifs théoriques pour le tableau de contingence des patients atteints de cirrhose.
2. Saisir les valeurs dans R et appliquer un test de Khi_2 : `chisq.test(matrice)`

4 Tests de rang

Lorsque 2 échantillons ne suivent pas les distributions normales, il est possible de comparer leur rang dans l'échantillon plutôt que la valeur.

4.1 Test de Wilcoxon

Définition de H_0 : la différence moyenne entre les deux mesures est nulle.

- Test 2 échantillons indépendants.
- Hypothèse : les 2 groupes de valeurs ont-ils des distributions identiques ?
- Utilise le rang des valeurs calculé dans un ensemble unique.
- Statistique de rang : s'affranchit de la différence de moyenne.

Méthode : Soit deux échantillons E_1 et E_2 de taille n_1 et n_2, on rassemble les deux échantillons et on range les valeurs dans l'ordre (de la plus petite à la plus grande). On calcule ensuite pour chaque échantillon la somme des rangs de ses valeurs (si plusieurs valeurs sont égales on leur attribuera la moyenne des rangs qu'elles auraient eu si elles n'avaient pas été égales). On note U la plus petite de ces deux sommes. Pour savoir si l'on rejete H_0 avec un risque d'erreur α il faut consulter une table du test de Wilcoxon qui suivant les effectifs n_1 et n_2 donne quelle est la valeur maximale attendue pour rejeter H_0 en fonction du risque choisi 0.005, 0.01, 0.25. ect...Si la valeur de la statistique de Wilcoxon est supérieure à cette valeur, on ne peut pas rejeter H_0.

NB : la somme des rangs doit être identique.

Exercice Pour ce jeu de valeurs, on ne peut pas rejeter H_0.
Des prélévements sanguins ont été réalisés sur des Natives Amercicains et sur des Caucasiens, voici les valeurs pour chaque échantillon :

- Native Americains : 8.5,9.48,8.65,8.16,8.83,7.76,8.63
- Caucasiens : 8.27,8.2,8.25,8.14,9.00,8.1,7.2

Commande R : `wilcox.test(Na, Ca)` - impossible d'utiliser `paired=TRUE` car un individu ne peut être à la fois Caucasien et Native Americain..
Résultats :

```
        Wilcoxon rank sum test

data:  Na and Ca
W = 35, p-value = 0.2086
alternative hypothesis: true location shift is not equal to 0
```

4.2 Test de Kruskal et Wallis

- Généralisation du test de Wilcoxon a N échantillons.
- Permet de déterminer si au moins un échantillon est différent.
- Commande R : `kruskal.test(vecteur,facteur)`
- Exercice : tester la co-localisation de protéines sur la membranne (analyse d'une image - qui sera fournie sous forme de 3 colonnes chacune representant une couleur - RGB).

4.3 Corrélations de Spearman

Le coefficient de corrélation de Spearman, appelé R ou ρ est similaire au coefficient de corrélation de Pearson mais est calculée sur les rangs. Pour calculer le rang R de Spearman il faaut que les variables aient été mesurées sur une échelle ordinale, pour pouvoir être ordonnées.

Commande R : `cor(V1,V2,method=''spearman'')`